Detailed Responses to Reviewer Comments

**Editorial comments:**

Changes to be made by the Author(s):

1. Please take this opportunity to thoroughly proofread the manuscript to ensure that there are no spelling or grammar issues.

**Response**: We have thoroughly proofread the manuscript for both spelling and grammar.

2. Figure 5c: Please change “sec” to “s”.  
 **Response**: “Sec” in Figure 5 has been changed to “s”.

3. Please rephrase the Introduction to include a clear statement of the overall goal of this method.  
 **Response**: The introduction has been modified to include the following statement of the overall goal: “The goal of this method is to capture both the natural visual environments of infants, and infants’ active visual exploration of those environments, as infants move freely. Such data can help to answer questions not only about visual attention, but also about a broad range of perceptual, cognitive, and social developments4–8.”

4. Please revise the protocol text to avoid the use of any personal pronouns (e.g., "we", "you", "our" etc.).  
 **Response**: The one instance of “our” in “our laboratory” has been revised to read “a laboratory”. No other use of “we”, “you”, or “our” was found.

5. Please revise the protocol so that all text in the protocol section is written in the imperative tense as if telling someone how to do the technique (e.g., “Do this,” “Ensure that,” etc.). The actions should be described in the imperative tense in complete sentences wherever possible. Avoid usage of phrases such as “could be,” “should be,” and “would be” throughout the Protocol. Any text that cannot be written in the imperative tense may be added as a “Note.” However, notes should be concise and used sparingly.  
 **Response**: The protocol has been revised to only include statements written in the imperative tense or in a “Note”.

6. The Protocol should contain only action items that direct the reader to do something. Please move the discussion about the protocol to the Discussion.  
 **Response**: We have moved portions of sections 1.2, 3.1, 3.2.1, 3.3.2.1, 3.3.2.3, 3.3.2.3.1, and 4.1 of the original manuscript to the Discussion section of the revised manuscript.

7. There is a 2.75 page limit for filmable content. Please highlight 2.75 pages or less of the Protocol (including headings and spacing) that identifies the essential steps of the protocol for the video, i.e., the steps that should be visualized to tell the most cohesive story of the Protocol. Remember that non-highlighted Protocol steps will remain in the manuscript, and therefore will still be available to the reader.  
 **Response**: Approximately 2.5 pages of the Protocol, including headings and spacing, is highlighted for filming.

8. Please ensure that the highlighted steps form a cohesive narrative with a logical flow from one highlighted step to the next. Please highlight complete sentences (not parts of sentences). Please ensure that the highlighted part of the step includes at least one action that is written in imperative tense. Please do not highlight any steps describing euthanasia.  
 **Response**: The highlighted portions of the revised manuscript are complete sentences that include at least one action written in imperative tense and that form a cohesive narrative. There are no steps describing euthanasia.

9. Please include all relevant details that are required to perform the step in the highlighting. For example: If step 2.5 is highlighted for filming and the details of how to perform the step are given in steps 2.5.1 and 2.5.2, then the sub-steps where the details are provided must be highlighted.  
 **Response**: Relevant sub-steps of the protocol are now highlighted.

10. Discussion: Please also discuss critical steps within the protocol.  
 **Response**: As noted in our response to comment 6 above, discussion about the protocol has been moved to the Discussion. Moreover, we now include in the Discussion notes on the most critical steps in the protocol, particularly several aspects of calibration.

11. References: Please do not abbreviate journal titles. Please include volume and issue numbers for all references.  
 **Response**: The references now include the full title of all journals, as well as volume and issue numbers.

**Reviewers' comments:**  
  
**Reviewer #1:**  
Manuscript Summary:  
This manuscript outlines a protocol for conducting real-world studies with infants and toddlers and recording their eye movements via head-mounted eye trackers. These methods are of great interest to the developmental science community and the authors of this protocol have been pioneering in this work so I am certain this protocol will be well received. However, the success of their work somewhat overshadows some major complexities to gathering good data with this target developmental population which I believe must be emphasised in a rewrite before submission.  
  
