Reviewer 1

We thank Reviewer 1 for their support of the manuscript.

Reviewer 2

We thank Reviewer 2 for their detailed comments and suggestions. They raised a number of important points, each of which we address below.

MAJOR COMMENTS

1. **The HMM method maps every gaze position to a list of pre-defined objects (hidden states in the HMM). The reliance on this pre-defined set of objects is potentially problematic when dealing with gaze data that is not directed to any of the defined objects. Possible examples might be gaze towards empty space in a display (similar to the "looking at nothing phenomenon") or gaze towards the centroid of multiple tracked objects. Although the authors mention the ability to add the centroid to the pre-defined set, this example nicely illustrates the issue with mapping each and every gaze position to a pre-defined set: It is always possible that there are target objects one is not aware of and that are thus not included in the pre-defined set. And this might then result in biased results. For example, increasing object speed increases the bias of gaze towards the centroid as compared with targets (Huff et al., 2010). Without taking the centroid into account, this shift in gaze could not be captured.**

   We agree with the reviewer - in some cases, the assumption that the participant is tracking exactly one of the prescribed objects may be too strong and this could cause hard-to-diagnose biases. We note, however, that this source of bias is present even when humans hand-code the data according to a pre-defined protocol. Hence, this difficulty is inherent not to the proposed HMM method, but rather to the problem of analyzing continuous eye-tracking data, which captures more complex and diverse behaviors than can be explicitly modeled.

   Further in the present manuscript, when asking human raters to manually map gaze data in Experiment 2 as a reference for the HMM method, their human coders were given the possibility to use the "Off Task" classification to indicate that gaze was not directed to any of the objects of the pre-defined object set. Their human raters used this possibility indicating that some gaze was difficult to map. Even though the authors discuss that adding an "Off Task" classification to their current HMM approach is not so straight forward and thus left for future work, I highly recommend the authors to deal with the described issue in one way or another. Currently, users will have no indicator of whether they are operating on interpretable or nonsense mappings. A rather straight forward approach (even without extending the current HMM method) might be to add some kind of quality indicator based on the distance between gaze and mappings resulting from HMM (or something the like).

   We note that, although only a minority of frames are classified as "Off Task" ($\approx 10\%$ by both coders, and $\approx 20\%$ by at least 1 coder), the majority ($\approx 77\%$) of frames on which coders disagree are those on which one, but not both, of the coders used the "Off Task" classification (this is illustrated by the gap between the black and white lines in Figure 5). Given this, it is not clear that adding an "Off Task" state to the HMM, even if practically feasible, is the best solution, since even human coders may not agree on whether a given frame is "Off Task".

   On the other hand, the ability to precisely quantify how well the data fit the model is an advantage of an automated procedure such as the HMM procedure, since asking humans to quantify their confidence in each classification could significantly increase the time and effort of hand-coding.

   Given these factors, we feel that the reviewer's suggestion to include a quality indicator is perhaps the best way to deal with this issue. We have added a new section (Section 5.3, pages 19-20) to the paper describing such a quality indicator, based on the computing the "trial log-likelihood" (TLL) of the gaze given the object sequence predicted by the HMM. A low value corresponds to data that was unlikely to be generated from that object sequence, alerting the user of a potential failure of the modeling assumptions.

   To provide a basic validation of this TLL statistic, we investigated how it correlates with human classification "Off Task" classifications; we found a moderate, statistically significant correlation of -0.36, across all experimental trials.

2. **The introduction should be extended to include existing approaches of dealing with eye-movements in dynamic scenes. I am aware of at least two approaches that are missing (both also published in BRM): Papenmeier & Huff (2010); Friedrich, Wußwinkel & Möhlenbrink (2017).**

   Thank you for bringing these highly-relevant papers (as well as other papers on multiple object tracking) to our attention. We have added this discussion under "Related Work" (Section 1.3, page 6) in the paper.

