PEER REVIEW HISTORY

This paper was submitted to a another journal from BMJ but declined for publication following peer review. The authors addressed the reviewers' comments and submitted the revised paper to BMJ Open. The paper was subsequently accepted for publication at BMJ Open.

(This paper received three reviews from its previous journal but only two reviewers agreed to published their review.)

ARTICLE DETAILS

TITLE (PROVISIONAL)	Improving Sensitivity to Eye Gaze Cues in Autism Using Serious Game Technology: Study Protocol for a Phase I Randomized Controlled Trial
AUTHORS	Scherf, K. Suzanne; Griffin, Jason W; Judy, Brian; Whyte, Elisabeth M; Geier, Charles F; Elbich, Daniel; Smyth, Joshua M

VERSION 1 – REVIEW

REVIEWER	Liz Pellicano Macquarie University, Australia
REVIEW RETURNED	03-Jun-2018

GENERAL COMMENTS	This manuscript reports a study protocol for a planned phase I randomised controlled trial designed to improve autistic adolescents' sensitivity to eye-gaze using an adaptive serious game. The manuscript is well written and well described – and I love the fact that the intervention is specifically designed for adolescents, unlike the majority of interventions and support programmes that focus on early childhood. I do, however, have some comments that the authors might wish to consider – particularly with regard to clarifying various aspects of the rationale and methods. These are listed below. I have one main concern that I note here, that there is no qualitative element included in this preoiminary trial. Including qualitative methods during and particularly post-intervention with the young people themselves and their parents could be enormously informative. Indeed, qualitative methods – ideally, interviews, but at the very least a questionnaire with open-ended questions – can help to (1) determine whether the intervention was delivered as intended, (2) understand adolescents and their parents' experiences of the intervention (including probing for potential harms – see below), and (3) unpack processes of implementation and change. They can also help to generate further questions or hypotheses. I realise that this trial is already planned but is there scope of including such methods following the post-intervention assessment? 1. p. 5, para 2. The authors state that no existing interventions

	designed to improve face processing have used eye-tracking measures as outcome behaviours. It would be helpful for readers if the authors could explain *why* such measures are superior to more standard measures. 2. p. 7, aims and objectives. The intervention (game) is to be played for 30 minutes three times per week for two months. Could the authors please provide a rationale for the dose (30 minutes, three times per week) and duration of the intervention (two months)? On what basis were these decisions made? 3. p. 8, para 3. One of the inclusion criteria is that the adolescent should have normal vision and hearing with correction. I assume this will be, as reported by parents rather than tested by the researchers? It would be helpful to specify this. 4. p. 10, para 2. Who is conducting the randomisation procedures? 5. p. 10, para 3. Will the same researcher be involved in the pre- and post-intervention assessments in an effort to minimise measurement error? 6. p. 12, para 2. It is commendable that the researchers are measuring feasibility – but feasibility should not be limited to the intervention itself, it should also include other features of the trial. For example, feasibility can include procedures for (1) estimating the standard deviation of the outcome measure, which is needed to estimate adequate sample size for a full trial and (2) determining the follow-up rates, response rates to questionnaires and adherence/compliance rates. I encourage the researchers to also record such details to inform a larger trial (should this preliminary trial yield positive effects). 7. p. 13, para 3. It is great that the researchers are recording adverse events and unintended effects (i.e., harms) during the course of the intervention. But the recording procedures are lacking in detail in the protocol description and seem rather ad-hoc to me. Do the researchers have any particular potential harms in mind (perhaps that they have gleaned from pilot work)? Again, qualitative interviews at follow-up would be very useful in this regard. 8. p. 13, para 2. The dependent variables of interest for the visual attention to faces task include the average gaze time to faces and the proportion of total gaze time to faces. It's unclear to me why the region of interest is the entire face, rather than the eye region, given that increased sensitivity to eye-gaze direction is the proposed active ingredient in this study. 9. p. 13, para 3. I am intrigued about the development of the eye-gaze sensitivity task. From the description, it seems that it would be rather easy for participants, especially for cognitively able autistic adolescents. Have the authors piloted this task with their target population to preclude the possibility of ceiling effects on this task? Some background on task development and assurances that the tasks is developmentally appropriate and sufficiently sensitive to capture change would be beneficial here. 10. p. 14, para 2. The researchers are proposing to administer the Social Skills Improvement System and the Social Responsiveness Scale. Could the authors please clarify their rationale to include both of these measures, which have overlapping content. In particular, I am concerned about the use of the SRS and whether it will be sufficiently sensitive to change over a 2 or 3-month period – especially given that the SRS asks parents/teachers to report on the young person's behaviour *in the past 6 months* (and will thus overlap with both the pre- and post-intervention time-points). 11. p. 14, para 3. While I appreciate the need to reimburse participants to take part in the intervention, I'm not sure that paying participants $5 for every 30 minutes of game play up to $200 will

