Dear authors

Your Stage 1 RR manuscript has now been reviewed by three experts in the field. As you will see several reviewers comment on how to be sure if participants indeed experience perceptual fading. I agree that this is a critical point to be addressed before commencing the study. However, I also advise caution when adopting a trial-sorting procedure, such as basing these analyses on perceptual ratings or continuous psychophysics as suggested by some reviewers. Such an approach could risk selection artifacts (related to regression-to-the-mean) as described in *Shanks 2017, Psychon Bull Rev*. This danger largely depends on how reliably participants report their actual percept. If there is some variability this results in erroneously assigning trials which can skew the summary statistics (this could be caused by the somewhat unintuitive response mapping- see below). So this will take some careful consideration. Would it make more sense to include all trials but incorporate the Likert rating in the model (as in the current Hypothesis 5)?

We have considered a trial-sorting procedure as suggested by R3 (and opted against this), and also the other reviewer suggestions, some of which we have adopted as we will discuss in more detail below. Before we get into the details, a general comment on the issue of perceptual Fading: the adaptation in the Fading block unfolds rapidly, often within just a few trials (e.g., after 10 trials). Moreover, once participants experience perceptual fading, the effect tends to remain robust throughout the duration of the block (and also resistant to small eye movements). In light of this, we anticipate that participants will consistently rate their experiences with 2 or 3, and thus that the reliability and variability of the responses will not be a problem.

I also included several other comments that should be addressed prior to in-principle acceptance:

- Typo in **Abstract**: "cued" instead of "cue".

  Thank you for bringing this to our attention. We changed it now.

- **Power calculation:** When using linear mixed models authors often use simulations for power calculations whereas your power estimate is based on an ANOVA instead. While you argue that this is a lower-bound estimate, you could consider using more direct simulations approach.

  We now estimated sample size using a more direct approach (simR package that allows power calculations for generalized linear mixed models). We explain the details of the analysis under the Sampling plan section.

- **Null hypothesis:** On a related note, I concur with one reviewer regarding claims relying on confirming the null hypothesis. Please consider alternative statistical approaches (such as Bayesian inference, equivalence testing, or a predefined confidence interval including zero).

We added an additional Bayesian test for null hypothesis testing (Hypothesis 3.1. And Hypothesis 4)

- **Stimulus details:** Please report the luminance of the grey background. Is the display linearised so this is the median between the white and dark patches? Related to that, the stimulus patches are presumably Gaussian blobs?

  We now report the luminance of the gray background and state that the display will be gamma-calibrated. And yes, the stimuli are gaussian blobs with no grating.

- **Response mapping in the Likert scale** seems somewhat awkward as the 1-4 options are mapped to the four arrows. Why is it done this way instead of using the left and right arrow to move a slider, or using number keys?

  We agree that the key mapping is not very intuitive, however, we wanted to prevent participants from moving their eyes (which is likely to happen if we use numbers 1-4). Therefore we decided to stick to the arrow keys and keep the numerical values next to the arrows on the display as a reminder. Also, we increased the number of practice trials to make sure that participants get used to the mapping.

- One reviewer also points out a logical disconnect in the **Abstract**. I don't personally perceive it as such, but for clarity and style you may wish to revise this. Either way, the abstract should be the same as it will be in Stage 2 (except for adding results/interpretation and changing verb tense). So if you will revise this, you should do it now.

  We revised the second sentence of the Abstract now. We hope that it flows better now.

- I commend you on the **Design Table**. This is very well organised, concisely written, and the colour coding is a very helpful visual aide.

  Thank you for your compliments!

Sam Schwarzkopf

*by **D. Samuel Schwarzkopf**, 04 Mar 2024 02:12*
Manuscript: **https://osf.io/m2xr5?view_only=99e4cfe64c344980ae3a03324ef85eb1**
version: 1

Review by Sander Nieuwenhuis, 30 Jan 2024 08:25

This is a very clear stage 1 RR. The research question makes sense and is clearly defined; the hypotheses are capable of answering the research question; the protocol is sufficiently detailed to enable replication by an expert in the field; the links between hypotheses, statistical tests and possible interpretations are clear; the sample size is sufficient; and the authors have built in several manipulation data quality checks.