3. **The authors state that their approach is suitable for analyzing smooth-pursuit eye-movements. For example, they state: "Because smooth pursuit is intimately tied to tracking smoothly moving objects, this model effectively provides a way of analyzing smooth-pursuit movement in many contexts". This is an over-generalization of the capabilities of their method. Their method is not able to classify or separate smooth pursuit eye-movements from fixations and saccades (at least to my understanding). Thus, their method cannot be used to analyze smooth-pursuit eye movements per se. Rather, their method uses raw gaze data instead of pre-classified fixation/saccade data, thus also working in contexts that also include (but not only!) smooth-pursuit eye-movements. That is, their method does not rely on a pre-classification and, thus, does not care whether eye-movements are mapped to moving objects due to there being smooth-pursuit eye-movements or a succession of short fixations and saccades.**

   Thank you for pointing out this over-generalization and suggesting a better phrasing. We have revised the relevant sentence (on page 3) to more accurately reflect how our method addresses the problem of analyzing eye-tracking data that may contain smooth pursuit movements as follows:

   > "Because our method uses raw gaze data instead of pre-classified fixation/saccade data, our method works in contexts that include either or both of smooth-pursuit and fixation/saccade eye-movements, bypassing the difficult problem of identifying smooth pursuit movements."

4. **How should users of their method determine the optimal value of σ with their own data? The experiments demonstrate that the HMM approach works particularly good with some values of σ in their paradigm. If applying to a new paradigm, how should we determine σ?**

   Thank you for raising this important question. In addition to the original brief sentence about this at the end of Section 2, we have now clarified and elaborated on this in a new paragraph (on page 22) as follows:

   > "We reiterate that at present, we do not have a general, automatic method for calibrating the tuning parameter σ in the HMM. σ depends on both the physical properties (e.g., display size and resolution, viewing distance, object speed) of the experimental setup and characteristics of the participant (e.g., age). Practical solutions include considering results over a range of σ values or calibrating σ, either by having human coders manually code a small subset of data from the task being studied or by directly estimating the variance of the participant's gaze data when tracking an object (e.g., using a calibration experiment consisting of TrackIt with no distractor objects). Statistical approaches, such maximum likelihood, may also be applicable. When in doubt, both intuition and our empirical results suggest that erring on the side of using a smaller σ value will minimize potential bias introduced by the HMM model, while still outperforming the SDM."

5. **In the General Discussion they discuss the application of their method to natural scenes. When dealing with natural scenes, we certainly also deal with tracking not just a single object but also with tracking multiple objects simultaneously, like in the MOT paradigm (see Meyerhoff et al., 2017 for a review on the task; also see Hyönä, et al., 2019 for a recent review on eye-movements in MOT and MIT). Thus, this discussion should also include some consideration on how their HMM approach deals with eye-movements in situations where observers track multiple objects simultaneously.**

   Thank you for this insightful point and for the helpful references. We have added some discussion (to Section 6.2, page 23) on how the HMM might be adapted to handle multiple object/identity tracking.

FURTHER COMMENTS

A. **Use APA style for citations (year should be separated with comma instead of putting it into another parentheses).**

   Thank you for pointing out this formatting error; it has been corrected.

B. **p.4 add a reference when first mentioning "TrackIt" (similar to how it is done on p.7) so readers know where it comes from.**

   Thank you for the suggestion; we have moved the reference on p. 4 to be immediately after the first mention of TrackIt to make it easier to find.

C. **p.8: 9 shapes x 9 colors = 81 objects -> So it is possible that multiple objects in the display share either the same color or shape, correct? I am just asking because the figures suggest that there are unique colors and shapes across objects in the display (that is neither a color nor a shape is repeated across objects).**

Thank you for pointing this out – the original text was actually incorrect, and the reviewer is correct in guessing that neither shapes nor colors are repeated within a trial. We have corrected this (now on page 9) with the following:

> "For each trial, the target and distractor objects are constructed with random colors (selected without replacement from a set of 9 distinct colors) and shapes (selected without replacement from a set of 9 distinct shapes); that is, out of 81 possible objects, target and distractor objects are selected randomly under the constraint that no color or shape is repeated within a trial."

D. **p.8: There was a minimum trial duration of 10 seconds in each trial? For Multiple Object Tracking experiments, identities are known to not only support tracking by establishing correspondence (Papenmeier et al., 2014) but also by allowing participants to use their memory when tracking objects with identities (Makovski & Jiang, 2009). With a minimum trial duration of 10 seconds, participants could look at the display initially, then not track any objects for 10 seconds, recover the target based on its identity and then track the trial until trial completion. Did this happen, and if so, what are the influence for the results (particularly of Experiment 2)?**

Indeed, it is possible that participants could stop tracking any objects in the middle of the trial and still correctly identify the ending grid cell of the target. A few aspects of the experiment should mitigate this to some extent:

(a) Participants were explicitly instructed to "follow the target object with your eyes".