	really allow the researchers to examine feasibility, i.e., whether the intervention is acceptable in the absence of a (rather lucrative) reward. 12. p. 16, para 1. The researchers state that they will analyse the groups (intervention vs. waitlist control) for potential differences in demographic characteristics. It is not clear, however, why they will do this (as far as I can tell, there are no stopping rules for the trial) or how they will avoid potential biases here and ensure that allocation remains concealed until post-intervention assessments are complete.

REVIEWER	Dr Jane McCarthy King's College London UK
REVIEW RETURNED	03-Jun-2018

GENERAL COMMENTS	Well thought out preliminary RCT. The main limitation is discussed in terms of the blindness of the researchers collecting the post intervention data.

REVIEWER	Dr. Pavel A. Orlov Imperial College London, UK
REVIEW RETURNED	07-Jun-2018

GENERAL COMMENTS	The paper "Improving Sensitivity to Eye Gaze Cues in Autism Using Serious Game Technology: Study Protocol for a Phase I Randomized Controlled Trial" provides a description of the future study of an effect of intervention game usage that designed to train individuals with ASD. Authors provide general description of the game and mostly focus on the study design. The paper consists of 7 sections including Introduction and Discussion sections. At the Introduction, section authors point out that Autism spectrum disorder (ASD) "characterized by impairments in social communicative behaviors. Core symptoms of these impairments are deficits in social looking behaviors including limited visual attention to faces and sensitivity to eye gaze cues." They briefly provide a background review of the underlying mechanism of these deficits and propose an intervention strategy based on "serious game" approach. And finally, authors end up the Introduction section with Aims & Objectives "The aims of this study are to assess the feasibility and safety of this serious game intervention and examine the initial evidence for its effectiveness to alter sensitivity to eye-gaze and social visual attention to faces in adolescents with ASD". At the Methods section, authors show study design with settings, describe participants, sample size and recruitment process with randomization procedure. Authors present the core principals of the intervention game design in the section Intervention Conducted in the Experimental Group. Then they discuss outcome measures, data collection and management processes. Finally they provide a short discussion. The paper has 1 figure and supplementary forms. The paper appears to be helpful for conduction future studies with ASD subjects. Authors fully describe the procedure, data management and carefully show aspects of the study. Here I provide my comments about issues that were found: 1. I did not found the dates of the future study. 2. Line 33-35 (p 7). It is good to clarify what "evidence-based game mechanics" do you mean there. 3. Line 18 (p 8). Please give the definition of RCT acronym. 4. The main issue with the paper I see the lack of game description.

	I suppose that some pictures and game logic or scenarios (via UML diagrams) will be helpful to see the potential of the intervention. At the line 3 (p 12) you have “joint attention episodes when two people engage in mutual gaze with each other and then engage in joint attention on the same specific object”, do you mean two avatars or two people? I guess that some screen picture will be helpful. The same with Line 18-26 (p 12). 5. Line 3-4 (p 13) you have “Second, in the early stages of the game, participants are provided with multiple kinds of non-verbal social cues in their interactions with the avatars” and then you have a few examples of such cues. Do you have a full list? And why did you select that one? 6. Line 16 (p 16). How did you get 70-90 minutes for the eye-tracking experiments? Take into account issues that were found I suggest a minor revision of the paper. At this stage, the paper is good in general.