Major Concerns:

-There should be some comments/caveats on data quality. Just because you get a good eye image and the system calibrates does not mean the data is good enough quality for your intended analysis. See Holmqvist et al (2011) and Saez De Urabain, Johnson and Smith (2014).  
 **Response**: This has been added to the Discussion as follows: “Finally, keep in mind that even when a good eye image is obtained and the system calibrates, this does not ensure that the data is of sufficient quality for the intended analyses. Differences in individual factors such as eye physiology, as well as environmental factors such as lighting and differences in eye-tracking hardware and software can all influence data quality and have the potential to create offsets or inaccuracies in the data. 18,19 provide more information and possible solutions for such issues (see also 20).”

-It should be highlighted more that manual ROI coding requires a lot of time and resources as automated tools are unreliable. Authors should include a time estimate for cleaning each minute of raw gaze data and for hand-coding ROIs. Also include links to possible automation techniques if they believe these may be possible.  
 **Response**: Time estimates depend strongly upon how the coding is done. We now include the following statements in the discussion: “Manually coding data frame-by-frame for ROI can require extra time compared to coding fixations. As a reference, it took highly trained coders between 5 and 10 minutes to manually code for ROI each minute of the data presented here, which was collected at 30 frames per second. The time required for coding is highly variable and depends on the quality of the eye tracking data; the size, number, and visual discriminability of ROI targets; the experience of the coder; and the annotation tool used.”

We are not aware of any reliable automation techniques and thus have not included any in the manuscript. Tools like the neural network YOLO (Redmon, Divvala, Girshick, & Farhadi, 2016) hold promise for automated object detection, but the output of such networks are bounding boxes, not object segmentation at the pixel level, and as mentioned in the Protocol we do not recommend trusting the crosshairs at the pixel level for various reasons, for instance there can be overlapping objects in the scene view or the pupil can be incorrectly located in the eye image.

-Lines 575-576: The authors point out that 70% of already collected data is usable/of high-quality, but given the focus on methods, the discussion should also point out the actual drop-out rate (i.e. how much usable/high-quality data is obtained of all participants who took part in the study, including those toddlers who did not start the study due to, e.g., fussiness as they did not tolerate the cap/ET gear). This should be noted as this may affect labs with limited resources or participant pools. Head-mounted eye tracking is specifically aversive to many young children whose resistance may not even allow the study to begin.  
 **Response**: We now realize that it was unclear to what the estimate of “70%” referred. We have revised this section to clarify that this number refers to the percentage to participants from whom we get sufficient high-quality data, as well as to note the causes for the failure to obtain good data from the remaining 30% of infants. Specifically, the revised manuscript reads: “Employing the recommendations included in this protocol, designed for use with infants from approximately 9-24 months of age, a laboratory can obtain high-quality head-mounted eye-tracking data from approximately 70% of participants18. The other 30% of participants may either not begin the study due to intolerance of the eye tracker, or fuss out of the study before sufficient data (e.g., >3-5 minutes of play) with a good eye track can be obtained. For the successful 70% of infant and toddler participants, these sessions typically last for upwards of 10 minutes…”

Minor Concerns:

Lines 34, 82, 546: The authors mention "infants". This protocol focuses on toddlers, and although the protocol is similar for infants, there are some additional challenges with infant testing that were not discussed in the manuscript, so the protocol should either mention this or focus on toddlers/young children only.  
 **Response**: We have maintained use of both of the terms “infant” and “toddler” because this protocol is designed for use with children aged 9-24 months, which we now state explicitly in the revised manuscript. We find that within this age range, there are large differences in attention and motor ability between individuals even of the same age, often larger even than mean differences between individuals of different ages, hence we have not specified particular ages for which the various methodological challenges might apply. Nevertheless, in addition to noting that the age of the participants may affect the length of the session, we have also added the following statements to the revised manuscript: “…these sessions typically last for upwards of 10 minutes, however much longer sessions may be infeasible with current technologies, depending on the age of the participant and the nature of the task in which the participant is engaged. When designing the research task and environment, researchers should keep in mind the developmental status of the participants, as motor ability, cognitive ability, and social development including sense of security around strangers, can all influence participants’ attention span and ability to perform the intended task. Employing this protocol with infants much younger than 9 months will also involve additional practical challenges such as propping up infants that cannot yet sit on their own, as well as consideration of eye morphology and physiology, such as binocular disparity, which differ from that of older children and adults19,21.

