Dear Dr. Karakashevska,

Thank you for your thorough reply to the reviewers and recommenders. We have now received positive feedback from three of our reviewers and will proceed with an invitation to reply to the few remaining comments.

We are delighted that the reviewers and recommenders are positive about our revision.

One of the reviewers highlights the necessity of a multiple comparison correction regarding the sub-hypotheses under hypothesis 3, especially considering how these pairwise comparisons come together in support of hypothesis 3. We agree with the reviewer and suggest you could use Bonferroni and accommodate the adjusted alpha for each comparison. Power should be calculated for the t-tests with respect to the adjusted alpha. As your study seems very highly powered, this should not be a problem.

Futhermore, we will reiterate our suggestion that you only run the pairwise comparisons to answer hypothesis 3 rather than include the RM ANOVA to examine the main effect of Block. The reason we suggest jumping straight to the four pairwise comparisons is twofold: 1) a main effect of Block does not add anything beyond the pairwise comparisons, each of which directly tests a claim of interest; 2) a Bonferroni already controls familywise error rate without the need for the omnibus main effect to be significant (the only function of testing the omnibus main effect before proceeding to specific comparisons is to control familywise error rate).

We have now removed the Repeated Measures ANOVA, and gone straight to the four pairwise comparisons under hypothesis 3. We have changed the section on hypothesis 3 to explain this.

> *"Hypothesis 3 predicts perspective cost differences across the four blocks (red bars in Figure 3). This can be stated as four predicted pairwise differences:*
>
> *a) Perspective cost will be reduced in the Monocular viewing block (as compared to Baseline).*
>
> *b) Perspective cost will be reduced in the Static frame block (as compared to Baseline).*
>
> *c) Perspective cost will be reduced in the Moving frame block (as compared to Baseline).*
>
> *d) Perspective cost will be reduced in the Moving frame block (as compared to Static frame block)."*

We have powered our study to find these pairwise differences with a Bonferroni correction for multiple comparisons (0.05/4 = 0.0125). Power analysis changes a little when we adopt this more conservative threshold (previously 0.02). The new power analysis section explains why the sample of 120 is still adequate.

> *We chose a sample size of 120. The most demanding part of our research, in terms of sample size, is testing the four pairwise perspective cost differences stated under hypothesis 3. N=120 gives 88% power to find significant pairwise differences with effect size d = 0.34, and with alpha = 0.0125. This conservative threshold of 0.0125 reflects a Bonferroni correction for 4 pairwise comparisons (0.05/4 = 0.0125). We applied the Bonferroni correction despite that fact we make a priori predictions, and we used two-tailed tests despite the fact these predictions were directional. By some conventions, one tailed tests with no correction for familywise error rate would be justifiable, but we lean towards 'overpowering' the study rather*

than missing real effects (Brysbaert, 2019). We also note that the median sample size in previous SPN research is just 24. Our sample of 120 is more than twice as large as any published or unpublished within-subjects SPN experiment. The sample of 120 is partly constrained by needs to balance block order (24 orders X 5 repeats of each). Sample size must be a multiple of 24, and N= 96 would be underpowered by these standards (power = 0.78, alpha = 0.0125, d = 0.34, two tailed) while N=144 exceeds our resources and time constraints.

Also be clear what error term you will be using when you specify t-tests (presumably the error term specific to that t-test rather than one derived from the ANOVA). The error term specific to the t-test is what you would get just by asking a package to perform a t-test, rather than a multiple comparison as a post-hoc test as part of an ANOVA.

Yes, we will be using the error term specific to the t test rather than the ANOVA.

For clarity, we also request that you separate the H3 hypotheses and their four t-tests in the design table.

We have now separated the H3 hypotheses and tests in the table.

Minor

Missing word in the additional sentence: "In a recent SPN study (Karakashevska et al., forthcoming), we found polygons slightly *** perspective cost, but do not eliminate it."

We have added ‘reduced’ to complete the sentence.

Please accompany revisions with a reply to reviewers and a tracked changes version of the manuscript. We look forward to hearing from you.

Best,

Grace & Zoltan

Review by Guillaume Rousselet, 24 Sep 2023 09:58

The revised version of the article and the reply addressed my questions. I only have a few minor suggestions left.

The new Figure 4 is very useful. Try to match font sizes across panels and align the ABCD labels. In panel A, font sizes are too small and the y axis seems to have two superimposed lines.

We have updated Figure 4 as recommended.

"We powered our experiment ..." should be changed to "our line of research". Power is not defined for a single experiment. Alternatively, explain how you determined your sample size, using a power analysis...