VERSION 1 – AUTHOR RESPONSE

Reviewer 1

I have one main concern that I note here, that there is no qualitative element included in this preoiminary trial. Including qualitative methods during and particularly post-intervention with the young people themselves and their parents could be enormously informative. Indeed, qualitative methods – ideally, interviews, but at the very least a questionnaire with open-ended questions – can help to (1) determine whether the intervention was delivered as intended, (2) understand adolescents and their parents' experiences of the intervention (including probing for potential harms – see below), and (3) unpack processes of implementation and change. They can also help to generate further questions or hypotheses. I realise that this trial is already planned but is there scope of including such methods following the post-intervention assessment?

Thank you for the recommendation. We agree that greater input from individuals enrolled in the intervention would be potentially valuable. Although we are not able (e.g., due to grant funding parameters) to implement a comprehensive qualitative element, we have now included a Usability questionnaire as part of the safety outcomes that will allow us to query participants about their intervention game experience. Participants will be able to rate multiple aspects of game play on a Likert scale. For example, they will respond to statements like, Experience was fun or I felt discouraged, using a scale with five possible ratings (strongly disagree, disagree, neutral, agree, strongly agree). We used a similar questionnaire to elicit feedback from adolescents with autism during development of the game. Given reports in the literature about limitations in metacognitive awareness in individuals with autism (e.g., Grainger et al., 2014, 2016) particularly in the domain of face processing (McHahon et al., 2015; Sawyer et al., 2014), we think an open-ended interview in which we ask participants to reflect on their experience might not generate the same kind of concrete, usable information as this structured, scaled questionnaire.

The Usability questionnaire is now described in the Safety outcomes section of OUTCOME MEASURES on p. 13.

Regarding the feasibility and “dose” of the intervention delivery, we will actually be collecting objective use data. Data from the intervention game will be uploaded from each participant's computer every 8 minutes of each training session (see Data Collection - p. 16). Log files will provide details about each

behavior executed in the game, time spent in each task, and total time spent playing the game (see Feasibility outcomes). These data will allow us to begin to determine whether the intervention is delivered as intended and also help unpack processes of implementation and change. Additional information will also be captured from the adolescents via the Usability questionnaire.

Additional Comments:
1. p. 5, para 2. The authors state that no existing interventions designed to improve face processing have used eye-tracking measures as outcome behaviours. It would be helpful for readers if the authors could explain *why* such measures are superior to more standard measures.

This is an important point. In the paragraph the Reviewer is referring to, we note several potential explanations for the lack of success of existing computer-based interventions to influence long-term changes in face-processing behavior, one of which was the lack of measurement of social visual attention itself via eye tracking metrics. Reduced visual attention to faces is one of the earliest indicators of autism in infancy, persists across the lifespan, and 3 is a reliable predictor of social impairments in ASD. Importantly, deficits in face identity recognition and emotion expression categorization, which are the other two behaviors often targeted in interventions of face processing in ASD, are not diagnostic features of ASD. Therefore, given that atypical social visual attention to faces is a diagnostic feature of ASD, it may be a critical dependent measure for assessing the success of interventions targeting face processing in ASD.

To clarify this issue, we now include a statement in the manuscript that, “none of the existing interventions evaluated changes in social visual attention, which is a diagnostic feature of ASD, using eye-tracking measures as outcome behaviors.” (p. 5)

2. p. 7, aims and objectives. The intervention (game) is to be played for 30 minutes three times per week for two months. Could the authors please provide a rationale for the dose (30 minutes, three times per week) and duration of the intervention (two months)? On what basis were these decisions made?

This is another good point; given the lack of direct evidence on this issue, we relied on related empirical evidence and our clinical research experiences. We now report on p. 8 (Study Design) that the “dose” of treatment was estimated based on the reported amount of training tolerated and required to evince learning in prior face processing intervention studies of children, adolescents and adults with ASD. The goal is for participants to obtain a minimum of 10 hours of training specifically on eye gaze tasks across the 2-month training period, which may require a total of 15-20 hours of total game play.