Lines 93-95, Eye-tracking equipment: To my knowledge, there is only one commercially available system for young children, namely Positive Science. The images in this manuscript suggest the authors have modified the Pupil Labs system to work with a custom made infant cap but this should be made clear so the readers do not leave with the false impression that such systems can be commonly bought off the shelf. This section and the following sections suggest that options are available on the market.  
 **Response**: The revised manuscript now clarifies this issue in several places. Protocol section 1.1 now states: “Select one of the several head-mounted eye-tracking systems that are commercially available, either one marketed as specifically for children or modify the system to work with a custom-made infant cap, for instance as shown in Figures 1 and 2.” More detail on how to do this is now provided in protocol section 1.1.3.1: “Embed the system into a cap by attaching the scene and eye cameras to a Velcro strap that is affixed to the opposite side of Velcro sewn onto the cap, and positioning the cameras out of the center of the toddler’s view.” Protocol section 2.2.2 has also been revised to read: “2.2.2 At the study, have different types of caps available to which the ET can be attached. Customize caps by purchasing different sizes and styles of caps, such as a ball cap that can be worn backward or a beanie with animal ears, and adding Velcro to which the eye-tracking system, fitted with the opposite side of the Velcro, can be attached. Also consider having hats to be worn by the caregiver and experimenters, to encourage the child’s interest and willingness to also wear a cap.”

Lines 112/208-209: How do you put the eye camera out of a toddler's view? It will always be in the view and is a major challenge for HMET testing.  
 **Response**: These sections have been revised to specify that the cameras should not be *in the center of* the child’s view.

Line 115: "more normal" Bit awkward.  
 **Response**: This sentence has been revised to read: “Also consider having hats to be worn by the caregiver and experimenters, to encourage the child’s interest and willingness to also wear a cap.”

Point 2.3.1.1 - You should also mention the benefits of binocular eye tracking if the gaze is expected to shift in depth during the testing protocol. Monocular tracking will create an offset on the world camera relative to the actually fixated point if monocular gaze is used as the actual point is a consequence of vergence between the two eyes.  
 **Response**: The revised manuscript includes the following note after Protocol section 2.4.2.1, including references for more information on binocular eye tracking should the reader be interested: “Note: Binocular eye tracking is a developing technology13,14 that promises advances in tracking gaze in depth.”

Calibration section, general comments:

Parts of the calibration section could be restructured in their order. Some parts are repetitive. The calibration section should be divided into what to do/consider during the study when getting points, and what to do/consider when calibrating in the software.  
 **Response**: We have chosen to organize the Protocol sequentially rather than topically and have therefore left the section on what to do during the study to obtain calibration points as part of Protocol section 2, as this step is part of the data collection process, rather than combining it with section 3 on performing the calibration in the software. We have, however, added the italicized words in the section titles pasted below to make it more clear when each of these steps take place: section 2.3 is now titled “Obtain Points *During the Study* for Offline Calibration” and section 3 is titled “Calibrate the ET Data *using Calibration Software*” (italics not used in the manuscript). We have also revised and/or deleted portions of both Protocol sections 2.3 and 3 to reduce redundancy and reordered portions of section 3 for better flow.

Example for repetition: Section 3.2.1 (line 279) is repeated in 2.4 (lines 252); Section 3.3.2.1. mentions accurate pupil tracking as a requirement when previous sections already mention it (e.g. 3.3.1.).  
 **Response**: These Protocol sections have all been revised to reduce redundancy. Note, however, that we have kept these sections separate rather than combining them because, despite referencing similar information, they instruct the reader to do different things at different stages. For instance, section 2.5 (2.4 in the original manuscript) instructs the reader to take note during the study of the timing of any changes in the ET’s position, whereas section 3.1.1 (3.2.1 in the original manuscript) instructs the reader to later create separate calibrations for those sections of the recordings. To better integrate these sections, we have added references between the sections, for instance we have added “(see Section 3.1.1)“ to Protocol section 2.5.

Example for structure: Section 3.2.2. (line 288) could go in the previous section for "obtaining points for offline calibration".  
 **Response**: As noted in our response to the previous comment by the Reviewer, we have opted not to combine these sections because, despite referencing similar information, they instruct the reader to do different things at different stages.

