**Peer review history**

**MS Title: Faces do not attract more attention than non-social distractors in the Stroop task**

Author names: Anna Henschel, Hannah Bargel, Emily S. Cross

Handling editor: Christopher R Madan

**First editor decision**

June 15, 2020

Dear Miss Anna Henschel,  
  
After review, we have reached a decision regarding your submission to Collabra: Psychology, "Faces do not attract more attention than non-social distractors in the Stroop task". Our decision is to request revisions of the manuscript prior to acceptance for publication. The full review information should be included at the bottom of this email.  
  
Three expert reviewers have provided comments on your work and find that some revisions are necessary before it would be suitable for publication. The reviewers have made suggestions that cover most of the work and many of these points are critical, while described in much more detail in the reviewer comments below, I would like to highlight a few that I found particularly important to be addressed. (1) Structure of the introduction, as detailed by Reviewer 2. (2) Reporting the number of excluded trials. (3) Further methodological details need to be included. All of the reviewers thought the figures and tables were well done.  
  
To access your submission account, follow the below instructions:  
1) login to the journal webpage with username and password  
2) click on the submission title  
3) click 'Review' menu option  
4) download Reviewed file and make revisions based on review feedback  
5) upload the edited file  
6) Click the 'notify editor' icon and email the confirmation of re-submission and any relevant comments to the journal.  
  
Please ensure that your revised files adhere to our author guidelines, and that the files are fully copyedited/proofed prior to upload. Please also ensure that all copyright permissions have been obtained. This is the last opportunity for major editing;, therefore please fully check your file prior to re-submission.  
  
If you have any questions or difficulties during this process, please do contact us.  
  
Please could you have the revisions submitted by September 1, 2020. If you cannot make this deadline, please let us know as early as possible.  
  
Kind regards,  
  
Dr Christopher R Madan  
University of Nottingham  
christopher.madan@nottingham.ac.uk  
------------------------------------------------------  
Reviewer 1:  
  
  
1) General comments and summary of recommendation  
Describe your overall impressions and your recommendation, including changes or revisions. Please note that you should pay attention to scientific, methodological, and ethical soundness only, not novelty, topicality, or scope. A checklist of things to you may want to consider is below:  
 - Are the methodologies used appropriate?  
 - Are any methodological weaknesses addressed?  
 - Is all statistical analysis sound?  
 - Does the conclusion (if present) reflect the argument, is it supported by data/facts?  
 - Is the article logically structured, succinct, and does the argument flow coherently?  
 - Are the references adequate and appropriate?:  
The authors were interested in human-robot social interaction with a focus on social motivation towards artificial agents. In two studies, they investigated the effects of distractors with varying social salience on an adaptation of a classic Stroop task. In both studies conducted, a classic Stroop effect emerged, yet there was no significant effect of salient social cues (human face distractors) capturing attention.  
  
Comments:  
  
The authors do a good job setting up the study's aims by providing a big picture question. However, the organization of the literature review is somewhat hard to follow. The studies reviewed contain details that do not seem necessary and distracts the reader from the main point. For example, the sample sizes each study had are unnecessary as well as step-by-step accounts of their experimental procedures. I think the introduction would benefit from succinct accounts of the main manipulations of the studies (e.g., direct vs averted gaze, open vs closed eyes) and relevant results (as the authors already do). Additionally, we do not see synthesis or commentary by the authors. It would be useful if they could provide their own interpretation of the literature and its implications.  
  
Refrain from using direct quotations as it takes away from the authors' original thinking. Where "centrally presented direct gaze delay[ing] attentional disengagement and recruit[ing] cognitive processing resources, and hence, processing times of the peripheral targets and the Stroop interference are increased” is written, perhaps the authors could paraphrase the central idea.  
  
Details such as “The experimenter explained the procedure of the study and ensured participants understood the task” and “During this part of the study, the light in the cubicle was still switched on, and was switched off when participants started the test phase of the experiment“ do not necessarily have to be in the body of the paper. If the authors would like to keep this, it would be better to move this in Supplementary Materials.  
  
It might be worth thinking about the importance of controlling for emotional valence when using human face stimuli and provide this as a motivation for using neutral stimuli, which is missing in the paper. This might be useful in the Introduction or in the General Discussion and set this up as a limitation or a note for future studies. There are several studies showing that emotional human faces (or emotional stimuli in general) have been found to capture attention faster than neutral when task-irrelevant (e.g., Theuuwes & Van der Stigchel 2006; Pessoa, McKenna, Gutierrez, & Ungerleider, 2002; Vuilleumier, 2002) This could influence the degree of social salience of social agents.  
  
Statistical analyses are sound.  
  
  
  
2) Figures/tables/data availability:  
Please comment on the author’s use of tables, charts, figures, ifrelevant. Please acknowledge that adequate underlying data is available to ensure reproducibility (see open data policies per discipline of Collabra here).:  
While sufficiently described, I would like to see a schematic of the Stroop task as it is always helpful to readers. The plots are beautiful.  
  
  
  
3) Ethical approval:  
If humans or animals have been used as research subjects, and/or tissue or field sampling, are the necessary statements of ethical approval by a relevant authority present? Where humans have participated in research, informed consent should also be declared.  
If not, please detail where you think a further ethics approval/statement/follow-up is required.:  
Ethical approval is present and informed consent is declared.  
  
  
  
4) Language:  
Is the text well written and jargon free? Please comment on the quality of English and any need for improvement beyond the scope of this process.:  
English is excellent. As per my comments above, the authors could be more concise when reviewing the literature, as well as describing the methods.  
  
------------------------------------------------------  
  
Reviewer 2:  
  
  
1) General comments and summary of recommendation  
Describe your overall impressions and your recommendation, including changes or revisions. Please note that you should pay attention to scientific, methodological, and ethical soundness only, not novelty, topicality, or scope. A checklist of things to you may want to consider is below:  
 - Are the methodologies used appropriate?  
 - Are any methodological weaknesses addressed?  
 - Is all statistical analysis sound?  
 - Does the conclusion (if present) reflect the argument, is it supported by data/facts?  
 - Is the article logically structured, succinct, and does the argument flow coherently?  
 - Are the references adequate and appropriate?:  
This manuscript featured two nearly-identical experiments designed to examine the impact of social salience during the Stroop Task. Specifically, the authors varied the social salience of distractors, including images of human faces, robot faces, objects that looked like faces, and flowers, and predicted that faces would amplify the stroop effect due to their high social salience. They found evidence for this in Experiment 1, but when they controlled for a stimulus confound in Experiment 2, they were not able to reject the null hypothesis. This investigation has some strengths. For example, comparing evaluations of human and robot faces is interesting, their analyses were appropriately simple and clear, and basic visual characteristics of the stimuli were well controlled for. The authors also took care to evaluate whether their data provided evidence in favor of the null hypothesis, which I appreciated. In general, I don’t have any concerns about the methods, the data, the analyses, or their interpretation. However, I did have quite a few major concerns about more general issues that tempered my enthusiasm for the work. I’ll explain these in more detail below.  
  
First, the majority of the Introduction is highly specific to research on gaze perception, as is the Discussion. And yet gaze direction is not examined, nor is it even important, in the current investigation. This reflected a larger issue with the Introduction, which seems to cover many topics before finally focusing on the hypothesis and aims of the investigation. It didn’t feel like the gaps in the literature (as described) necessarily led to the current work and its design. Rather, it felt a bit like a literature review was forced around the current experiment. In this sense, I did not think that the article was as logically structured as it could have been.  
  
Second, it’s not clear to me that the design the authors have selected is the best one to examine their main question. In describing previous work by Conty and then Chevallier, the authors state that “the lack of difference in arousal would lead to “centrally presented direct gaze delay[ing] attentional disengagement and recruit[ing] cognitive processing resources, and hence, processing times of the peripheral targets and the Stroop interference are increased”. They then go on to state that testing this claim was the goal of the current investigation. But as far as I can tell, their design does not examine arousal, nor is it confirmed that the stimuli themselves differ in the extent to which they arouse the participants. I don’t have an issue with their stimulus choices per se (they’re rather clever), but they don’t seem to fall out of the literature reviewed, and it seems like not manipulating gaze direction was a missed opportunity. Thus, although I have no issues with the analysis or the data, it’s not clear to me that the conclusions reflect the underlying question, at least as it’s framed in the introduction.  
  
Third, critical information about the task is missing. Yes, the Stroop Task is well known and quite simple, but it isn’t adequately described in the Methods, nor is any background provided about the history of the task or its mechanisms. This wouldn’t be too difficult to rectify, but as it stands, it’s a curious omission.  
  
Finally, I found it hard to process the takeaway message of the manuscript. The authors found evidence against the null hypothesis in Experiment 1, but there were issues with a stimulus confound, and then in Experiment 2, there appears to be no effect of category, but the authors were at the same time not able to support a case in favor of the null--of faces not drawing more attention in the Stroop task. In other words, it’s just really difficult to get a clear sense of what the study demonstrates, and thus what it’s impact will be.  
  
  
  
2) Figures/tables/data availability:  
Please comment on the author’s use of tables, charts, figures, ifrelevant. Please acknowledge that adequate underlying data is available to ensure reproducibility (see open data policies per discipline of Collabra here).:  
The tables and figures are nice. Well done.  
  