3. p. 8, para 3. One of the inclusion criteria is that the adolescent should have normal vision and hearing with correction. I assume this will be, as reported by parents rather than tested by the researchers? It would be helpful to specify this.

The reviewer is correct. We now clarify that the inclusion criteria regarding normal vision and hearing with correction is verified via parental report (see p. 8).

4. p. 10, para 2. Who is conducting the randomisation procedures?

We have now clarified that the Principal Investigator, who will not be involved in testing participants, will randomize participants and notify staff who are working with the participants about the condition assignment (see p. 10).

5. p. 10, para 3. Will the same researcher be involved in the pre- and post-intervention assessments in an effort to minimise measurement error?

Another excellent point to clarify. Unfortunately, due to our desire to provide maximal flexibility for research participant scheduling, together with our need to have the assessments on a particular interval, we cannot always ensure that exactly the same person will conduct assessments at both time-points. Rather, we now clarify that, “Measurement error of the eye tracking data will be minimized by having the same small number of highly trained researchers collect the data at both the pre- and post-intervention sessions.” (see p. 16)

6. p. 12, para 2. It is commendable that the researchers are measuring feasibility – but feasibility should not be limited to the intervention itself, it should also include other features of the trial. For example, feasibility can include procedures for (1) estimating the standard deviation of the outcome measure, which is needed to estimate adequate sample size for a full trial and (2) determining the follow-up rates, response rates to questionnaires and adherence/compliance rates. I encourage the researchers to also 4 record such details to inform a larger trial (should this preliminary trial yield positive effects).

Thank you for the recommendation to include strategies for estimating feasibility of the outcome measures. We have now included this information on p. 13. It reads:

“The feasibility of the testing procedures will also be assessed. We will report adherence rates, means, and standard deviations for each outcome measure separately for each group in each of the pre- and post-intervention testing sessions. This will allow us to assess potential floor or ceiling effects in any of our measures, collect information relevant for determining effect sizes, and estimate sample sizes for a full trial.”

7. p. 13, para 3. It is great that the researchers are recording adverse events and unintended effects (i.e., harms) during the course of the intervention. But the recording procedures are lacking in detail in the protocol description and seem rather ad-hoc to me. Do the researchers have any particular potential harms in mind (perhaps that they have gleaned from pilot work)? Again, qualitative interviews at follow-up would be very useful in this regard.

The reviewer correctly notes that we have tried to be attentive to this issue and picks up on the fact that we do not have particular potential harms in mind that might result from participating in this intervention. Rather, we expect the intervention to be minimal in risk for several reasons. It is designed from an empirically informed approach, administered remotely, designed to flexibly accommodate participants’ schedule, and is semi-supervised. At the same time, we are mindful that unanticipated risks might arise. As a result, we have instituted a DSMB and we have included self-report and behavioral measures that allow us to broadly monitor for any unanticipated risks. We also include a qualitative usability survey that we conduct at the post-intervention session that is described in the Safety Outcomes section. We now provide this more complete rationale about the expected minimal risk on p. 13 in the section called Safety outcomes. We have also moved the information about our procedures for monitoring suicidal ideation and self-injurious behavior to this section from the Study Monitoring section.

8. p. 13, para 2. The dependent variables of interest for the visual attention to faces task include the average gaze time to faces and the proportion of total gaze time to faces. It’s unclear to me why the region of interest is the entire face, rather than the eye region, given that increased sensitivity to eye-gaze direction is the proposed active ingredient in this study.