We have now explained how we determined sample size using power analysis. The power section has now been shaped by other reviewers and editors. It now reads:

We chose a sample size of 120. The most demanding part of our research, in terms of sample size, is testing the four pairwise perspective cost differences stated under hypothesis 3. N=120 gives 88% power to find significant pairwise differences with effect size d = 0.34, and with alpha = 0.0125. This conservative threshold of 0.0125 reflects a Bonferroni correction for 4 pairwise comparisons (0.05/4 = 0.0125). We applied the Bonferroni correction despite that fact we make a priori predictions, and we used two-tailed tests despite the fact these predictions were directional. By some conventions, one tailed test with no correction for familywise error rate would be justifiable here, but we lean towards 'overpowering' the study rather than missing real effects (Brysbaert, 2019). We also note that the median sample size in previous SPN research just 24. Our sample of 120 is more than twice as large as any published or unpublished within-subjects SPN experiment. The sample of 120 partly constrained by needs to balance block order (24 orders X 5 repeats of each). Sample size must be a multiple of 24, and N= 96 would be underpowered by these standards (power = 0.78, alpha = 0.0125, d = 0.34, two tailed) while N=144 exceeds our resources and time constraints.

"are significant " -> "are statistically significant"

Done.

About filtering: ideally, add details about the filter characteristics. FIR or IIR, specific kernel... Or at least report the name of the function and the version of EEGLAB it came with. This is important for reproducbility, because for years EEGLAB provided very poor default filter settings.
As for the 25 Hz distortions, your figure is convincing. I would just cite your 2020 paper to support the lack of signal distortions.

We have now cited the paper and filter version in the manuscript (see EEG processing section).

pop_eegfiltnew is the FIR filter and the intended as a replacement for the deprecated pop_eegfilt function. Required filter order/transition band width is estimated with the following heuristic in default mode: transition band width is 25% of the lower passband edge, but not lower than 2 Hz, where possible (for bandpass, highpass, and bandstop) and distance from passband edge to critical frequency (DC, Nyquist) otherwise.

Channel interpolation: I didn't mean not to remove bad channels, but simply not to interpolate them after removal. You use the example of ICA: interpolation doesn't recover any information, so your ICs will be the same with or without interpolation. It is up to you but from experience, interpolation will give you nothing. For topographic maps, interpolation is built into the algorithm, so again, no gain.

We appreciate this advice, and we are not experts on channel interpolation vs simple removal. However, following Delorme's eeglab guidelines and the Harvard Automated Processing Pipeline recommendations, we choose to interpolate missing and bad channels rather than remove them. This has the minor advantage of keeping our scripts consistent with other SPN research. Moreover, missing channels create inconsistencies in matrix dimensions (e.g. 64 channel X 512 time becomes 63 electrode X 512 time for participants where there was a missing channel). This causes numerous 'matrix dimensions must agree' errors in our pipelines.

"However, the validity of these tests is questionable." -- I'm not sure if the tests can be described as invalid. What we know from simulations is that these tests have very low power, so failure to detect deviation from normality is not conclusive. Also, accepting normality based on p>0.05 is a statistical fallacy.

In the text, I would phrase your conclusion more carefully, suggesting that assuming normality is a reasonable *approximation* to the population distribution. Afterall, we know that ERPs cannot be normally distributed -- values are necessarily bounded. And if you suspect skewness, you could always run simulations using a small amount of skewness (g-and-h distributions are great for that). One-sample t-tests using trimmed means will increase power in the presence of skewness.

We have changed the section on Normality as recommended.

> *"The ridgeplot in Figure 4A shows that individual participant SPNs are often normally distributed around the mean. Indeed, only 8-9% of the 249 SPNs violate the assumption of normality according to Shapiro-Wilk and Kolmogorov-Smirnov tests ($p < 0.05$). However, median sample size in this analysis was just 24, and these tests will often miss departures from normality. Meanwhile accepting normality based on $p > 0.05$ is a statistical fallacy (confirming the null hypothesis). We therefore analysed the distribution of 249 skewness statistics associated with the 249 SPNs. There is a small but significant mean negative skew when regularity is task relevant (Figure 4C, mean = -0.174 microvolts, SD = 0.529, t (124) -3.665, $p <$ 0.001). However, this is less pronounced when regularity is not task relevant (Figure 4D, mean – 0.081 microvolts, SD = 0.561 t (123) = -1.609, $p = 0.110$). We there assume that the normality is a reasonable approximation of the population distribution."*

We have run the power simulations with a small amount of skewness, using the sn rather than rnorm to create distributions. As you say, this slightly increases power, but by very little. We do not expect skewness, so we have not added this complication to the manuscript.

