No evidence that late-sighted individuals rely more on color for object recognition: Reply to "Impact of early visual experience on later usage of color cues"

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: Vogelsang et al. (*1*) claim that individuals treated for congenital blindness via cataract removal surgery (Prakash patients) rely more on color cues for object recognition than age-matched controls. The evidence presented for this claim is based on an inappropriate statistical analysis of proportion data in a recognition task. We show that a variety of more suitable analyses provide, if anything, slight evidence in favor of the null hypothesis that patients and controls are similarly impaired by the removal of color information in an object recognition task.

Vogelsang et al. (*1*) claim that individuals treated for congenital blindness via cataract removal surgery (Prakash patients) rely more on color cues for object recognition than age-matched controls. The evidence for this claim is based on the result of an independent samples t-test, which found a significant difference in the group means of a "color > gray benefit measure", calculated by subtracting the percent of correctly identified grayscale images from the percent of correctly identified color images for each participant. However, we find that this comparison is undermined by a ceiling effect in the control data (see Figure 1A). Additive reasoning about percentage changes near the bounds is difficult because the same percentage change has different meaning in terms of the odds of success (*2*). This also means that the statistical tests the authors use are inappropriate, since they assume Gaussian error distributions (*3*, *4*).

One way to mitigate this issue is to first transform the data and then conduct inferential tests. Here we compute tests on ranked data and proportions converted to the log odds scale (*2*, *3;* see Supplementary Material for the complete analyses). Using these transformations, we repeated the t-test critical to the study's claim that the effect of image color on recognition differs between Prakash patients and controls, (see Figure 1B, C; for log-odds we converted 100% scores to 99% to avoid infinity). We found that the reported effect was no longer significant (rank data: $t(18) = 0.843$, $p = 0.41$; log odds: $t(18) = 0.317$, $p = 0.75$). In fact, Bayes Factors suggest weak evidence for the null hypothesis (rank data: $BF10 = 0.51$; log odds: $BF10 = 0.41$).

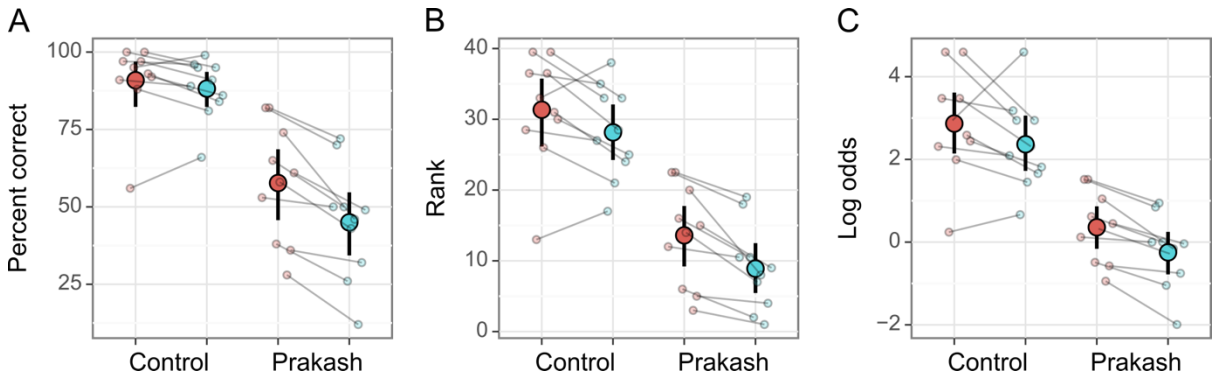

Fig 1. Recognition performance for color (red) and grayscale (blue) images, for patients and controls, using three measurements. (A) Percent correct, as in Vogelsang et al. **(B)** Rank-ordered. **(C)** Log odds (logit). Small points show individual participant data. Larger points show mean; error bars show 95% bootstrapped confidence intervals.

A more comprehensive and accepted statistical approach to handle proportion data (e.g. see 5, chapter 11) is to perform a regression with a binomial error model (e.g. logistic or probit regression). In this case, we model each trial separately as a binomial outcome (either a success or failure in identifying an image). One advantage of this approach compared to the data transformation plus test approaches is that these regressions take the uncertainty about each proportion into account. For example, a participant who has a proportion correct of 0.9 would be treated the same in the transform-plus-test approach, whether that proportion was based on 9/10 or 900/1000 successes. However, in the binomial regression case, the model has more certainty in the latter case, since the total number of trials contribute to the likelihood of the data.

Given that we have hierarchical structure (multiple trials from individual participants from which we wish to make an inference about the population average), an appropriate modelling approach is to use a mixed-effects extension of a logistic regression (i.e. a Generalized Linear Mixed Effects model). This models the probability of the outcome as a linear combination of the predictors (image type, group, and the interaction between image type and group), while the 'mixed effects' take into account the correlated residuals present in the data (i.e. that each participant contributes a score for the gray and color conditions, which are correlated). When we run this model, we once again find no significant difference for the interaction effect, suggesting that the groups do not significantly differ in their recognition performance of gray vs. color images ($\chi 2(1) = 1.95$, $p = 0.162$; see Supplementary Material for more details). Furthermore, indices of model fit (AIC, BIC) suggest, if anything, slight evidence in favor of the null model. Consistent with the rank and logit transform approaches above, this analysis suggests that the data in Vogelsang et al (2024) provide no evidence in favor of the claim of a group difference.

The empirical data presented by Vogelsang et al. (*1*) therefore do not support the study's core claim motivating the subsequent hypothesis and neural network modelling. Rather, until further evidence is provided, we should accept the null hypothesis that Prakash patients and controls are equally affected by color removal. We believe that it is particularly important to correct the scientific record in this instance not only due to the implications of this study for understanding visual development, but also because of the potential for the study's results to inform clinical practice in treating ophthalmic disorders.

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 Pawan Sinha and Lukas Vogelsang for their comments on an earlier version of this article, and for making their dataset available. Thomas Wallis and Joshua Martin contributed equally to this work. The order of authorship was determined by a coin toss.

References and Notes

1. M. Vogelsang et al., Impact of early visual experience on later usage of color cues. *Science*, **384**, 907-912 (2024).
2. J. D. Emerson, "An Introduction to Transformation" in *Fundamentals of Exploratory Analysis of Variance*, D. C. Hoaglin, F. Mosteller, J. W. Tukey, Eds. (Wiley, 1991), pp. 365-400.
3. D. I. Warton, & F. K. C. Hui, The arcsine is asinine: The analysis of proportions in ecology. *Ecology*, **92**, 3–10 (2011).
4. M. Šimkovic, & B. Träuble, Robustness of Statistical Methods When Measure Is Affected by Ceiling and/or Floor Effect. *PLOS ONE*, **14***, e0220889* (2019).
5. R. McElreath, "God Spiked the Integers" in *Statistical rethinking: A Bayesian course with examples in R and Stan*, (Taylor and Francis, ed. 2, 2020), pp. 323-366.

Supplementary Material

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