(b) Participants were not informed of the 10 second minimum trial duration.

(c) While the minimum trial length was 10 seconds, the exact length of the trial was variable, typically between 10-20 seconds, making it more difficult for participants to estimate the minimum trial duration and thereby anticipate trial end, within the span of only 10 trials.

The effect of this behavior on the results on Experiment 1 might be quite significant, perhaps partially accounting for the relatively low measured performance in children; this was one of the motivations for performing the additional validation Experiment 2. In Experiment 2, this behavior should not have an appreciable effect on the results, since frames that coders classified as "Off Task" were omitted when with comparing model (HMM or Naive) predictions.

Finally, we would like to note that we have some work in progress in which we explicitly manipulate trial duration to measure its effect on tracking behavior in children. However, these data are not yet available.

E. **It is great that the authors make the data and scripts available online. While the data and scripts of Experiment 2 are well structured, the corresponding data of Experiment 1 is hard to use. I strongly recommend re-structuring the files of Experiment 1 similar to Experiment 2. As an example: The manuscript mentions on p.10 that a Python script used to collect eye-movements synchronized with their TrackIt task is available at Github (they provide a link). Unfortunately, I could not find it there, either because it is missing or due to the unstructured way the scripts and data is deposited at Github. While cleaning up the repositories, please consider to make available everything on repositories that are expected to be available for a long time, such as OSF. I am particular concerned about the data of Experiment 2 because I experienced too many times that data made available on websites of institutes disappeared once the corresponding researcher moved to another university.**

We apologize for the confusing organization of the Experiment 1 supporting files. As the reviewer suggested, we have reorganized the files from Experiment 1 in a similar way to Experiment 2, and we have uploaded both to OSF (`https://osf.io/u8jbs/`). When appropriate, the URLs within the text now point to individual files, rather than the entire project directory, to make it easy to find the files described.

The original submission had separate paragraphs on "Source Code and Reproducibility" for each experiment. Since the supporting files for both experiments are now available in the same location, we have removed these separate paragraphs and added a short section to the introduction (Section 1.5, page 7).

F. **p. 10: "10 consecutive frames ($\approx 16.7\,\mathrm{ms}$)" -> 10 frames correspond to $167\,\mathrm{ms}$, not $16.7\,\mathrm{ms}$.**

We have corrected this typo, thank you.

G. **p.11/12: paragraph on leave-one-out cross-validation is hard to comprehend. Please elaborate (and make clear whether Table 2 presents the corresponding data, also in the title of Table 2).**

We have added a few sentences of clarification to this paragraph (now on p. 12) and to the caption of Table 2, elaborating on the cross-validation procedure and making explicit the connection to Table 2.

H. **Data of Experiment 2 was averaged across conditions that differ quite a lot regarding their visual properties. This does not feel right to me.**

Thank you for pointing this out. We initially felt that the condition difference was largely orthogonal to the main contrast (HMM versus Naive) being studied in the paper, and that presenting results for both conditions would complicate the presentation of the results.

To justify this decision, we have re-run the analyses separately under each condition. Figure 1 (below) shows the main plots of the results, side-by-side for the two conditions. Although there are small quantitative differences, we do not observe major qualitative differences, and we feel that these differences do not affect the conclusions of the paper. By most measures, both the HMM and Naive models perform slightly (but not significantly) better in the Exogenous condition than in the Endogenous condition (as expected, since participant task compliance is typically better in the Exogenous condition, which is less demanding). We have added a few sentences to this effect (on page 14).

Other Changes

In addition to the changes suggested by reviewers, we made two minor modifications of the manuscript:

1. Throughout the text, we have replaced the name "Naive model" with "shortest-distance model (SDM)". This is in reference to a paper (Zelinsky and Neider (2008); now cited in the main paper), which suggested using a shortest-distance rule (that is equivalent to our Naive model) for eye-tracking data analysis in dynamic settings. We have added a short paragraph under "Related Work" (on page 6) referencing this paper.

2. We have added (on pages 2-3 and page 8) references to a few papers (Ross et al., 1993; Katsanis et al., 1998; Luna et al., 2008) that studied the development of oculomotor control in children. These papers suggest that oculomotor control in children during smooth pursuit can be quite a bit noisier than in adults. This helps motivate the use of a "soft" (Gaussian) model for gaze behavior when tracking an object and also helps elucidate the role of the σ parameter, suggesting why its optimal value differs between children and adults.