  
  
3) Ethical approval:  
If humans or animals have been used as research subjects, and/or tissue or field sampling, are the necessary statements of ethical approval by a relevant authority present? Where humans have participated in research, informed consent should also be declared.  
If not, please detail where you think a further ethics approval/statement/follow-up is required.:  
This seemed adequate.  
  
  
  
4) Language:  
Is the text well written and jargon free? Please comment on the quality of English and any need for improvement beyond the scope of this process.:  
In general, yes. The quality of English was good.  
  
------------------------------------------------------  
  
Reviewer 3:  
  
  
1) General comments and summary of recommendation  
Describe your overall impressions and your recommendation, including changes or revisions. Please note that you should pay attention to scientific, methodological, and ethical soundness only, not novelty, topicality, or scope. A checklist of things to you may want to consider is below:  
 - Are the methodologies used appropriate?  
 - Are any methodological weaknesses addressed?  
 - Is all statistical analysis sound?  
 - Does the conclusion (if present) reflect the argument, is it supported by data/facts?  
 - Is the article logically structured, succinct, and does the argument flow coherently?  
 - Are the references adequate and appropriate?:  
Review of 329-4052-1-RV. Henschel et al. “Faces do not attract more attention than non-social distractors in the Stroop task”  
  
In this preregistered study, the question was investigated whether human faces automatically attract attention more than other types of distractors (human-like faces or non-faces) while participants solve a Stroop task. Two studies are presented, where in study 1 (N=39) a small effect seemed to favour the prediction with slightly increased Stroop effects in the presence of human faces, but a second study (N=51) that increased the number of unique distractor images failed to find differences between response times to the different distractor types.  
I enjoyed reading this well-written manuscript, the theoretical background is nicely developed, the methods are sound and the statistical analyses are sophisticated. I have a few observations nevertheless that I would like the authors to consider.  
  
1. The task involved the concurrent presentation of a Stroop colour-word interference test and distractors. While this setup seems to follow methods by Conty et al. (2010), the main measure involves a form of “dual distraction” - distraction from the colour-incongruent words and distraction from the faces. It would have been nice to have baseline trials in which no distractors were shown, in order to evaluate people’s Stroop effect per se, without imposing a second task.  
  
2. The visual layout of these stimuli on the screen was not entirely clear to me as it is not shown in the figures, although described in the text. Was the distance between the words and the distractors different or the same as in the original study? In other words, was it perhaps easier to ignore the distractors here than in Conty et al., especially given that ignoring the distractors was indeed what participants were asked to do.  
  
3. The skewed RTs (as nicely shown in Figure 4, for study 2) were analysed in raw format without further transformation (log) – have the authors tried to analyse log-transformed RTs?  
  
4. Perhaps most critically, it seems that trials in which response times larger than 2 SD above the sample mean were excluded. This is likely too stringent since individual response times are quite variable and in fact, the most interesting trials in this task would be those in which distraction was maximal (i.e., response times are long). In order to avoid “overcleaning”, I would strongly recommend either not to exclude long trials, or to use individual response time distributions - exclude trials that are 2 (or 3) SDs above each participant’s own mean RT instead of the sample mean. Numbers of excluded trials and excluded trials per condition are also not reported and should be added. As a result, the remaining trials could be biased towards those that were not distracting (no matter which condition).  
  
5. The Bayesian analyses are sufficiently esoteric for me that I require more clarification here.  
  
a. Page 13 states “Given the results of Study 2, we explored the extent to which our data provided compelling evidence for the null hypothesis (no difference in reaction times in the incongruent and neutral conditions when human faces are presented)“. This implies to me that the null hypothesis would predict no Stroop effect when the human faces were presented. I believe this is not what the authors meant, but instead that the size of the Stroop effect would not differ between distractor conditions. Is this the case? If so, this needs to be changed in the text.  
b. How are R-hat values of 1.00 for each of the tested parameters in Table 3 to be understood?  
c. The ROPE outcomes do not support the presence of a Stroop effect at all, if I understood this correctly. The size of the general Stroop effect was sufficiently large, in both studies, based on the conventional outcomes (study 1: F(1, 38)= 39.24, p<.001, ηG2= .03; study 2: F(1, 50)=70.31, p<.001, ηG2=.06). Can the authors comment more directly on this discrepancy? And if the outcomes of the Bayesian analysis are taken seriously, what are the consequences for the rest of the paper? For example, page 15 in the discussion states “While we again observed the predicted Stroop effect” – did you? The different outcomes need to be reconciled better, in my opinion.  
d. Figure 5 is unclear to me. What is zero on the x-axis – this can’t be “reaction time (s)” ? Also, going back to point 4 a) is this testing the presence/absence of any Stroop effect or the slow-down of RTs (i.e., bigger Stroop effect) for human faces compared to the other conditions?  
  
Minor comments  
  
6. Using mirror-images also in study 2 arguably may not have created unique distractors. A mirror image could act as a particularly strong distractor, as it would appear familiar but not identical. This could be considered in the limitations section.  
  
7. The decision to move the stimulus rating into a supplement abbreviated the rating outcomes presented in the paper, but I would still have liked to see some details. In fact, the supplement also does not state what exactly was being judged regarding these stimuli. The paper states on page 11, “mind perception of different kinds of agents” – what does this mean and what was the actual outcome of the ratings? Is it relevant or irrelevant for this paper?  
  
8. Some missing details on the Stroop task itself included the number/ratio of congruent and incongruent trials, and any restrictions regarding the switch between the two (e.g., no more than 2 incongruent trials after each other etc.).  
  
9. Since several participants were excluded, I wonder whether these criteria were too stringent. At least the method of excluding participants should be detailed. For example, excluding participants with ASD diagnoses – how was this done?  
  
  
  
2) Figures/tables/data availability:  
Please comment on the author’s use of tables, charts, figures, ifrelevant. Please acknowledge that adequate underlying data is available to ensure reproducibility (see open data policies per discipline of Collabra here).:  
Very nice use and high quality of Figures  
  
  
  
3) Ethical approval:  
If humans or animals have been used as research subjects, and/or tissue or field sampling, are the necessary statements of ethical approval by a relevant authority present? Where humans have participated in research, informed consent should also be declared.  
If not, please detail where you think a further ethics approval/statement/follow-up is required.:  
Ethical approval was obtained from the University of Glasgow ethics review board (300170224).  
  
  
  
4) Language:  
Is the text well written and jargon free? Please comment on the quality of English and any need for improvement beyond the scope of this process.:  
English is appropriate

Author response to reviewers

August 7, 2020

**Editor**

*Summary: Three expert reviewers have provided comments on your work and find that some revisions are necessary before it would be suitable for publication. The reviewers have made suggestions that cover most of the work and many of these points are critical, while described in much more detail in the reviewer comments below, I would like to highlight a few that I found particularly important to be addressed. (1) Structure of the introduction, as detailed by Reviewer 2. (2) Reporting the number of excluded trials. (3) Further methodological details need to be included. All of the reviewers thought the figures and tables were well done.*

Response: We are very grateful for the reviewers’ detailed comments and suggestions and the editor’s synthesis of the overarching points on how best to improve this paper. In the revised manuscript and our responses below, we detail how we have taken this feedback onboard. In line with the suggestions of the reviewers we have placed the strongest emphasis on reworking the introduction and the discussion section, providing more synthesis, critical commentary and making more explicit our study’s rationale. We have detailed the exact number of excluded trials in the supplementary materials, as well as added the overall number of excluded trials in each experiment to their respective results sections. In order to ensure full transparency, we included the alternative reaction time pre-processing method which was recommended by one of the reviewers in the supplementary materials as well. Finally, we have carefully followed the reviewers’ suggestions to include more details on our methods while at the same time removing some information that was flagged as redundant. We highlighted the changes to the manuscript in bold typeface.

**Reviewers**

*Reviewer D:* *The authors were interested in human-robot social interaction with a focus on social motivation towards artificial agents. In two studies, they investigated the effects of distractors with varying social salience on an adaptation of a classic Stroop task. In both studies conducted, a classic Stroop effect emerged, yet there was no significant effect of salient social cues (human face distractors) capturing attention.*

**D.1a: The authors do a good job setting up the study's aims by providing a big picture question. However, the organization of the literature review is somewhat hard to follow. The studies reviewed contain details that do not seem necessary and distracts the reader from the main point. For example, the sample sizes each study had are unnecessary as well as step-by-step accounts of their experimental procedures. I think the introduction would benefit from succinct accounts of the main manipulations of the studies (e.g., direct vs averted gaze, open vs closed eyes) and relevant results (as the authors already do). Additionally, we do not see synthesis or commentary by the authors. It would be useful if they could provide their own interpretation of the literature and its implications.**

Response: As the introduction has been criticized by all three reviewers, we have made major changes to its structure and content, now focusing less on very detailed accounts of each study’s experimental procedure and proving a bird’s eye view on the current state of the art social attentional capture research.

Thus, we have removed superfluous details such as participant numbers in the revised introduction, for example:

P.8, l. 237-239: Importantly, the authors tested the paradigm in two groups of children: typically developing boys and a group of male adolescents with Autism Spectrum Condition (ASC).

We agree that our evaluations of the literature should include more critical reflection, which we have added as well in rewriting the introduction.

P.6, l. 187-191: While the evidence on how deeply illusory faces are perceived as social is mixed, they constitute an ideal control for human facial features in social attentional capture tasks. This also raises the question how deliberate pareidolic faces, such as humanoid robots, might engage our visual attention, as these agents are capable of at least some interactions with the physical world.