This is, of course, a critical issue and something that we have given a lot of thought to. We determined that focusing the Areas of Interest (AOI) for the eye tracking metrics on the eyes, and not the face more generally, is likely to underestimate the effects of the intervention for two reasons. First, as the Reviewer made clear, the goal of the intervention to help adolescents with autism learn to attend to and use eye gaze cues to solve problems. However, we do not know whether the participants will be able to do so. It may be that adolescents with autism are only able to learn to solve puzzles in the earliest levels of the game that provide multiple non-verbal cues, including turns of the head, which include the face. Turns of the head provide directionally important social communicative information that are often (not always) related to gaze direction in the real world. If adolescents cannot learn highly specific information about gaze trajectories from the game but they can learn general trajectory information that is correlated with head direction, this will still be a successful beginning component of learning that may be reflected in fixation time to faces, but not eyes specifically. Second, the literature investigating whether adolescents with autism and TD adolescents exhibit differences in looking time to eyes and/or mouths is highly inconsistent (see Guillion for 5 review, 2014). This may be due to many factors including the specificity of AOIs, especially in dynamic stimuli. Therefore, using larger face AOIs, which are inclusive of the eyes, and are much easier to define in dynamic stimuli in a reliable way, may help resolve some of these methodological issues and also test the effectiveness of the intervention. We now provide a brief explanation about this approach on p. 14.

9. p. 13, para 3. I am intrigued about the development of the eye-gaze sensitivity task. From the description, it seems that it would be rather easy for participants, especially for cognitively able autistic adolescents. Have the authors piloted this task with their target population to preclude the possibility of ceiling effects on this task? Some background on task development and assurances that the tasks is developmentally appropriate and sufficiently sensitive to capture change would be beneficial here.

We agree that ceiling effects would be problematic in this context. The eye-gaze tasks are adapted versions of tasks developed by Riby and colleagues (Riby & Doherty, 2009; Riby, Handcock, et al., 2013), who originally reported that adolescents with autism perform 2 SD worse than age-matched typically developing adolescents on their versions of the tasks. In our adapted version of the task described in this work, we found that typically developing (TD) adult males with high autism-like traits also show relative deficits on this task compared to TD adult males with fewer autism-like traits (Whyte & Scherf, 2017).

In addition, we tested 50 TD adolescents (ages 11-17 years) on the screening version of this task. Critically, they perform above chance and below ceiling levels of performance (M = 85%, SD = 9%). Therefore, we similarly do not expect that adolescents with ASD will exhibit ceiling effects on this task given their reported deficits in sensitivity to eye gaze. Given the importance of this issue as noted by the reviewer, we have provided more detail regarding the development of the screening version of this task in the recruitment section, including more explicit information about testing with the TD adolescents (p. 10).

10. p. 14, para 2. The researchers are proposing to administer the Social Skills Improvement System and the Social Responsiveness Scale. Could the authors please clarify their rationale to include both of these measures, which have overlapping content. In particular, I am concerned about the use of the SRS and whether it will be sufficiently sensitive to change over a 2 or 3-month period – especially given that the SRS asks parents/teachers to report on the young person's behaviour *in the past 6 months* (and will thus overlap with both the pre- and post-intervention time-points).

The Reviewer is correct that there is some overlap in the content measured in the SSIS and the SRS. However, there are also important differences captured by these measures and they are differentially associated with intervention outcomes in ASD. We address these issues on p. 17 in the DATA COLLECTION section. Briefly, the SSIS provides separate indices of adaptive social skills and problem behaviors that have been tested on a wide range of clinical populations, including but not limited to ASD. In contrast, the SRS-2 is specifically designed to characterize social impairments and repetitive/stereotypical behaviors of ASD. In addition, there are items on the SRS-2 specifically about eye gaze and face processing behaviors that tend to be problematic in ASD that are not represented on the SSIS.

The SSIS has been used extensively in program evaluation, including to evaluate behavioral and social skills interventions for ASD (see Anagnostou et al., 2015). As the Reviewer has pointed out, the existing data are inconclusive regarding whether SRS-2 total scores will be sensitive to change over the time frame of the intervention (although in our view, recent work is quite positive: Duku et al., 2013; Geretsegger et al., 2016; LaGasse et al., 2014), which is part of our rationale for including the use of the SSIS. We have added additional information 6 about the effectiveness of the SRS-2 as an outcome measure in intervention studies of children with ASD on p. 17 to help address this measurement issue.