We thank Sander Nieuwenhuis for his compliments and valuable suggestions.

I have only a few minor comments:

- The text does not contain any information about eye movements right now: no instructions for subjects and no plans about how to deal with (unwanted) eye movements. These should be added.

We now elaborate that the participants' task is to keep fixating at the center of the screen and only covertly attend to the cued patch.

As regards dealing with eye-movement artifacts, we now state in the Data preprocessing section that we will exclude all trials where the horizontal gaze deviation from the center of the screen was larger than 7.42 degrees (distance to the inner edge of the patches) and lasted longer than 10ms.

- Two of the hypotheses rely on conventional null hypothesis significance testing to conclude evidence of absence from null results. For example: "Hypothesis 3.1 will be confirmed by the absence of a main
effect of Covertly attended brightness." The authors should use a method that allows one to draw such a conclusion, such as Bayesian hypothesis testing.

We agree with Sander's point, and we now include an additional (Bayesian) test for Hypothesis 3.1. and 4.

- About the writing, something to keep in mind for the stage 2 RR: The second sentence of the abstract does not follow from the first sentence.

We revised the second sentence of the Abstract now. We hope that the Abstract flows better now.

Review by Martin Rolfs, 03 Mar 2024 23:29

The proposed study pursues the question how the change in perceived luminance for a perceptually fading, but physically unchanged stimulus affects pupil size. I applaud the authors for this intriguing idea and for submitting a Registered Report before starting the study. I list a number of thoughts and suggestions below.

1A. The scientific validity of the research question(s).

The literature review is highly informative with respect to the study, resulting in a scientifically valid and intriguing set of research questions.

1B. The logic, rationale, and plausibility of the proposed hypotheses, as applicable.

The authors specify five clearly spelled out hypotheses, including replications of previous studies (Hypothesis 1), sanity checks (Hypothesis 2), novel predictions (Hypotheses 3, 4, 5). The main hypothesis is that the attentional effect (attending to the location of a black vs white luminance patch) is reduced in the fading conditions. The rationale of the study makes sense, the hypothesis is derived from previous research and plausible based on pilot data included in the RR.

1C. The soundness and feasibility of the methodology and analysis pipeline (including statistical power analysis or alternative sampling plans where applicable).

The analysis plan is straight-forward and clearly described, including pre-processing steps that the authors have used in previous work. The authors provide a power analysis (90% power, alpha level = 2%) yielding an estimate of N = 24 given the effect size of a previously published study with a similar design, and round up to a planned N = 30. They specify the inferential analysis (LMM) they intend to use for each hypothesis in a way that leaves few researcher degrees of freedom.

We thank Martin Rolfs for his compliments and valuable suggestions.

There are a few points regarding the experimental procedure that need to be addressed, however.

A report of perceptual fading follows each trial, but is not continuous in time. The authors should consider specifying to the participants for which moment in time they are supposed to report their perceptual state (e.g., at the time of target offset). Alternatively, they could track the degree of fading over time using a continuous psychophysical task (e.g., using a fader or a volume knob). This continuous report, however, would require a different set of analyses than planned so far. Another alternative would be to have participants press one of two buttons when the stimulus disappears or reappears, respectively, throughout each trial. This procedure is well established in the literature on rivalry and other perceptual alternations, including Troxler fading (e.g., Martinez-Conde et al., Neuron, 2006).

We thank Martin for his suggestions. We now clarify that participants should report the level of perceptual fading at the moment that the question appears (which is right after the target offset).

Speaking of which, it is not clear how the authors will deal with trials that contain microsaccades, saccades or blinks, which are known to counteract fading in this stimulus (Martinez-Conde et al., Neuron, 2006). Will these trials be excluded? Keeping track of perceptual state through continuous reports would provide an opportunity to analyse these occasions as a function of the perceptual state they elicit.