The more strongly emphasized synthesis and critical commentary is especially evident in the revised discussion, for example:

P.21, l. 522-529: Many studies report effects based on very small samples (some as small as 8 participants per experiment; Ariga & Arihara, 2017; Miyazaki, Wake, Ichihara, & Wake, 2012; Sato & Kawahara, 2015), make bold statements based on modest statistical evidence (“the three-way interaction approached significance, F(2,76) = 2.46, p<.10”, p. 1103, Hietanen et al., 2016) or use small sets of distractor images which are repeated across many experimental trials (Bindemann et al., 2007; Theeuwes & Van der Stigchel, 2006). Indeed, some of these problematic confounds have been highlighted and tested by Pereira and colleagues (2019; 2020).

And:

P.21, l. 539-545: While a different task was used in these studies, the authors’ findings closely align with ours: faces are not reliably capturing attention and impairing the performance on an unrelated cognitive task. Interestingly, in a direct replication of Bindemann and colleagues (2007), using less well-controlled stimuli, the authors were able to replicate the effect of attentional capture by task-irrelevant faces, providing convincing evidence for systematic confounds obscuring the true picture in the existing literature.

**D.1b: Refrain from using direct quotations as it takes away from the authors' original thinking. Where "centrally presented direct gaze delay[ing] attentional disengagement and recruit[ing] cognitive processing resources, and hence, processing times of the peripheral targets and the Stroop interference are increased” is written, perhaps the authors could paraphrase the central idea.**

Response: This point is well taken, and similar points have been raised by other reviewers. In the revised version of the manuscript, we have completely restructured/rewritten the introduction taking this feedback onboard.

**D.1c: Details such as “The experimenter explained the procedure of the study and ensured participants understood the task” and “During this part of the study, the light in the cubicle was still switched on, and was switched off when participants started the test phase of the experiment“ do not necessarily have to be in the body of the paper. If the authors would like to keep this, it would be better to move this in Supplementary Materials.**

Response: We agree that this information is unnecessary and have removed it.

**D.1d: It might be worth thinking about the importance of controlling for emotional valence when using human face stimuli and provide this as a motivation for using neutral stimuli, which is missing in the paper. This might be useful in the Introduction or in the General Discussion and set this up as a limitation or a note for future studies. There are several studies showing that emotional human faces (or emotional stimuli in general) have been found to capture attention faster than neutral when task-irrelevant (e.g., Theuuwes & Van der Stigchel 2006; Pessoa, McKenna, Gutierrez, &** **Ungerleider, 2002; Vuilleumier, 2002). This could influence the degree of social salience of social agents.**

Response: Indeed, emotional valence of faces plays a crucial role in social interaction and has been repeatedly shown to influence attention differentially. Thank you for the helpful literature recommendations. We added our rationale for selecting neutral faces to the methods:

P.10, l. 299-306: The rationale behind including only neutral faces was that emotional facial cues have been shown to draw attention, especially in comparison to neutral facial expressions (Pessoa, McKenna, Gutierrez, & Ungerleider, 2002; Theeuwes & Van der Stigchel, 2006; Vuilleumier, 2002). An independent sample rated the first pool of human and robot images, resulting in a pre-selection of more neutrally perceived faces (more details can be found in the Supplementary Materials).

To illustrate the point of possibly more varied robot and object images, we show below additional stimulus examples for Reviewer D’s information, which, due to copyright restrictions we cannot include in the manuscript:

Ein Bild, das drinnen, Foto, haltend, darstellend enthält.

Automatisch generierte Beschreibung

In the revised discussion, we write:

P.23, l. 605-614: Despite our best efforts to only include neutral faces, the emotional content of the social stimuli could not be controlled to a fine-grained degree, as it was limited by the design and availability of the robots and objects that were identified through our Google search. In the emotion rating experiment, which we undertook prior to Experiment 1, the robot faces were not rated as unambiguously neutral as the human faces, even after excluding the outliers. Human faces were selected from the neutral category of the Radboud and London faces database, so these stimuli would have contained inherently less variance in perceived emotionality than the robot and object faces. However, given the scarcity of frontally oriented and high-quality robot and object faces, we chose to operate within those constraints.

**D1e: Statistical analyses are sound.**

**D.2: (Figures, Tables, data availability) While sufficiently described, I would like to see a schematic of the Stroop task as it is always helpful to readers. The plots are beautiful.**

Response: In line with this helpful suggestion, we have added a schematic representation of the Stroop task in the Methods section (p.11).

**D.3: (Ethical approval) Ethical approval is present and informed consent is declared.**

**D.4: (Language) English is excellent. As per my comments above, the authors could be more concise when reviewing the literature, as well as describing the methods.**

Response: We hope to have addressed this concern by making the changes we outline in the above responses.

*\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*

*Reviewer H: This manuscript featured two nearly-identical experiments designed to examine the impact of social salience during the Stroop Task. Specifically, the authors varied the social salience of distractors, including images of human faces, robot faces, objects that looked like faces, and flowers, and predicted that faces would amplify the stroop effect due to their high social salience. They found evidence for this in Experiment 1, but when they controlled for a stimulus confound in Experiment 2, they were not able to reject the null hypothesis. This investigation has some strengths. For example, comparing evaluations of human and robot faces is interesting, their analyses were appropriately simple and clear, and basic visual characteristics of the stimuli were well controlled for. The authors also took care to evaluate whether their data provided evidence in favor of the null hypothesis, which I appreciated. In general, I don’t have any concerns about the methods, the data, the analyses, or their interpretation. However, I did have quite a few major concerns about more general issues that tempered my enthusiasm for the work. I’ll explain these in more detail below.*

**H.1a: First, the majority of the Introduction is highly specific to research on gaze perception, as is the Discussion. And yet gaze direction is not examined, nor is it even important, in the current investigation. This reflected a larger issue with the Introduction, which seems to cover many topics before finally focusing on the hypothesis and aims of the investigation. It didn’t feel like the gaps in the literature (as described) necessarily led to the current work and its design. Rather, it felt a bit like a literature review was forced around the current experiment. In this sense, I did not think that the article was as logically structured as it could have been.**

Response: Thank you for this critical reflection – as we have written in response to the editor and the other reviewers, we have reframed both our introduction and discussion to better explain our rationale for designing the conceptual extension of the eye contact effect. We agree wholeheartedly with this criticism. The revised introduction includes a general overview of the social attentional capture focusing on whole faces (rather than gaze perception) and relates this literature back to human-robot interaction research, as this was one of the main motivators for designing this conceptual extension.

For example:

P.4, l. 101-105: Given their prioritization in our visual environment, it is unsurprising that faces have been the central focus of many visual attention studies. Collectively, these studies point towards faces ranking above objects in capturing automatic attention. Using a change blindness paradigm, Ro, Russel and Lavie (2001) found that participants detected changes in temporarily presented faces more quickly than changes in any other object.

And:

P.5, l.149-153: Hence, and as Geiger and Balas (2020) point out, robot faces, which we have presented here as a special case of intentional pareidolia, constitute a border category of face processing, and while some research exists on attentional capture by pareidolic faces, less is known about the social relevance of robot faces.

**H.1b: Second, it’s not clear to me that the design the authors have selected is the best one to examine their main question. In describing previous work by Conty and then Chevallier, the authors state that “the lack of difference in arousal would lead to “centrally presented direct gaze delay[ing] attentional disengagement and recruit[ing] cognitive processing resources, and hence, processing times of the peripheral targets and the Stroop interference are increased”. They then go on to state that testing this claim was the goal of the current investigation. But as far as I can tell, their design does not examine arousal, nor is it confirmed that the stimuli themselves differ in the extent to which they arouse the participants. I don’t have an issue with their stimulus choices per se (they’re rather clever), but they don’t seem to fall out of the literature reviewed, and it seems like not manipulating gaze direction was a missed opportunity. Thus, although I have no issues with the analysis or the data, it’s not clear to me that the conclusions reflect the underlying question, at least as it’s framed in the introduction.**

Response: This comment is very much in line with the comments of the other reviewers and editor, so we decided to reframe the introduction to clarify the rationale for our task, reduced the discussion of the follow-up experiment by Hietanen and colleagues (2016), removed the direct quote, which was criticised by another reviewer as well (see **comment D.1b**), and moved the entire section to the discussion.

In this paragraph, we were trying to establish that another conceptual extension of the eye-contact effect by Hietanen and colleagues failed to show the predicted effect: these researchers reported an effect in the opposite direction (reaction times speeding up) and credit levels of arousal in their experiment with an embodied confederate as an explanation. They describe studies with pictorial stimuli (as our studies, or the studies by Conty, Chevallier and colleagues) as low-arousal situations, in which the original effect should hold. However, we were of course unable to provide convincing evidence for a social salience effect in this version of the Stroop task. While we did not measure arousal directly, we wanted to pick up this point by Hietanen and colleagues, to continue the conversation on why a null effect could be observed in conceptual extensions of this paradigms.

P.20, l. 499-506: Hietanen and colleagues (2016) found a main effect of direct gaze speeding up the RTs of the participants, as compared to averted gaze. The authors reconcile their contradictory findings by relating them to the higher arousal produced by their stimuli: eye contact with a real person should be more engaging than pictorial representations thereof. In so-called low arousal contexts, they argue, salient social cues should recruit attentional resources and interfere with participants’ performance on cognitive tasks. In our experiments, even in a context that Hietanen and colleagues (2016) describe as “low arousal”, it is most probable that any social salience effect is practically equivalent to zero.

