Still no evidence that late-sighted individuals rely more on color for object recognition: Reply to Vogelsang et al.

Authors: Thomas S. A. Wallis[1,2], Joshua M. Martin[1]

Affiliations:

[1*]Centre for Cognitive Science and Institute for Psychology, Technical University of Darmstadt.

[2]Center for Mind, Brain and Behavior (CMBB), Universities of Marburg, Giessen and Darmstadt.

*Corresponding author. Email: thomas.wallis@tu-darmstadt.de

Abstract: We thank Vogelsang et al. for their reply [*1*] to our critique [*2*]. However, we remain unconvinced that the experimental data presented in their paper supports their claim. First, applying a non-parametric test to the differences in percent correct, as in their reply, does not solve the ceiling problem. Second, there is indeed enough data to apply more appropriate models that assume binomial rather than Gaussian errors; these models provide evidence for the null hypothesis. Third, we note a design issue that undermines the causal inferences that can be drawn from the experimental data. Based on these statistical and design considerations, we reiterate that in our view, the empirical data do not support the claim that object recognition in Prakash patients is more impaired by color removal than controls.

We thank Vogelsang et al. for their reply [*1*] to our critique [*2*] of their original paper [*3*]. However, we remain unconvinced that the experimental data presented in their paper supports their claim. We conclude this based on both statistical and experimental design considerations. We address their key points below.

Vogelsang et al. present a non-parametric Mann-Whitney U test on the color – gray difference scores, which has a significant p-value of 0.004, and conclude that "the results presented and the inferences drawn in the original study are reliable". This test takes the difference score in percent correct, ranks these differences, then computes a test statistic based on the median ranks. This approach reverses the order of the rank analysis we performed in our reply, in which we first transformed percent correct into ranks (see Figure 1 of [*2*]) and then performed an independent-samples t-test on the difference in rank. Vogelsang et al. [*1*] write that our rank-based analysis "[is] likely to mask the inter-group difference". We instead argue that due to the ceiling effect, the data do not provide evidence for an inter-group difference in the first place.

We believe the difference-then-rank approach adopted by Vogelsang et al. [*1*] does not solve the ceiling problem, because it bakes the (possibly) restricted range of the Control group into the subsequent analysis. Comparing differences in percentage scores makes assumptions regarding the underlying distributions of the group data; namely, that changes in percentage at different levels of the performance bound are equivalent, e.g. a change from 50-60% is treated the same as 90-100%. We believe this assumption is inappropriate, since a number of participants performed at or near ceiling, meaning that there is hard limit on the extent to which their color performance

could improve. Showing that the difference scores themselves do not significantly violate normality according to tests or using rank-based analyses based on these difference scores does not address our critique, since what we take issue with are the assumptions underlying the difference of percentages metric itself. We acknowledge that our original response could have been clearer on the central issue.

In our critique, we argued that the most appropriate analysis strategy for these data was to use a Generalized Linear Mixed Effects model (GLMM), to model the hierarchical data structure using a binomial error distribution. This model provided no evidence for an interaction between image condition and group. Vogelsang et al. [*1*] write that "our dataset may be less amenable to more complex analytical models that benefit from larger datasets", implying that the dataset size is insufficient to reliably apply these models. Here we note two points: first, the model diagnostics indicated no problems with numerical optimization convergence, indicating that the dataset is sufficient to constrain the model parameters in a maximum-likelihood setting [*4*]. Second, using a Bayesian statistical approach allows a principled treatment of low-data regimes by using prior information [*5,6*]. Bayesian GLMMs yield the same qualitative conclusion: the data do not provide evidence that Prakash patients benefit more by inclusion of color information than controls. In fact, Bayes Factors over a range of sensible priors favor the null model (ranging from BF10=0.007 to BF10=0.124, depending on prior).

Finally, we note a related issue in experimental design that we did not raise in our initial reply. The presentation order of color and grayscale images in the recognition task was not counterbalanced, nor were trials of color and gray randomly interleaved. Instead, all participants first viewed and named 100 images shown for an unlimited time in grayscale, followed by viewing and naming the same 100 images shown in color (see supplementary methods in [*3*]). This means that the experimental data presented do not provide strong causal evidence that including color improves recognition for either group: the average performance improvement for color images could instead be driven by either general learning (improving on the task) or by learning about the stimuli themselves via the second exposure. Indeed, color information has been previously shown to have little influence on object identification accuracy (though some influence on response times and recognition memory, [*7–9*]). Nevertheless, if the experiment had indeed provided evidence that Prakash patients improved more with color compared to gray images than controls (the original claim), due to these design issues this effect could be caused by differences in learning between the groups, rather than differences in visual experience of color-reduced inputs early in life.

In summary, given the reasons stated above we stand by the original arguments of our reply letter [*2*], and conclude that the experiment presented by Vogelsang et al. in [*3*] does not provide evidence of the claimed difference in reliance on color information for object recognition between Prakash patients and controls.

Acknowledgements

Funded by the European Union (ERC, SEGMENT, 101086774). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them. We thank the authors of [*3*] for making their dataset available and for engaging with our critique. Thomas Wallis and Joshua Martin contributed equally to this work.

References

1. M. Vogelsang et al. (2024, September 4). Assessing the impact of color on recognition performance in late-sighted and controls: Response to Thomas Wallis and Joshua Martin. *OSF Preprints*. https://doi.org/10.31234/osf.io/upj2d
2. T. S. A. Wallis & J. M. Martin. (2024, August 16). No evidence that late-sighted individuals rely more on color for object recognition: Reply to Vogelsang et al. *OSF Preprints*. https://doi.org/10.31234/osf.io/sv4pw
3. M. Vogelsang et al. (2024). Impact of early visual experience on later usage of color cues. *Science*, **384**, 907-912.
4. D. Bates et al. (2015). Fitting linear mixed-effects models using *lme4*. *Journal of Statistical Software*, 67(1), 1–48. https://doi.org/10.18637/jss.v067.i01
5. P. C. Bürkner. (2018). Advanced Bayesian multilevel modeling with the R package *brms*. *The R Journal*, **10**, 395-411. https://doi.org/10.32614/RJ-2018-017
6. B. Carpenter et al. (2017). Stan: A probabilistic programming language. *Journal of Statistical* Software, **76**. https://doi.org/10.18637/jss.v076.i01
7. A. L. Ostergaard & J. B. Davidoff. (1985). Some effects of color on naming and recognition of objects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **11**, 579–587.
8. I. Biederman & G. Ju. (1988). Surface versus edge-based determinants of visual recognition. *Cognitive Psychology*, **20**, 38–64.
9. F. A. Wichmann et al. (2002). The contributions of color to recognition memory for natural scenes. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **28**, 509–520.

Supplementary Material

The code for all the statistical analyses reported here can be found at a Github repository under the following URL (https://github.com/ag-perception-wallis-lab/vogelsang_science_reply).