In order to test Hypothesis 3, we will include only the trials from the Fading block where participants reported high levels of adaptation (ratings of 2 and 3). That way, we ensure that perceptually, participants experienced fading.

As regards microsaccades, we use stimuli with fuzzy edges, for which microsaccades do not disrupt perceptual fading.

As regards larger eye-movement, we now clarify in the Data preprocessing section that we will exclude all trials where the horizontal gaze deviation from the center of the screen was larger than 7.42 degrees (distance to the inner edge of the patches) and lasted longer than 10ms.

As regards blinks, it's not practically feasible to exclude trials with blinks, because those will be too many. Admittedly, blinks may indeed counteract perceptual fading to some extent. We will rely on subjective reports as mentioned above to ensure that this effect is minimal.

The authors use a detection task, but if I understand correctly, the staircase they use appears to focus on hits only (aiming for 70% correct), which means that performance (as defined by the

authors) depends on participants' criterion. It is inherently difficult to staircase d', which would be solution to this dilemma, but that requires monitoring hits and false alarms/correct rejections. The authors should consider using the single-interval adjustment-matrix (SIAM) procedure (Kaernbach, 1990), which has been successfully applied in detection tasks relating stimulus visibility to eye movement responses (White & Rolfs, 2016; White et al., 2022).

We thank Martin for bringing this up to our attention and providing valuable references. We agree that the SIAM procedure is a more elegant way to staircase d', and therefore we are going to implement it instead of the originally planned 3 up - 1 down procedure. We now specify details of the procedure in the manuscript.

1D. Whether the clarity and degree of methodological detail is sufficient to closely replicate the proposed study procedures and analysis pipeline and to prevent undisclosed flexibility in the procedures and analyses.

The methods are described in sufficient detail to implement the study, with a few small exceptions.

The hexadecimal codes for colors are not very useful, as they depend on the specific display device being used (and its settings), and do not translate into wavelength spectra that would specify the stimulus.

We thank Martin for bringing this up to our attention. We deleted the hexadecimal codes now and left only cd/m2 measurements.

As stated above, it is unclear how exactly participants are instructed to report the perceptual state. What do they report, for instance, if fading set in during the trial? This should be specified to reduce differences between participants.

We now clarify in the manuscript that participants will report the level of fading experienced at the moment that the question appears (which is right after the target offset). We will make sure that the instructions at the beginning of the experiment are also clear.

1E. Whether the authors have considered sufficient outcome-neutral conditions (e.g. absence of floor or ceiling effects; positive controls; other quality checks) for ensuring that the obtained results are able to test the stated hypotheses or answer the stated research question(s).

The authors propose contrasting two conditions, one of which aims at a replication of previous work (Non-Fading condition) and one of which is new. There are a couple of manipulation checks that would be advisable to make this data set as convincing as it should be.

First, we need to know that the attentional manipulation worked. An analysis of performance is not reported yet, but should be included in the Registered Report. Yet, this analysis alone will not speak to the successful manipulation of spatial attention, as there is no neutral or invalid condition to compare performance to. Introducing such conditions would be desirable. Note that given previous work showing effects of attention on pupil size, the authors could use a difference in pupil size as an indicator that attention was successfully deployed. But if Hypothesis 3.1 is true, this would only work in the Non-Fading condition.

We agree with Martin's point and we now manipulate Cue Validity (cue is 80% valid), include the behavioral analysis regarding Cue Validity, and include this as an additional Quality check hypothesis.

Second, and more crucially, we need to know if attention is allocated to the same degree in both conditions. It could well be that the absence of a stimulus makes it easier to process the targets at the cued location. As a consequence, the attention shift might not be required (to the same degree) in the Fading as compared to the Non-Fading condition. The authors will thus need to show that performance in the detection task is comparable between the Fading and the Non-Fading condition. Given that they intend to equate performance across conditions (70% hits, but see my comment on using hit rates above), it would be necessary to show that the circles' opacity was the same (on average).