In case you are interested, the distributions and power curves are shown below:

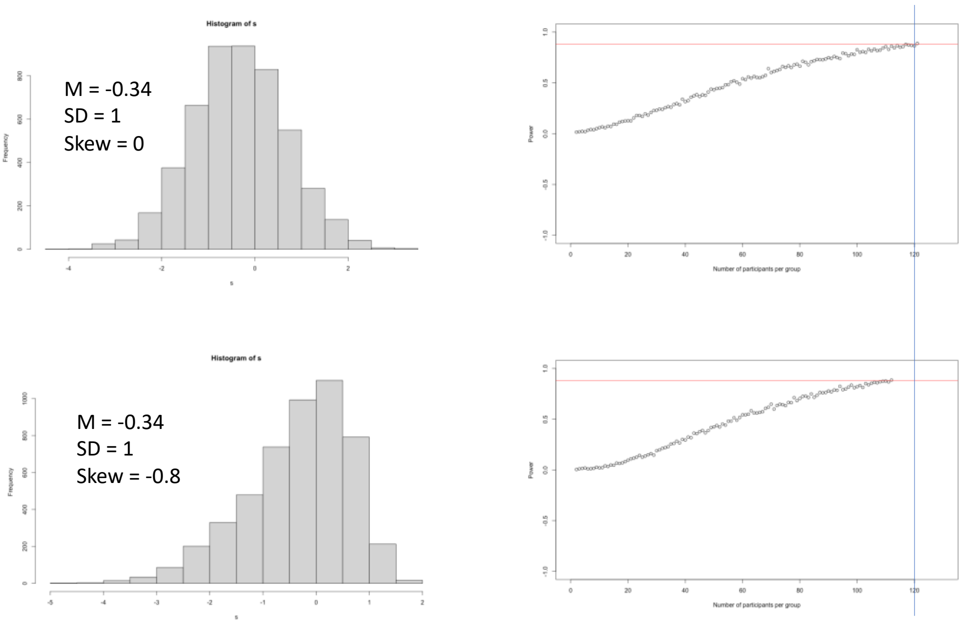

Figure 1. Top row shows power simulation where data are sampled from a normal distribution with a mean of -0.34 and a SD of 1. The power curve shows that 90% of t tests are significant when we select samples of 120. The bottom row shows the same analysis when we sample from a population with a negative skew (gamma = -0.8, the mean < median). The power curve is shifted right, and 90% of t tests are significant when we select samples of 112. The vertical blue line aids comparison.

Review by Tadamasa Sawada, 27 Sep 2023 18:52

In the revision of the manuscript, entitled "Putting things into perspective: Which visual cues facilitate automatic extraretinal symmetry representation?", by Elena Karakashevska, Marco Bertamini, and Alexis D.J. Makin, the authors addressed issues raised by the reviewers. Now, I have only a few minor concerns.

p.9. Hypothesis 3, which is a set of four sub-hypotheses, is discussed on the basis of the results of statistical tests of these sub-hypotheses. How will the authors combine the results of these statistical tests about the sub-hypotheses to discuss Hypothesis 3? If the authors use a logical operator “or” (at least one of the tests show results consistent with the predictions), A Bonferroni correction of the statistical tests is necessary. If the authors use a logical operator “and” (all of the tests show results consistent with the predictions), the correction is not necessary.

Our treatment of hypothesis three is shaped by your recommendations here, but also by the editors (see above). They recommend that we do not use the repeated measures ANOVA with main effect of Block but go straight to the four pairwise comparisons. The editors also recommend that we use Bonferroni correction for these four tests, and this an alpha level of 0.05/4 = 0.0125. While we could argue that the Bonferroni correction is not necessary because we predict that all four tests will show results consistent with predictions (and not or), there are advantages to using stricter thresholds anyway and powering the study accordingly.

p.14. "Lux" is the unit of illuminance and not of luminance. Luminance cannot be computed from the illuminance without using some additional information.

We now report cd/m^2.

p.14. > *two visual transformations, first adopting the position of the virtual camera, and then correcting for perspective distortion (Sawada & Pizlo, 2008).*

I do not see any part of Sawada & Pizlo (2008) discussing such a mechanism in the visual system.

We have moved this reference to a different location. We wanted to cite another paper which mentioned the importance of eye position, not imply Sawada and Pizlo (2008) had proposed a specific two visual transformations mechanism.

p.15. > *The angles of 60 and 15 degrees were also chosen to follow recommendations in Sawada and Pizlo (2008).*

Sawada & Pizlo (2008) used slant between 50 and 70 degrees for their visual stimuli but they were not making any general recommendation about the slant.