Like Chevallier and colleagues (2013), we chose not to manipulate gaze direction in this task, but rather include the neutral flower distractors (just as Chevallier and colleagues did) and vary the levels of socialness of the distractor agents. Based on the findings of Chevallier and colleagues, we expected a human distractor-dependent enhancement of the Stroop effect in the incongruent condition, compared to the flower distractor. Despite the experiments not explicitly investigating eye gaze, the gaze direction of all “social” stimuli (humans, robots, objects) was direct, towards the observer. Thus, despite taking up a smaller region in the distractor image, the direct eye gaze was controlled across social distractors, and any one of these categories should then draw more attention than the flower images. We added this point to our discussion:

P.20, l.507-519: How can our results then be explained? Of course, the stimuli we presented were more complex than those used in the original studies, so it is possible that the eye-contact effect only holds in (more) simplified contexts. The eye region in our stimulus set appeared smaller than in the original experiments, due to it taking up a smaller percentage of pixels in our distractor images. While the eye region itself was smaller, all of our social stimuli (the human, robot and object faces) depicted direct gaze and a frontally oriented face. They only varied in their potential as a social interaction partner. So, if the eye-contact effect were to hold, we should have seen a consistent difference between our most salient social stimuli with direct eye gaze (the human faces) and the neutral control condition (flowers). The fact that our data did not support this pattern is especially surprising given that past studies examining direct gaze have also used full-face stimuli in similar, cognitively demanding tasks (Burton, Bindemann, Langton, Schweinberger, & Jenkins, 2009; Conty, Russo, et al., 2010).

**H.1c: Third, critical information about the task is missing. Yes, the Stroop Task is well known and quite simple, but it isn’t adequately described in the Methods, nor is any background provided about the history of the task or its mechanisms. This wouldn’t be too difficult to rectify, but as it stands, it’s a curious omission.**

Response: We have added detailed information and figures on the number of discarded trials in the main text and supplementary materials (see also Reviewer comment **I.4**) and have added more information on the design of the Stroop task (see also Reviewer comment **I.8**).

Further to Reviewer H’s request, we have added a section discussing the history of the task and its mechanisms in the introduction.

P.7, l. 208-222: Despite the above reviewed variety of paradigms which probe (social) attentional capture, the Stroop task has proven to be a particularly popular vehicle. Named after the psychologist who discovered the effect, hundreds of studies have shown that naming the ink colour of an incongruent colour word (i.e., the word “RED” presented in green) produces slower reaction times than determining the colour of a control word (the letters “XXX” presented in green). This interference effect, which highlights the fact that task-irrelevant information is processed concomitantly and automatically, has inspired a multitude of extensions, including pictorial, spatial, and social versions (MacLeod & MacDonald, 2000). For example, in the facial-emotional Stroop, participants name the ink colour of emotional, compared to neutral faces, which are overlaid with a coloured filter. Past research has shown that sad participants and participants with higher trait anger are slower to name the colour of angry versus neutral faces (Isaac et al., 2012; van Honk, Tuiten, de Haan, vann de Hout, & Stam, 2001; Van Honk et al., 2000). Thus, the Stroop task has been validated as a suitable paradigm to assess the distracting power of task-irrelevant information, such as facial cues.

**H.1d: Finally, I found it hard to process the takeaway message of the manuscript. The authors found evidence against the null hypothesis in Experiment 1, but there were issues with a stimulus confound, and then in Experiment 2, there appears to be no effect of category, but the authors were at the same time not able to support a case in favor of the null--of faces not drawing more attention in the Stroop task. In other words, it’s just really difficult to get a clear sense of what the study demonstrates, and thus what it’s impact will be.**

Response: Following comments from Reviewer I, we have revised the section on the Bayesian re-analysis of the data (including Figure 6). We hope that our interpretation of the results is now clearer: while the ROPE analysis does not offer compelling evidence in support of the null hypothesis, we can quantify our uncertainty. The 95% credible interval of the posterior distribution contains zero and overlaps to ~ 50% with our region of practical equivalence. Thus, if human faces draw more attention in the incongruent condition than the flower distractors, this effect is much smaller than expected and the evidence for it is not very strong. By providing our posteriors, other Bayesians may include them as priors and collect enough evidence to support one decision over the other. Science is cumulative, and Bayesian statistics give us an important advantage of quantifying our uncertainty, which would have not been possible if we stopped at the point of describing the null effect of the Frequentist analysis.

P.19, l.462-468: In summary, in defining our Bayesian regression model, we have increased the uncertainty of our estimates by including more random variance in the form of subject-level random effects. This increased uncertainty is expressed in Figure 5. Based on the ROPE analysis, we cannot definitively support the null hypothesis. However, considering that zero is contained in the 95% interval of credible values of the parameter’s posterior distribution, the evidence for an effect is not very strong, and if real, goes in the opposite direction: -10ms [-10, 40].

**H.2: (Figures, Tables, data availability) The tables and figures are nice. Well done.**

**H.3: (Ethical approval) This seemed adequate.**

**H.4: (Language) In general, yes. The quality of English was good.**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

*Reviewer I: In this preregistered study, the question was investigated whether human faces automatically attract attention more than other types of distractors (human-like faces or non-faces) while participants solve a Stroop task. Two studies are presented, where in study 1 (N=39) a small effect seemed to favour the prediction with slightly increased Stroop effects in the presence of human faces, but a second study (N=51) that increased the number of unique distractor images failed to find differences between response times to the different distractor types.*

*I enjoyed reading this well-written manuscript, the theoretical background is nicely developed, the methods are sound and the statistical analyses are sophisticated. I have a few observations nevertheless that I would like the authors to consider.*

**I.1: The task involved the concurrent presentation of a Stroop colour-word interference test and distractors. While this setup seems to follow methods by Conty et al. (2010), the main measure involves a form of “dual distraction” - distraction from the colour-incongruent words and distraction from the faces. It would have been nice to have baseline trials in which no distractors were shown, in order to evaluate people’s Stroop effect per se, without imposing a second task.**

Response: Indeed, our experiments were designed as a conceptual extension of Conty and colleagues (2010), and we followed the original procedure as closely as possible. Seeing as the Stroop effect is considered robust in the literature, we did not include another control condition without any distractor images to establish this as a ground truth. Given that we find a main effect of target in the pre-registered analysis of both experiments, we can assume that the task itself worked and, overall, induced the desired Stroop interference effect (with some variance between participants, of course).

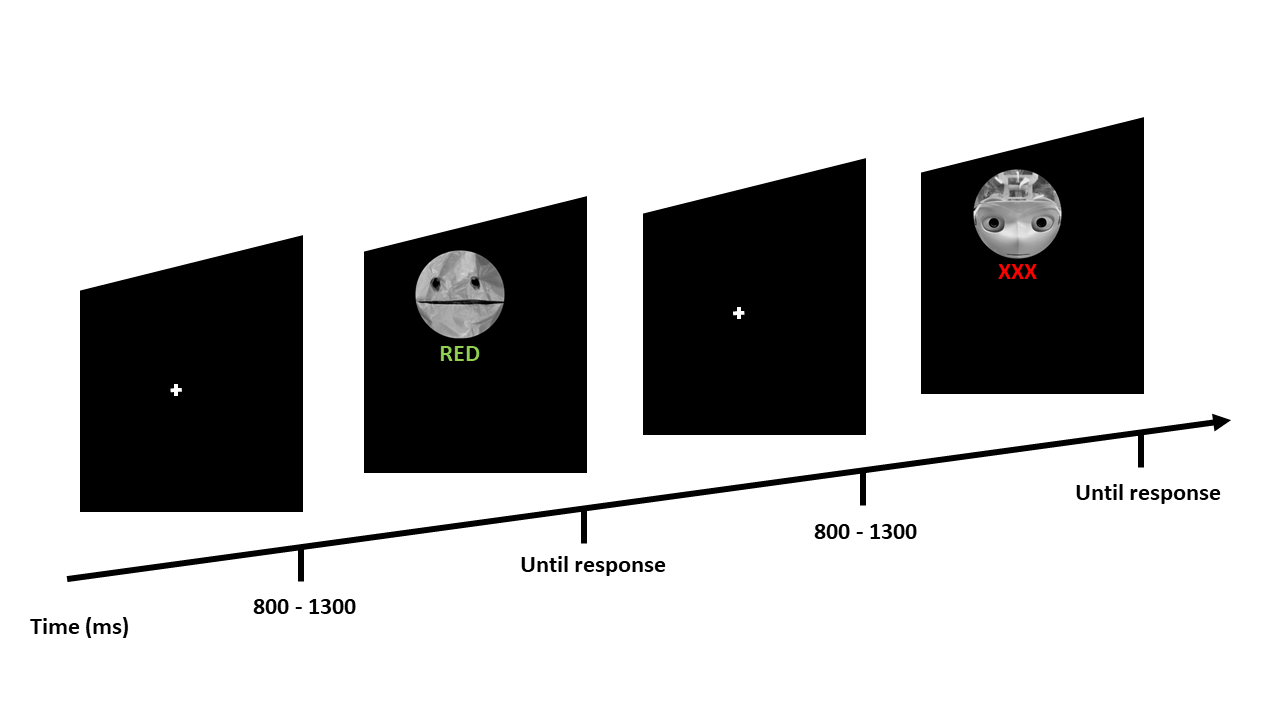
We reemphasized this point (in addition to raising it in the abstract), by including it in the Results sections of Experiments 1 & 2:

P.12, l. 355-356: This finding confirms that our modified task was still effective at inducing a Stroop interference effect.