We agree with Martin's point that it is important to show that attention was deployed to the same degree in both blocks. However, even though this is a valid concern, we believe (in part because of our experience as test subjects) that our paradigm requires participants to deploy their attention effectively due to the thin outlines of the circles, making it very challenging to detect the offset of a single circle in both blocks. It is a highly attention-demanding task.

To specifically address Martin's concern, we will now manipulate Cue Validity and compare the cueing effect across blocks. This analysis will check whether attention is deployed to a comparable degree in both Fading and Non-Fading conditions. We hypothesize that there will be a cueing effect in both blocks (Hypothesis 6). If it turns out that the cueing effect, while present in both blocks, did differ between blocks, we will address this transparently in the Discussion.

Third, if the putatively reduced attentional effect on pupil size in Fading compared to the Non-Fading condition is indeed due to Fading, it would be important to show that this reduction is absent when no fading occurs, but the protocol is otherwise identical. The authors use alternating luminance blocks to prevent the buildup of fading over time. But it would be important to know if the alternation is also affecting the pupil response per sé, that is, as a consequence of alternating, not because on reduced Troxler fading. A negative control of this kind would be to use sharp-edged circles in a different set of blocks (either alternating or not alternating). This point is the least important of the three points so far, as it is also addressed to some extent by Hypothesis 5

We agree with Martin's point and we tried the version with sharp edges on ourselves. However, we observed a slight fading of the contours around the edges and occasional fading of circles, plus pronounced (and difficult to describe) perceptual effects of the edges themselves. In other words, while it is true that the sharp edges to a large degree prevent perceptual fading, they introduce perceptual artifacts of a different nature, which we feel is undesirable.

Given that we want the Non-fading block to remain free of adaptation, we decided to proceed with the initial setup (swapping trials). However, to minimize the impact of swapping on our window-of-interest, we now extend the duration of the fixation dot period to 1s. This adjustment ensures that all changes induced by visual alterations (i.e., swapping) manifest during the fixation dot interval, while the window-of-interest that we will analyze starts from cue onset. Moreover, changing the procedure would leave us with no quality checks, and thus we believe that it is important to keep the initial setup.

Review by anonymous reviewer 1, 09 Feb 2024 14:46

In this study, the authors propose to investigate the relationship between the attentional modulation on pupil size and the subjective experience of brightness using perceptual fading. I find the scientific question extremely interesting and definitely worth studying. I believe the experimental strategy is overall valid and well designed. The procedure is clearly explained.

We thank the reviewer for his or her compliments and constructive suggestions.

My main concern is about the non-fading block. I believe that it is not an ideal control, because of the different experimental procedure employed (no swapping vs. swapping), possibly introducing confounding factors. I think that a more rigorous and efficient solution would be to design white and black stimuli so as to have a sufficiently high number of trials where participants report fading (3-4) and to use the remaining trials, where participants report 1 or 2 on the fading scale, as a control.This way the two conditions would differ only because of the degree of fading, leaving the experimental procedure unaffected.

We agree with the Reviewer's point, which is similar to the issue raised above by Martin Rolfs. Please see the answer to the last point raised by Martin Rolfs. We do prefer to stick to a design where the presence of fading is experimentally induced, rather than purely correlational and derived from subjective reports as suggested by the Reviewer (if we understand correctly).

In any case, it would be useful if the authors would report the rate of fading of the selected stimuli on 2-3 pilot datasets.

The mean fading self report (ratings: 0-3) was 2.88 for the Fading block and 0.004 for the Non-fading block.

Finally, two minor comments:

- Considering their role in adaptation, I believe that reporting eye movements data and their statistics across conditions would be useful.

We agree with the Reviewer's comment and we now state in the Data preprocessing section that we will exclude all trials where the gaze deviation from the center of the screen was larger than 7.42° and lasted longer than 10ms.

- The authors might be a bit more specific on how to quantify the target sample size for studies with linear-mixed effects models.

We now estimated sample size using a more direct approach (simR package that allows power calculations for generalized linear mixed models).