Calibration section, specific comments:

Line 232, Calibration: It could be recommended to use a few more points than needed for offline calibration (if this is feasible for the study). This ensures that if a calibration point turns out unusable later, an alternative one can be used. Also, it should be noted not to go too fast between points (e.g. the laser should not be immediately moved to the next location once the researcher sees the child focused on the point). This makes it easier to detect the moment of fixation later, when calibrating.  
 **Response**: Both of these recommendations have been added to Protocol section 2.4.1, which now reads: “2.4.1. Alternate between drawing attention to different locations that require large angular displacements of the eye. Cover the field of view equally and do not move too quickly between points, which will aid in finding clear saccades from the child during offline calibration to help to infer when they looked to the next location. If the child does not immediately look to the new highlighted location, get their attention to the location by wiggling the laser, turning off/on the LEDs, or touching the location with a finger. If feasible, obtain more calibration points than needed in case some turn out to be unusable later.

Line 249, adding calibration points: If feasible, calibration points could be added during the recording if the session is longer.  
 **Response**: We have added this to Protocol section 2.4.3, which now reads: “2.4.3 To accommodate for drift or movement of the ET during the study, collect calibration points at both the beginning and end of the study at minimum. If feasible, collect additional calibration points at regular intervals during the session.”

Line 302: "use pupil only" - this can, however, often lead to inaccurate tracking and occasionally result in displacements  
 **Response**: In cases where the pupil and corneal reflection are both reliably and consistently detected, using both leads to higher accuracy compared with using the pupil only. Nevertheless, we have found in practice it is much easier to position the eye camera to detect the pupil, and much more difficult to detect both the pupil and corneal reflection consistently when we run a study for much longer than 5 minutes, a relatively long period of time for infants/toddlers. Thus, we have kept this point in the manuscript, but qualified it with the phrase in parenthesis as follows: “3.2.1 Assist the calibration software in detecting the pupil and CR in each frame of the eye camera video to ensure that the identified POG is reliable. In cases where the software cannot detect the CR reliably and consistently, use the pupil only (note, however, that data quality will suffer as a result).“

Line 360, assessing quality: "indicate known gaze positions, such as the dots produced by a laser pointer during calibration" - however, if calibration points themselves are used for calibration, then these points cannot be also used for checking data quality, unless post-hoc calibration points are provided.  
 **Response**: This is a great point, which we have now added to Protocol section 3.3 as follows: “3.3 Assess the quality of calibration by observing how well the POG corresponds to known gaze locations, such as the dots produced by a laser pointer during calibration, and reflects the direction and magnitude of the child’s saccades. Avoid using points to assess calibration quality that were also used as points during the calibration process.”

ROI coding

Lines 390-392: Consider rewriting for clarity.  
 **Response**: This section has been heavily revised for clarity and moved to the discussion. We have also added an example to illustrate our point, as follows: “Head-mounted eye tracking can also pose the additional challenge of relatively more time-consuming data coding. This is because, for the purpose of finding ROIs, head-mounted eye-tracking data is better coded frame by frame than by “fixations” of visual attention. That is, fixations are typically identified when the rate of change in the frame-by-frame x-y POG coordinates is low, taken as an indication that the eyes are stable on a point. However, because the scene view from a head-mounted eye tracker moves with the participant’s head and body movements, the eye’s position can only be accurately mapped to a physical location being foveated by considering how the eyes are moving *relative to* *head and body movements.* For instance, if a participant moves their head and eyes together, rather than their eyes only, the x-y POG coordinates within the scene can remain unchanged even while a participant scans a room or tracks a moving object.Thus, “fixations” of visual attention cannot be easily and accurately determined from only the POG data…”

Line 423, Use the POG track as a guide, not as ground-truth: This section should come before the previous section so that the previous section is more readily understandable in terms of why there is a need to use the pupil as information (since the crosshair is available).  
 **Response**: This section now comes before the section on using pupil movement, as recommended.

Point 4.1.- This point is important but will only be understood if "fixation" is properly defined and the issues associated with classifying it in mobile data are expanded. Please do so with reference to Holmqvist et al (2011) and Smith & Saez De Urabain (2017)

**Response**: We have moved this information to the Discussion section to allow for elaboration, including a definition of fixations and an example to illustrate the complexity of the issue. This section is pasted below for your reference. We also refer the reader to the citations the Reviewer mentioned, but have chosen not to go into more detail in the manuscript because our focus is on coding ROI rather than fixations.