P.15, l. 421-422: Again, this showed that the task worked as expected.

**I.2: The visual layout of these stimuli on the screen was not entirely clear to me as it is not shown in the figures, although described in the text. Was the distance between the words and the distractors different or the same as in the original study? In other words, was it perhaps easier to ignore the distractors here than in Conty et al., especially given that ignoring the distractors was indeed what participants were asked to do.**

Response: We agree with the Reviewer that a visual representation of the experimental paradigm would be helpful, which we have added to the Methods section and is also visualised below:



The distance between the distractors in our experiments and the original studies was matched as closely as possible given the difference in shape.

In trying to emulate the stimulus size, we faced the following problem: to compute size based on reported visual angle, information on the distance at which the stimulus is viewed is also necessary. This information was missing from the 2010 paper. As a workaround, we referred to the later paper by the same group, which used a similar paradigm (exchanging the averted gaze control condition for flower distractors): Chevallier et al. (2013). This allowed us to calculate the size of the distractor images using the following code in R:

*desiredSize <- function(visAngle, distance){  
 Rad = visAngle/(180/pi)   
 size = 2\*distance\*tan(Rad/2)  
return(size)  
}*

*dist=50  
ang=6  
desiredSize(visAngle = ang, distance = dist)   
5.24*

*(Code taken from:* [*http://stephenrho.github.io/visual-angle.html*](http://stephenrho.github.io/visual-angle.html)*)*

Thus, we can be confident that the target words and distractor images had a comparable size and were at the same distance from each other as in the original studies.

**I.3: The skewed RTs (as nicely shown in Figure 4, for study 2) were analysed in raw format without further transformation (log) – have the authors tried to analyse log-transformed RTs?**

Response: As the Reviewer correctly observed, here we only report the untransformed reaction times, as we did not pre-register any data transformations. However, upon initially inspecting the skew, we did try the log-transform, thus achieving an approximately normal distribution:

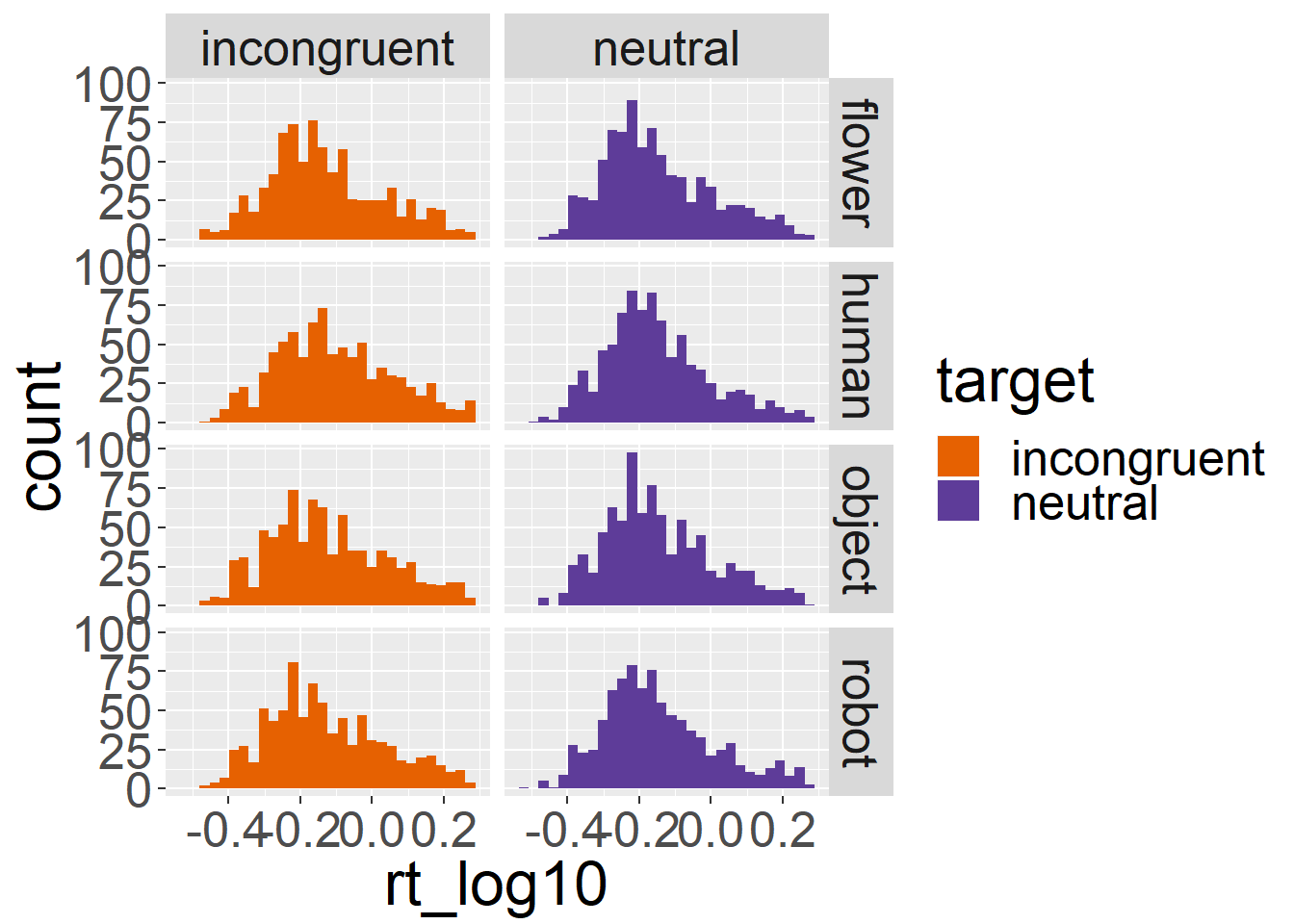


Figure 1) Log-transformed reaction time data of Experiment 1.

The log-transformation did not change the results of either of the two studies, and as a recent preprint questions the usefulness of this convention (Schramm & Rouder, 2019), we decided to focus our exploratory report on the Bayesian re-analysis of Experiment 2. In the Bayesian analysis we fit an exgaussian distribution to the data, which represents the inherent right-skew better (Baayen & Milin, 2010).

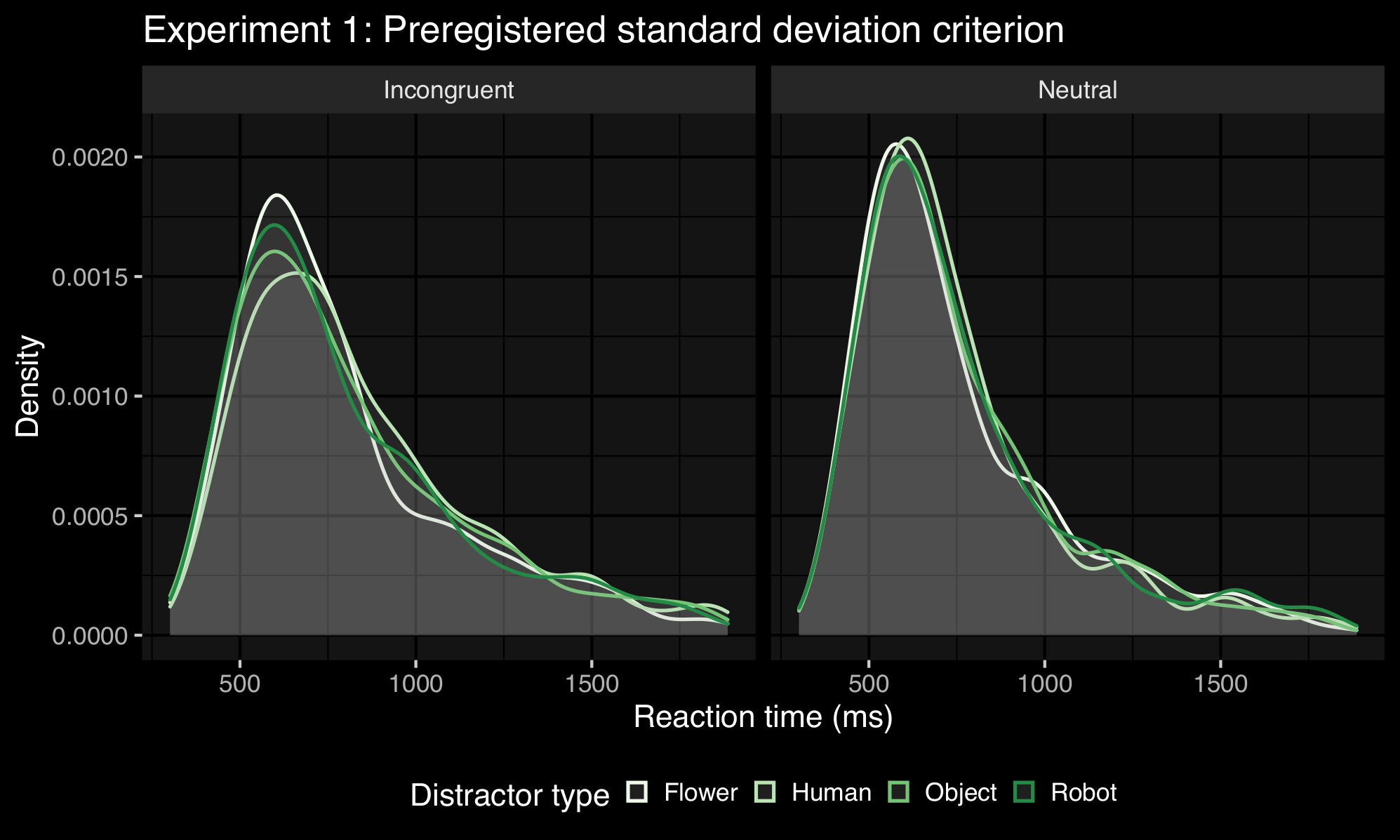
**References**

* Baayen, R. H., & Milin, P. (2010). Analyzing reaction times. *International Journal of Psychological Research*, 3(2), 12-28.
* Schramm, P., & Rouder, J. (2019). Are Reaction Time Transformations Really Beneficial? <https://psyarxiv.com/9ksa6/>.