We have now changed this paragraph to avoid misrepresentation of Sawada and Pizlo (2008):

> *These perspective stimuli have several advantages over those used by Makin et al. (2015). In Makin et al. (2015), the position of the participant's eye and virtual camera were not matched (see Sawada & Pizlo, 2008 for more on importance of eye position). This is a limitation, because the participants in Makin et al. (2015) had to do two visual transformations, first adopting the position of the virtual camera, and then correcting for perspective distortion. Furthermore, symmetry around the vertical axis was not substantially disrupted by the perspective in Makin et al. (2015). This feature can be seen by inspecting the stimuli in Figure 2. Consequently, if participants focused spatial attention on the axis region, they would have near-perfect retinal symmetry to guide judgements. In the new study, slant and tilt were used to reduce retinal symmetry around the axis. The angles of 60 and 15 degrees were chosen for consistency with previous work (same as Karakashevska et al., forthcoming, and the centre of the slant range used by Sawada and Pizlo (2008).*

Figure 6B. I do not see any purple dots in this figure. In this figure, there is an arrow, dotted grids, circles, and a horizontally-long rectangle but none of them are explained in the caption of the figure.

It is likely that you are not viewing the intended Figure 6B, we apologise for the confusion. It should look a semi-circular protractor superimposed on a blue horizontal rectangle labelled 'screen'. This should be aligned with a circle inside a square. There should be coloured dots around the perimeter of the protractor, representing possible virtual camera positions, including the ones used in this study (purple). This diagram is important for understanding the stimulus rendering algorithm in Python. We have saved as an image file and superimposed below.

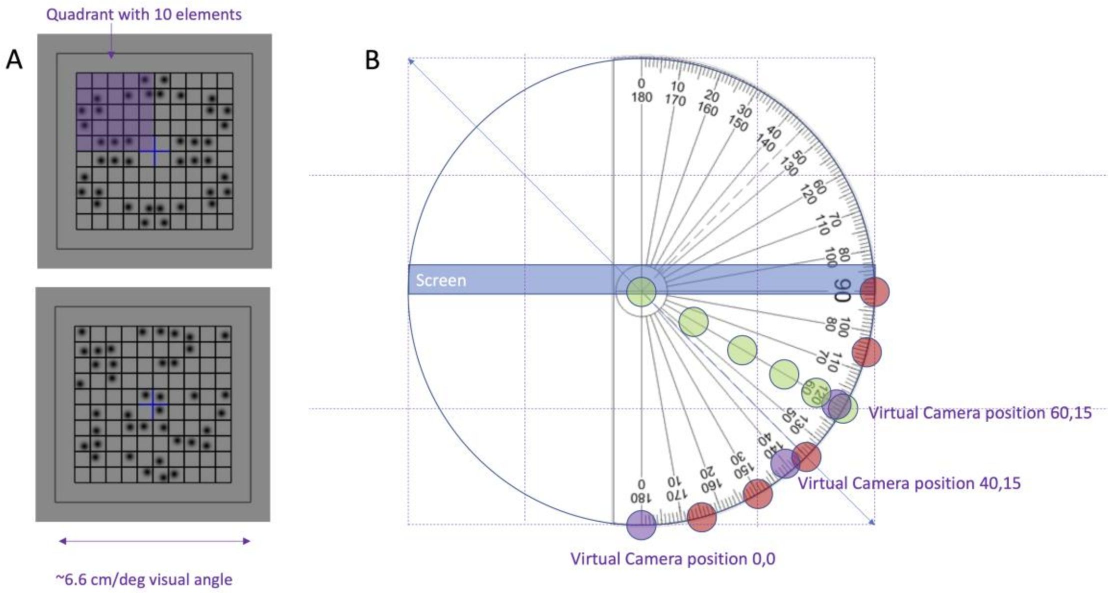

Figure 6. Stimulus construction diagram. A) Arrangement of dots in symmetrical and asymmetrical exemplars. *The top panel shows construction of a symmetrical exemplar. A quadrant is populated with 10 small Dot elements (highlighted in purple). This quadrant was reflected across horizontal and vertical axes. The bottom panel shows construction of an asymmetrical exemplar. Here all 4 quadrants are independent. B) The protractor diagram represents virtual view angles used to generate perspective stimuli. This is a top-down view of a screen. The centre of the screen is in the centre of a virtual sphere. The protractor represents the equator of this sphere. Purple dots are virtual camera positions used in stimulus rendering. The camera is always focused on the centre of the screen/sphere.*

Reading the authors' reply to my comments in the last review, I see that the authors want to discuss which of the projective transformations or the perspective transformation are "superior." As I mentioned in the last review, the perspective transformation is a sub-set of the projective transformation so, we cannot say which is "superior." When we use these transformations for a particular application, we can discuss which transformation is valid for this application.

We are pleased to have feedback from a reviewer who is an expert on visual transformations, and we are keen to avoid any erroneous claim that perspective transformations are superior if this is not the case. We were probably using the terminology incorrectly, so we have taken all mention of 'projective' transformations' out of our manuscript.

What we are trying to say is that our stimuli are optically correct given the participants viewing distance (57 cm) and the stated levels of slant and tilt (e.g. -60, +15). This was not always the case in our previous research (Makin et al. 2015).