Paragraph 5 of the Discussion now reads: “Head-mounted eye tracking can also pose the additional challenge of relatively more time-consuming data coding. This is because, for the purpose of finding ROIs, head-mounted eye-tracking data is better coded frame by frame than by “fixations” of visual attention. That is, fixations are typically identified when the rate of change in the frame-by-frame x-y POG coordinates is low, taken as an indication that the eyes are stable on a point. However, because the scene view from a head-mounted eye tracker moves with the participant’s head and body movements, the eye’s position can only be accurately mapped to a physical location being foveated by considering how the eyes are moving *relative to* *head and body movements.* For instance, if a participant moves their head and eyes together, rather than their eyes only, the x-y POG coordinates within the scene can remain unchanged even while a participant scans a room or tracks a moving object.Thus, “fixations” of visual attention cannot be easily and accurately determined from only the POG data. For further information on issues associated with identifying fixations in head-mounted eye tracking data, please consult 15,22…”

Additionally, because fixation detection is so difficult in head-mounted eye tracking and we do not want readers to get caught up in the issue of fixations versus saccades, we have also removed Protocol section 4.2.2.2, pasted here, as it mentions saccades and is not critical to the protocol: “4.2.2.2 Code POG during saccades – the several frames it may take for the eye to shift from visually attending one object to attending another object – as a separate ROI category rather than as the ROI the POG track may happen to indicate on that particular frame. Use contextual information from the preceding and following frames to identify these transitional eye movements.”

New references:  
Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & Van de Weijer, J. (2011). Eye tracking: A comprehensive guide to methods and measures. OUP Oxford.  
Saez de Urabain, I.R. and Johnson, Mark H. and Smith, Tim J. (2014) GraFIX: a semiautomatic approach for parsing low- and high-quality eye-tracking data.Behavior Research Methods 47 (1), pp. 53-72. ISSN 1554-3528.  
Smith, Tim J. and Saez de Urabain, I.R. (2017) Eye tracking. In: Hopkins, B. and Geangu, E. and Linkenauger, S. (eds.) Cambridge Encyclopedia of Child Development. Cambridge, UK: Cambridge University Press, pp. 97-101. ISBN 9781107103412.  
 **Response**: Thank you for providing these full references.

**Reviewer #2:**  
Manuscript Summary:  
This protocol provides guiding principles and practical recommendations for researchers using head-mounted trackers in the laboratory and in naturalistic settings. It also includes the kind of fine-grained data generated from this type of eye-tracking equipment in free-play.  
  
Major Concerns:

Introduction could be improved in documenting research showing how head-mounted eye tracking could provide unique information about infant cognitive, language and motor development. It is unclear what UNIQUE information the ET system brings to infant cognition. In other words, is there any evidence that infants who pay more attention to a certain object learn the word for it earlier? Categorize these kind of objects earlier? **Response**: We have revised the final paragraph of the introduction to note some of the unique contributions of head-mounted eye tracking. Specifically, this paragraph now reads: “…The goal of this method is to capture both the natural visual environments of infants and infants’ active visual exploration of those environments as infants move freely. Such data can help to answer questions not only about visual attention, but also about a broad range of perceptual, cognitive, and social developments4–8. The use of these techniques has transformed understandings of joint attention7-9, sustained attention10, changing visual experiences with age and motor development4,6,11, and the role of visual experiences in word learning12. The present paper provides guiding principles and practical recommendations for carrying out head-mounted eye-tracking experiments with infants and toddlers and illustrates the types of data that can be generated from head-mounted eye tracking in one natural context for toddlers: free-flowing toy play with a parent.”