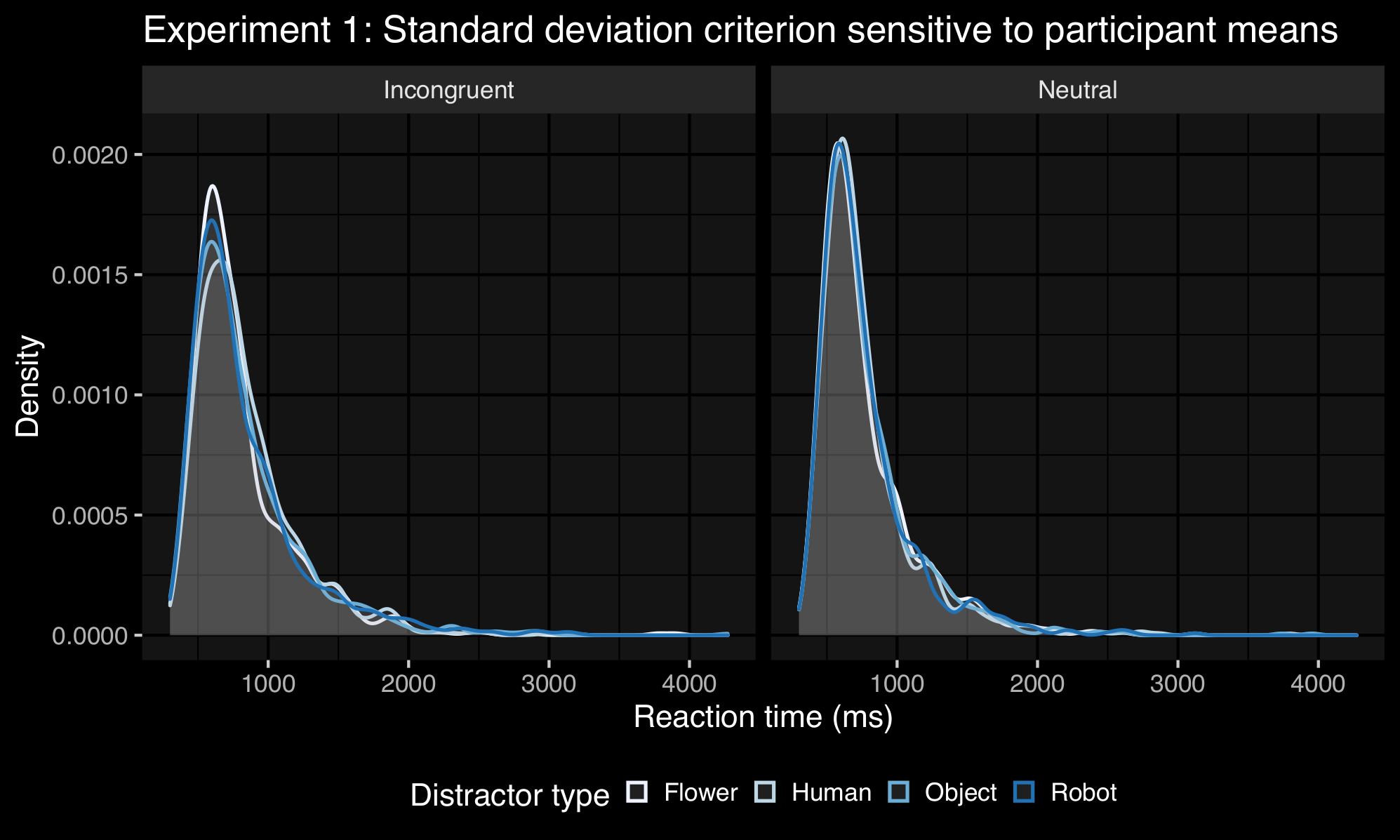
**I.4: Perhaps most critically, it seems that trials in which response times larger than 2 SD above the sample mean were excluded. This is likely too stringent since individual response times are quite variable and in fact, the most interesting trials in this task would be those in which distraction was maximal (i.e., response times are long). In order to avoid “overcleaning”, I would strongly recommend either not to exclude long trials, or to use individual response time distributions - exclude trials that are 2 (or 3) SDs above each participant’s own mean RT instead of the sample mean. Numbers of excluded trials and excluded trials per condition are also not reported and should be added. As a result, the remaining trials could be biased towards those that were not distracting (no matter which condition).**

Response: We agree with the Reviewer that the criterion for pre-processing the reaction time data was perhaps too inflexible and did not accommodate between-participant variability. As outlined in our previous responses, we followed the procedure of the original studies and thus specified in our [pre-registration](https://osf.io/39eqd/): “Outlier data are defined as reaction times below 200 ms and as more than 2 standard deviations above the mean.”

To investigate the concerns of the Reviewer, we used the {trimr} package, which allows the implementation of various response time trimming criteria. We used a standard deviation criterion sensitive to the participants’ own means (8.9% of all trials were removed), as well as the standard deviation criterion we pre-registered (8.09% of all trials were removed). We included two tables side by side that show the means for both methods. Using the suggested standard deviation per participant criterion resulted in overall slower reaction times (Table 2), compared to the original analysis (Table 1).



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Table 1. Mean reaction times (in ms) per condition using the pre-registered standard deviation criterion (Experiment 1).** | | | | |
|  | Distractor | | | |
|  | Human | Robot | Object | Flower |
| Incongruent target | 843 | 807 | 815 | 796 |
| Neutral target | 753 | 768 | 763 | 760 |



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Table 2. Mean reaction times (in ms) per condition using the standard deviation per participant criterion (Experiment 1).** | | | | |
|  | Distractor | | | |
|  | Human | Robot | Object | Flower |
| Incongruent target | 840 | 833 | 833 | 809 |
| Neutral target | 769 | 770 | 780 | 776 |

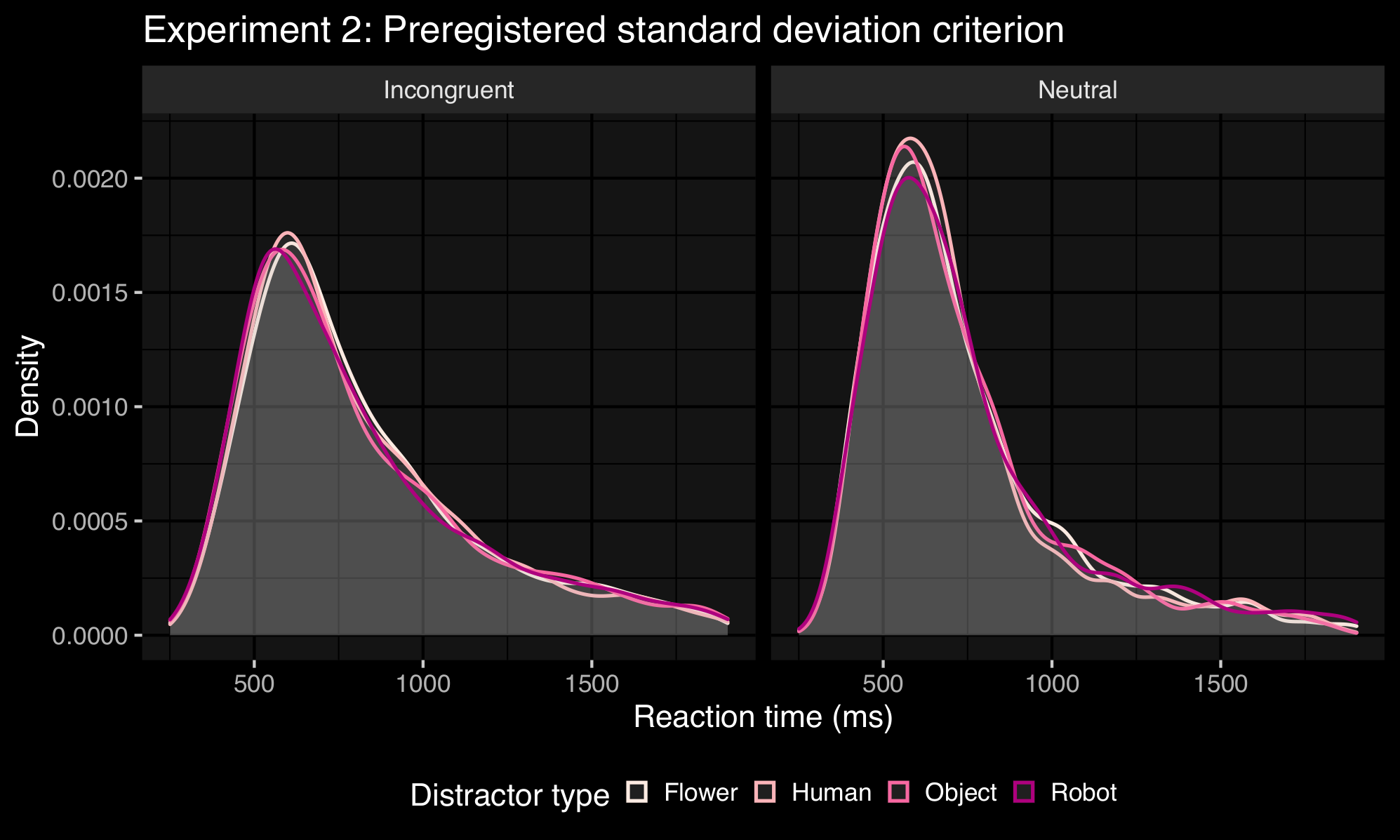
However, the pattern of results nonetheless looks similar to what we originally reported, and in the interest of adhering to our pre-registration, we decided to keep the current results section of Experiment 1 as it is, but add a table on the number of discarded trials in the supplementary materials, as well as add the percentage of the total amount of discarded trials in the main text:

P.12, l. 338-340: As a result, 606 trials (8.09%) were discarded (a detailed breakdown of the trial number per condition can be found in the Supplementary Materials).

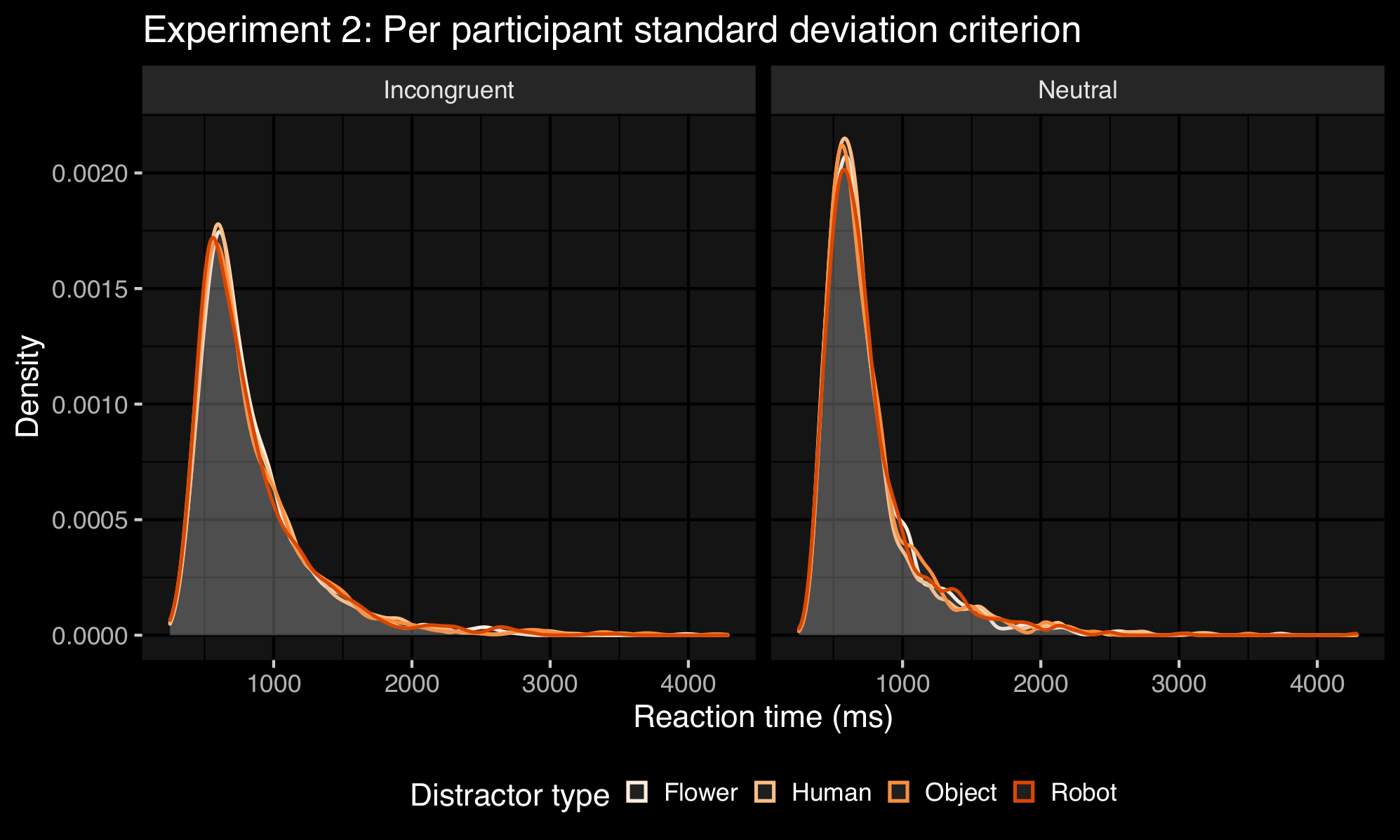
Repeating the same procedure for Experiment 2, we find that with our pre-registered standard deviation criterion we discard 1061 trials (10.84%) in total (with the participant sensitive criterion we discard 11.35% of all trials). We added the information on discarded trials in the main text:

P,15, l. 410-412: With this reaction time trimming criterion, we discarded 1061 trials (10.84%). A detailed breakdown of the number of trials remaining per condition can be found in the Supplementary Materials.

For comparison, we list Table 5 and 6 with the mean reaction times for Experiment 2 using our reported method and the method recommended by the Reviewer:



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Table 5. Mean reaction times (in ms) per condition using the pre-registered standard deviation criterion (Experiment 2).** | | | | |
|  | Distractor | | | |
|  | Human | Robot | Object | Flower |
| Incongruent target | 811 | 808 | 809 | 816 |
| Neutral target | 723 | 747 | 730 | 735 |



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Table 6. Mean reaction times (in ms) per condition using the standard deviation per participant criterion (Experiment 2).** | | | | |
|  | Distractor | | | |
|  | Human | Robot | Object | Flower |
| Incongruent target | 833 | 828 | 834 | 825 |
| Neutral target | 748 | 755 | 750 | 749 |

Again, we see that the overall patterns of results remain the same across these two pre-processing methods.

**References**

Grange, J. A. (2015). trimr: An implementation of common response time trimming methods. R package version 1.0. 1

**I.5: The Bayesian analyses are sufficiently esoteric for me that I require more clarification here.**

Response: We agree with the Reviewer that the report of the exploratory Bayesian analysis in its original form was not as clear as it could (or should) have been. We hope we have addressed this concern sufficiently in the following responses.

**I.5a: Page 13 states “Given the results of Study 2, we explored the extent to which our data provided compelling evidence for the null hypothesis (no difference in reaction times in the incongruent and neutral conditions when human faces are presented)“. This implies to me that the null hypothesis would predict no Stroop effect when the human faces were presented. I believe this is not what the authors meant, but instead that the size of the Stroop effect would not differ between distractor conditions. Is this the case? If so, this needs to be changed in the text.**

Response: We thank the reviewer for pointing out this mistake. We have amended the text accordingly:

P.16, l. 428-433: Given the results of Experiment 2, we explored the extent to which our data provided compelling evidence for the null hypothesis (no enhanced Stroop effect when human faces are presented compared to the control flower condition) by using a Bayesian regression modelling approach ({brms} package in R and Stan (Version 2.9.0), Bürkner, 2017), as the null cannot be confirmed with Frequentist statistics.

**I.5b: How are R-hat values of 1.00 for each of the tested parameters in Table 3 to be understood?**

Response: The R-hat value provides information on how well the algorithm could estimate the posterior distribution of each parameter. Since we already provided information in the main text on the convergence of the model, this column has been removed to avoid redundancy.