11. p. 14, para 3. While I appreciate the need to reimburse participants to take part in the intervention, I'm not sure that paying participants $5 for every 30 minutes of game play up to $200 will really allow the researchers to examine feasibility, i.e., whether the intervention is acceptable in the absence of a (rather lucrative) reward.

We recognize the Reviewer's concern about whether a potential difficulty is introduced with respect to our ability to assess feasibility of the intervention in the absence of participant payment. We also considered this concern very carefully, particularly in the context of multiple factors. First, our broad motivation and view is that this intervention was designed to be a serious game, which is meant to foster intrinsic motivation to learn a difficult skill. We thus hope (and expect) that the participants will want to play because the game is interesting and motivating. Second, in order to determine eligibility for the study, participants have to be evaluated in multiple ways that require lots of commitment on their part. The ADOS, IQ, and reading assessments alone take 90 minutes. Many participants will drive from out to town to come to Penn State University to do this assessment and the pre-intervention testing, which is another 120 minutes. They will have already shown an interest in and commitment to the project long before they start playing the intervention game. Finally, in order to earn the full $200 during the intervention, participants will have to play 100 30-minute sessions. We are only asking them to play 24-72 30-minute sessions. Therefore, when considering all the aspects of participation in the study together with the engaging nature of the intervention, we think that the financial compensation may be less of a motivating factor than it might be in other studies. However, we do recognize that ultimately, it will be essential to evaluate the feasibility of the intervention in the absence of financial compensation to the participants, and we hope to conduct that work in the future should this initial trial provide positive results.

We now address this issue in the Discussion section.

12. p. 16, para 1. The researchers state that they will analyse the groups (intervention vs. waitlist control) for potential differences in demographic characteristics. It is not clear, however, why they will do this (as far as I can tell, there are no stopping rules for the trial) or how they will avoid potential biases here and ensure that allocation remains concealed until post-intervention assessments are complete.

We apologize for the ambiguity in this description of the analysis process. To be clear, we will utilize a stratified randomization technique based on sex and FSIQ to assign participants to a condition. We will not use an adaptive randomization procedure. We plan to enroll 17 participants per group, which we explain in the sample size estimates on p. 9. When we finish enrolling the designated number of participants, enrollment will end.

Once the randomization is complete, it is still possible that the intervention and waitlist groups are not balanced on the full set of demographic variables. To investigate this possibility, we will evaluate potential group differences on the full set of demographic variables. If any exist, we will submit those scores as co-variates to the subsequent analyses of group differences on the outcome variables of interest. We have now clarified this approach on p. 17 of the manuscript.

Reviewer: 2
Well thought out preliminary RCT. The main limitation is discussed in terms of the blindness of the researchers collecting the post intervention data.

As the Reviewer notes, we acknowledge that we cannot guarantee that the data collection team will be blinded to condition assignment at the post-intervention testing session. We have made efforts to reduce any bias, but do recognize that this is a limitation of the current design. However, it is important to note that the primary outcome measures are likely to be fairly robust to investigator bias. We have addressed this in the Discussion.

Reviewer: 3

1. I did not found the dates of the future study.

The study is currently ongoing and in a pre-result stage. We are not clear where this is to be mentioned in the manuscript beyond designating that this is a pre-results manuscript.

2. Line 33-35 (p 7). It is good to clarify what "evidence-based game mechanics" do you mean there.

We have added more information about the range of evidence-based mechanics that are useful for enhancing motivation in serious games, including "immersive storylines, goals directed around targeted skills, rewards and feedback about goal progress, increasing levels of difficulty, individualized training", and the provision of choice [for review see 32].

3. Line 18 (p 8). Please give the definition of RCT acronym.

We apologize for this oversight and have now clarified that RCT stand for Randomized Clinical Trial.