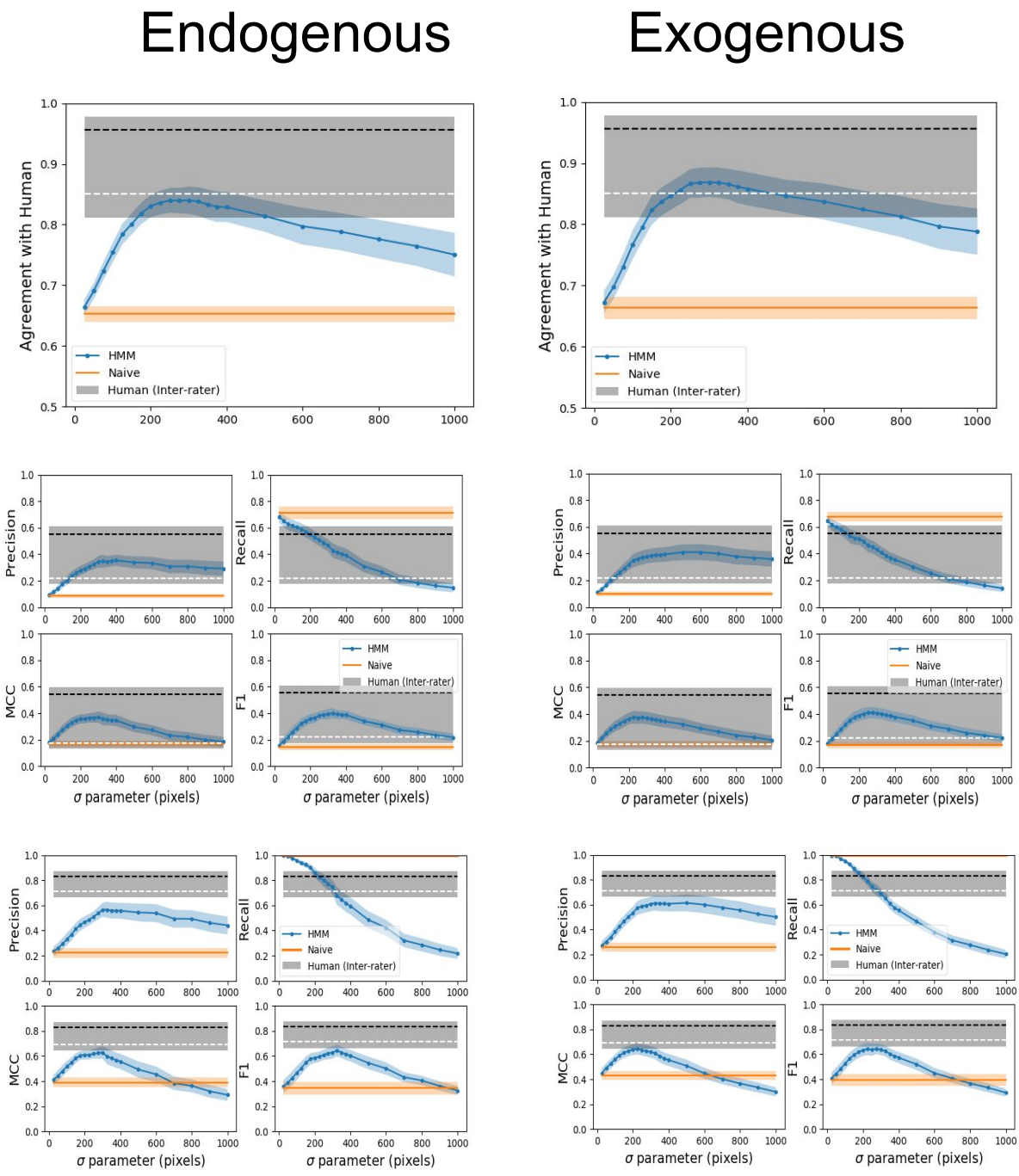

Figure 1: Versions of Figure 5 (Row 1), Figure 6 (Row 2), and Figure 7 (Row 3) from the original submission, separated by TrackIt condition (Endogenous/Exogenous).