**I.5c: The ROPE outcomes do not support the presence of a Stroop effect at all, if I understood this correctly. The size of the general Stroop effect was sufficiently large, in both studies, based on the conventional outcomes (study 1: F(1, 38)= 39.24, p<.001, ηG2= .03; study 2: F(1, 50)=70.31, p<.001, ηG2=.06). Can the authors comment more directly on this discrepancy? And if the outcomes of the Bayesian analysis are taken seriously, what are the consequences for the rest of the paper? For example, page 15 in the discussion states “While we again observed the predicted Stroop effect” – did you? The different outcomes need to be reconciled better, in my opinion.**

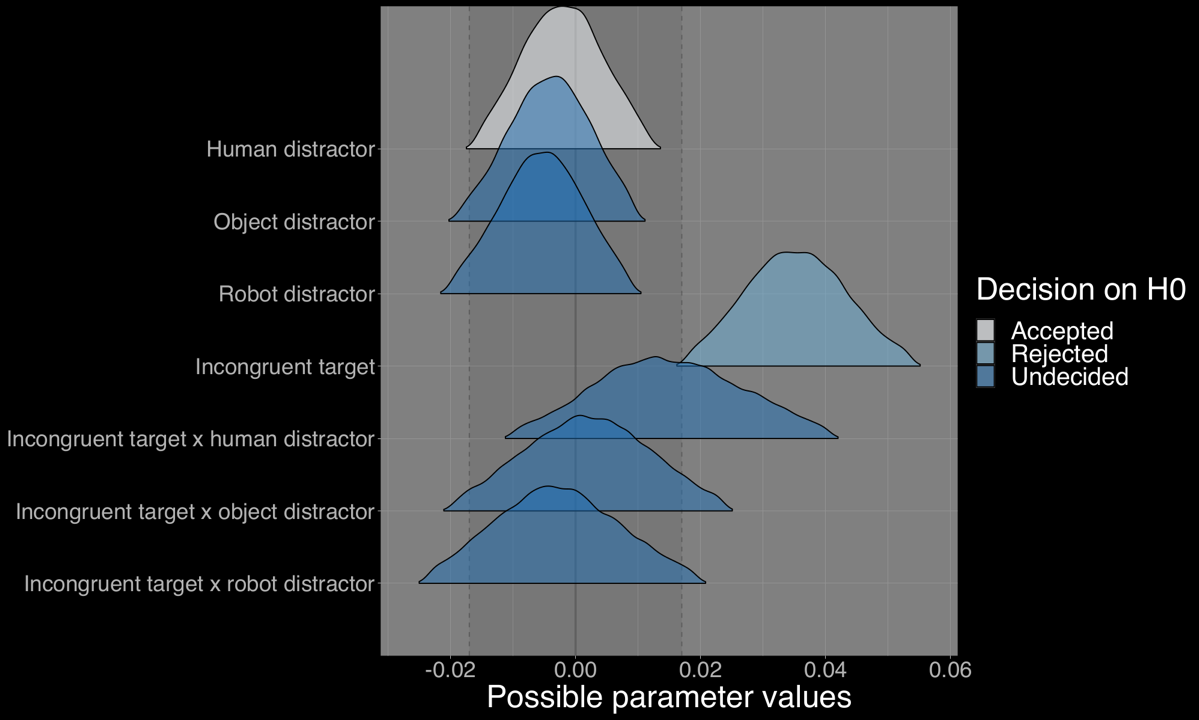
Response: We thank the Reviewer for raising this issue. Determining the limits of the ROPE is a somewhat controversial issue in the literature (Kruschke, 2018; Kelter, 2020). Instead of relying on the automated procedure we opted for initially via the {BayesfactoR} package’s rope\_range() function, in the revised manuscripts, we choose the limits based on half of what is considered a small effect. The rope\_range() function returned a range that is considered a large effect and because of this, the robust Stroop effect was classified as “undecided”. Based on Experiment 1 (∆ 47ms) and the findings of Conty et al. (∆ 34ms) and Chevallier et al. (∆ 41ms), we set the ROPE limits to [-.017, .017] and observe a plot that is easier to reconcile with the Frequentist analysis.

P.18, l. 451-457: In determining the ROPE range, we set the limits following the procedure based on half of what we consider a small effect (Kruschke, 2018). A small effect in our first experiment was an average difference of 47ms between the incongruent social and incongruent control distractor, compared to a difference in 34ms in Conty and colleagues’ task and 41ms in Chevallier and colleagues’ version (2010, 2013). Choosing the most conservative small effect, we set the ROPE limits to [-.017, .017].

**I5.d: Figure 5 is unclear to me. What is zero on the x-axis – this can’t be “reaction time (s)” ? Also, going back to point 4 a) is this testing the presence/absence of any Stroop effect or the slow-down of RTs (i.e., bigger Stroop effect) for human faces compared to the other conditions?**

Response: We included an updated version of Figure 5, now with a corrected x-axis label. Thank you for pointing out this issue. The graph depicts all parameters estimated in the Bayesian regression model. So, for instance, the estimated effect of the incongruent target on the reaction time (the Stroop interference effect), for which H0 is rejected. We also see the parameter estimates for the different distractor types, as well as the interactions. The effect of interest (as outlined in the figure description in the manuscript) is the incongruent target with the human distractor type. This effect is now shaded in yellow, and we do not have a clear decision on H0 based on the ROPE analysis. However, as we have written in our updated Bayesian analysis section, the estimated effect is small and likely not very strong (if present at all). Moreover, it is smaller than the smallest effect we consider interesting (based on our previous experiment and the literature), and in the 95% CI zero is contained as a likely value.

P.19, l. 462-468: In summary, in defining our Bayesian regression model, we have increased the uncertainty of our estimates by including more random variance in the form of subject-level random effects. This increased uncertainty is expressed in Figure 5. Based on the ROPE analysis, we cannot definitively support the null hypothesis. However, considering that zero is contained in the 95% interval of credible values of the parameter’s posterior distribution, and more than 50% of its values are practically equivalent with zero, the evidence for an effect is not very strong and even goes in the opposite direction: 10ms [-.01, .04].



**I.6: Using mirror-images also in study 2 arguably may not have created unique distractors. A mirror image could act as a particularly strong distractor, as it would appear familiar but not identical. This could be considered in the limitations section.**

Response: We agree that this point should be raised in the limitations section and have added it:

P.23, l. 593-604: A further stimulus-based limitation was that in Experiment 1, distractors were not controlled by their mirror and presented twice. Thus, the repeat presentation could have led to a particularly memorable stimulus set. In Experiment 2, the unique distractors in the incongruent condition were controlled by their mirror images. Of course, on the other hand, the repeat presentation of distractor images is common practice in the social attentional capture literature (for example, a set of four unique human and pareidolic face images used for an experiment consisting of 450 trials, Ariga & Arihara, 2017). Takahashi and colleagues (2013) used stimuli with three unique identities over many trials, and only four unique stimuli in another study (Takahashi & Watanabe, 2015). Theeuwes et al. (2006) presented 12 unique distractor images across 96 trials. To put it differently, based on the conventions of the social attentional biasing literature, it is unlikely that we did not observe the expected effect due to the number of unique distractor images we presented.

**I.7: The decision to move the stimulus rating into a supplement abbreviated the rating outcomes presented in the paper, but I would still have liked to see some details. In fact, the supplement also does not state what exactly was being judged regarding these stimuli. The paper states on page 11, “mind perception of different kinds of agents” – what does this mean and what was the actual outcome of the ratings? Is it relevant or irrelevant for this paper?**

Response: We have tried to clarify our rationale for the ratings of our distractor images in Experiment 2. As we have written, we wanted to establish that the 4 different categories were perceived differently with regard to “having a mind”, which we implicitly equated with the agent’s potential for socialness. The two items (and their descriptions), which we called “agency” (ability to plan and act) and “experience” (the ability to sense and feel), were derived from Gray, Gray & Wegner (2007).

P.14, l. 402-407: We used mind perception as a socialness proxy to distinguish between the control condition (flowers), inanimate (robot and pareidolic faces) and agents with a mind (humans). The analysis of the ratings confirmed that the stimulus categories were perceived differently: the human images received the highest agency and experience ratings. A detailed report of the stimulus ratings can be found in the Supplementary Materials.

**I.8: Some missing details on the Stroop task itself included the number/ratio of congruent and incongruent trials, and any restrictions regarding the switch between the two (e.g., no more than 2 incongruent trials after each other etc.)**

Response: We have added this information in the methods section:

P.11, l. 327-329: There were equal numbers of incongruent and neutral Stroop trials, and no restrictions regarding the switch between incongruent and neutral trials were put in place (as they were presented randomly). The target word and distractor image pairs were fixed.

**I.9: Since several participants were excluded, I wonder whether these criteria were too stringent. At least the method of excluding participants should be detailed. For example, excluding participants with ASD diagnoses – how was this done?**

Response: We have added this information.

The rationale for specifying these exclusion criteria was that Chevallier and colleagues (2013) found diverging results for the ASD participant group in their sample, and we wanted to ensure that all participants were equally naïve towards robots (as the initial goal was to establish this as a robust measure for social motivation, and then in future experiment integrate this task following prolonged human-robot interaction. We were curious about seeing any potential differences between a robot-naïve group of participants and a group that has encountered humanoid robots on this task).

P.9, l. 284-289: We recruited 50 participants, however, based on our pre-registered exclusion criteria (diagnosis of ASD and having had a previous interaction with a robot) we excluded 9 participants. Two additional participants had insufficient English language skills, and thus the total number of exclusions was 11. The pre-registered exclusions were made based on participant answers on the experiment questionnaires’ self-report items (for example: “Do you have a diagnosis of Autism Spectrum Disorder?” and “Have you interacted with a robot before?”). The other exclusions had to be made in addition, based on difficulties participants had with the task. We report a final sample size of N=39.

**I.10: (Figures, tables, data availability) Very nice use and high quality of Figures.**

**I.11: (Ethical approval) Ethical approval was obtained from the University of Glasgow ethics review board (300170224).**

**I.12: (Language) English is appropriate.**

**Final Editor decision**

Nov 13, 2020-12-20

Dear Miss Anna Henschel,  
  
I am happy to accept your submission, "Faces do not attract more attention than non-social distractors in the Stroop task", for publication at Collabra: Psychology, pending the completion of copyediting and formatting processes.  
  
As there are no further reviewer revisions to make, you do not have to complete any tasks at this point. The accepted submission will now undergo final copyediting. You will be contacted once this is complete to answer any queries that may have arisen during copyediting and to allow a final chance to edit the files prior to typesetting.  
  
Kind regards,  
  
Dr Christopher R Madan  
University of Nottingham  
christopher.madan@nottingham.ac.uk