4. The main issue with the paper I see the lack of game description. I suppose that some pictures and game logic or scenarios (via UML diagrams) will be helpful to see the potential of the intervention. At the line 3 (p 12) you have "joint attention episodes when two people engage in mutual gaze with each other and then engage in joint attention on the same specific object", do you mean two avatars or two people? I guess that some screen picture will be helpful. The same with Line 18-26 (p 12).

Thank you for this suggestion; we certainly want readers to understand the process and gameplay logic models. We now try to represent a fairly complex dynamic training structure in a (hopefully) easy

to understand format. To help visualize progress through the game we have added several figures. First, we include Figure 2, which represents screen shots of multiple training conditions from the game, including an episode of joint attention to help visualize the nature of the tasks.

In the first paragraph of the section the Reviewer is referring to we now more clearly describe the general goals of the intervention and provide a general definition of joint attention:

“joint attention episodes when two people (i.e., avatars) engage in mutual gaze with each other and then engage in joint attention on the same specific object.” We also hope the reference to Figure 2 will help clarify the confusion as well, as the reviewer suggests.

Second, we now include Figure 3, which illustrates a schematic representation of the game structure. It illustrates the way the game is organized to train learning about three functional uses of eye gaze cues (referencing locations, referencing objects, joint attention) in three sequential phases that increase in difficulty. It also shows that within each phase there are multiple levels that are defined by the number of non-verbal cues avatars use to guide participants to solve puzzles in the game. Easy levels have multiple cues. Level progression increasingly focuses learning to use eye gaze cues exclusively by stripping away other cues.

Within each level, there are 6 stages. Each stage represents the number of potential objects or locations that the participant has to discriminate between based on the cue from the avatar. In the easiest stage, the participant choses between two objects or locations that the avatar is pointing, directing shoulders, head and gaze to (as in Level 1), whereas in Stage 6, the participants chooses between 6 possible objects or locations that the avatar could be referring to with the non-verbal cue(s).

Finally, we have included the full UML diagrams of each phase of the game in separate figures in Supplementary materials. These diagrams provide the full range of progress that participants can make through the eye gaze training.

5. Line 3-4 (p 13) you have “Second, in the early stages of the game, participants are provided with multiple kinds of non-verbal social cues in their interactions with the avatars” and then you have a few examples of such cues. Do you have a full list? And why did you select that one?

The comprehensive list of nonverbal behavioral cues includes, pointing, orienting of shoulders, head turns, and gaze cues, which are now clearly represented in Figure 3. These behaviors are communicative behaviors themselves (i.e., pointing, gaze cues) or are associated with and often predictive of these communicative cues (shoulder and head turning). The full list of behaviors is now described in the Current Serious Game Intervention on p. 6 and are represented in the game figures (Figures 3 and Supplementary Figures 1-3).

6. Line 16 (p 16). How did you get 70-90 minutes for the eye-tracking experiments?

The total time of the eye tracking protocol is 32 minutes. In addition, based on best-practice standards in the field, we developed a 10-minute introduction to the eye-tracking procedure and an overview of the schedule to help participants adjust to the room and develop an understanding about the order of events for the testing procedure. We also carefully piloted the procedure with typically developing young children. Additionally, as we want to be minimize participant burden, we have multiple scheduled breaks to prevent fatigue and there are also required eye tracker calibrations before each task; accordingly, we have estimated that the entire protocol will take between 70-90 minutes. We have clarified this on p. 16.

VERSION 2 – REVIEW

REVIEWER	Liz Pellicano Macquarie University, Australia
REVIEW RETURNED	01-Aug-2018

GENERAL COMMENTS	The authors have done an excellent - and very thorough - job of responding to my comments. I particularly appreciated the inclusion of Figures 2 and 3 and the supplementary figures, in response to Reviewer 3's comments. I have no further comments for the authors to consider.

REVIEWER	Pavel A. Orlov Imperial College London
REVIEW RETURNED	15-Aug-2018

GENERAL COMMENTS	Authors have done a good job of updating the manuscript. They carefully addressed all my comments and provide the necessary figures, diagrams, and details. Also, I am glad to see their answers as well. And I think that the manuscript becoming much better now and the study is much clear for readers. It could be published.