We also describe in the Discussion how head-mounted eye tracking provides greater opportunity for capturing individual differences in children’s visual attention and has yielded new insights into infants’ looking to parents’ faces, compared to screen-based eye tracking research. That section is pasted here for your reference: “…[head-mounted eye tracking] increases the opportunity for participants to exhibit individual differences in looking behavior, because participants have control not only over where and for how long they focus their visual attention in a scene, as in screen-based eye tracking, but also over the composition of those scenes through their eye, head, and body movements and physical manipulation of elements in the environment. The two participants’ data presented here demonstrate individual differences in how long toddlers look and what objects toddlers sample when they are able to actively create and explore their visual environment. Additionally, the data presented here, as well as other research employing this protocol, suggest that in naturalistic toy play with their parents, toddlers look to their parent’s face much less than suggested by previous research4,5,7-10.”

Fig 5 shows some data on 2 different children. How old were the children? Were was the parent located? The low LT at the face will be directly affected by these variables. There is huge variability in the number of faces infants see during the first few months of life, for example.  
 **Response**: The children were both approximately 18 months of age, which is now noted in the results section of the manuscript. We agree that the position of the parent relative to the child will affect amount of looking to the parent’s face and for that reason parents were instructed to play freely and were not told to stay in a particular position relative to their child. Child 1’s parent chose to sit at a 90 degree angle to the right and in front of the child. Child 2’s parent chose to sit to the left of the child, turned at a 45 degree angle from the child.

Not quite clear how Fig 5A is a stream of data points.  
 **Response**: This is now clarified in the Legend for Figure 5, which reads: “(A)Sample ROI streams for Child 1 and Child 2 during 60 s of the interaction. Each colored block in the streams represents continuous frames in which the child looked at an ROI for either a specific toy or the parent’s face…”

Discussion: "this method is an alternative.." to what? Babies also have control over where and how long they focus their visual attention in traditional free play sessions. They typically code how long they manually explore objects.  
 **Response:** The revised sentence now reads “Due to these advantages, this method provides an alternative to screen-based looking time and eye-tracking methods for studying development across domains such as visual attention, social attention, and perceptual-motor integration, and complements and occasionally challenges the inferences researchers can draw using more traditional experimental methods.”

Minor Concerns:

No mention of Fig 2DEF in text.  
 **Response**: Figure 2C-F is now referenced in Protocol section 2.3.2.1.

Please clarify what you mean by cap (no visor?) vs bandana.  
 **Response**: The first time we mention a “cap” in Protocol section 1.1 we now direct the reader to Figures 1 and 2 for a visual example of what we are referring to. We have also revised Protocol section 2.2.2, which gives other examples of what we mean by “cap”: “Customize caps by purchasing different sizes and styles of caps, such as a ball cap that can be worn backward or a beanie with animal ears, and adding Velcro to which the eye-tracking system, fitted with the opposite side of the Velcro, can be attached.”

Any data on tolerance of the system as a function of age? The pupil camera must be distracting starting at a certain age.  
 **Response**: Please see our Response to a similar comment from Reviewer 1 above, which we also paste here for easy reference: “We have maintained use of both of the terms “infant” and “toddler” because this protocol is designed for use with children aged 9-24 months, which we now state explicitly in the revised manuscript. We find that within this age range, there are large differences in attention and motor ability between individuals even of the same age, often larger even than mean differences between individuals of different ages, hence we have not specified particular ages for which the various methodological challenges might apply. Nevertheless, in addition to noting that the age of the participants may affect the length of the session, we have also added the following statements to the revised manuscript: “…these sessions typically last for upwards of 10 minutes, however much longer sessions may be infeasible with current technologies, depending on the age of the participant and the nature of the task in which the participant is engaged. When designing the research task and environment, researchers should keep in mind the developmental status of the participants, as motor ability, cognitive ability, and social development including sense of security around strangers, can all influence participants’ attention span and ability to perform the intended task. Employing this protocol with infants much younger than 9 months will also involve additional practical challenges such as propping up infants that cannot yet sit on their own, as well as consideration of eye morphology and physiology, such as binocular disparity, which differ from that of older children and adults19,21.”

How long should the parent have the child wear a cap before testing?

**Response**: As stated in Protocol section 2.2.1, the goal of having the child wear a cap before testing is to “get them accustomed to having something on their head”. We intentionally do not specify a particular duration of time for this as children will take different amounts of exposure to a cap to become comfortable with it, and because this step will often be constrained by how far in advance an appointment can be scheduled. Moreover, leaving the recommendation open to the parent’s discretion helps to ensure that the parent does not feel burdened by this recommendation.