Reply to: Failure to replicate a superiority effect in crowding

Guido Marco Cicchini [1], Giovanni D'Errico [1] & David Charles Burr [1,2] ✉

It is unclear to us why Ozkirli et al.[1] fail to replicate our findings[2] of higher precision in orientation judgments in the presence of flankers. The authors[1] suggest that the use of different groups for the two conditions may have weakened our findings, as thresholds in orientation tasks can vary across participants. They further suggest that using different stimuli for the two conditions may introduce spurious effects. Our original choice was to avoid the repeated presentation of unflanked stimuli of similar orientation could induce response stereotyping[3,4]. However, we do acknowledge that there may have been an issue, and we have replicated the main effect of improved precision with similar target and flankers, in three ways, twice in the orientation domain and once testing a different attribute, color.

Firstly, we tested seven new participants on the two crucial conditions: target alone, and flankers of the same orientation as the target, as this condition should maximize integration. The task was orientation reproduction as in our original paper and Ozkirli's, the target was broad (low reliability) and the flankers narrow (high reliability). The navy symbols of Fig. 1a show the results, plotting thresholds measured with the flankers against those measured without, for all participants. All except one show lower average thresholds in the flanker condition, as we originally reported. The effect is significant: mean unflanked std 8.0 ± 0.6, flanked 6.2 ± 0.5, a reduction of 22% (statistics for the ratio against 1: $t(6) = 2.8$, $p = 0.029$, Cohen's $d = 1.16$, 95% CI = [0.62, 0.97], $\beta_{power} = 0.66$, two tails). This reinforces our original results with a new set of naïve participants.

This result reinforces our previous study, but there is an obvious limitation. While measuring only this condition maximizes the expected integration of the target and flankers, when the flankers have the same orientation as the test, they are fully informative about the target orientation, so simply attending to them, rather than the target, would improve performance. We, therefore, extended the replication with a range of flankers, like the original study. We also used a different psychophysical technique: two alternative forced choice (2AFC) judgments, rather than reproduction. With this technique—often considered the gold standard in psychophysics—nine participants make a simple judgment of whether the target seemed clockwise or counterclockwise from a memorized standard. As before, targets had low reliability and flankers had high reliability. Relative flanker orientation ranged from

−75° to 90° (15° steps), fully randomized, together with the unflanked target, within the same session. The target orientation straddled the learned standard (45°), ranging from 15° to 75°. For each flanker condition, the responses are plotted as a function of target angle, to yield psychometric curves. Figure 1b shows the curves of the *aggregate* participants (pooling all participants). The data of each condition have been fitted with cumulative gaussians, whose median estimates the bias introduced by the flankers (point of subjective equality (PSE)) and standard deviation of the threshold (also termed just noticeable difference: JND). We performed this operation separately for all participants, then averaged the PSEs and JNDs for each flanker orientation (ref. 5 gives more details of the general procedure).

Figure 1c and d show how the average PSEs and JNDs vary with flanker orientation. The PSEs show the same pattern of biases we observed previously for crowding[2], very similar to the pattern of biases reported for serial dependence paradigms[6–8]. JNDs also vary systematically with flanker orientation, showing an "M" pattern (as Ozkirili et al.[1] reported). Crucially for the current discussion, the flanked JNDs are lower than the unflanked JNDs at 0° and 15° and identical at −15°. The individual results for the 0° flankers for all participants are shown by the blue symbols of Fig. 1a. Average JNDs were 10.3 ± 1.5 for the unflanked and 7.8 ± 1.0 for the same orientation flanker condition, corresponding to a significant reduction of 21% in JND ($t(8) = 3.15$, $p = 0.013$, Cohen's $d = 1.11$, 95% CI = [0.64, 0.94], $\beta_{power} = 0.79$, two tail). The improvement brought by identical flankers can also be appreciated by looking at the total root mean squared errors (RMSE), which considers both bias and scatter errors (their Pythagorean sum). This is shown in brown in Fig. 1d. In our experiment, the RMSE curve closely follows the JND curve because biasing errors are generally smaller than scatter errors, so in a Pythagorean sum, the latter dominates. Even with this metric, which combines bias and scatter errors, the condition with identical flankers has a lower total error than the unflanked condition (9.6 ± 1.1 vs. 14.3 ± 1.1, $t(8) = 4.9$, $p = 0.001$, Cohen's $d = 1.73$, 95% CI = [0.53, 0.83], two tail). However, it is worth noting that while the crucial conditions where a benefit is predicted are replicated, dissimilar flankers show rather broad psychometric curves, leading to an M-shaped pattern. It is unclear why the M-shaped pattern (which implies suboptimal integration) emerges here, but not in our original report.

[1]Institute of Neuroscience, CNR, via Moruzzi, 1, 56124 Pisa, Italy. [2]Department of Neurosciences, Psychology, Drug Research and Child Health, University of Florence, viale Piercaccini, 6, 50139 Firenze, Italy. ✉e-mail: davidcharles.burr@unifi.it
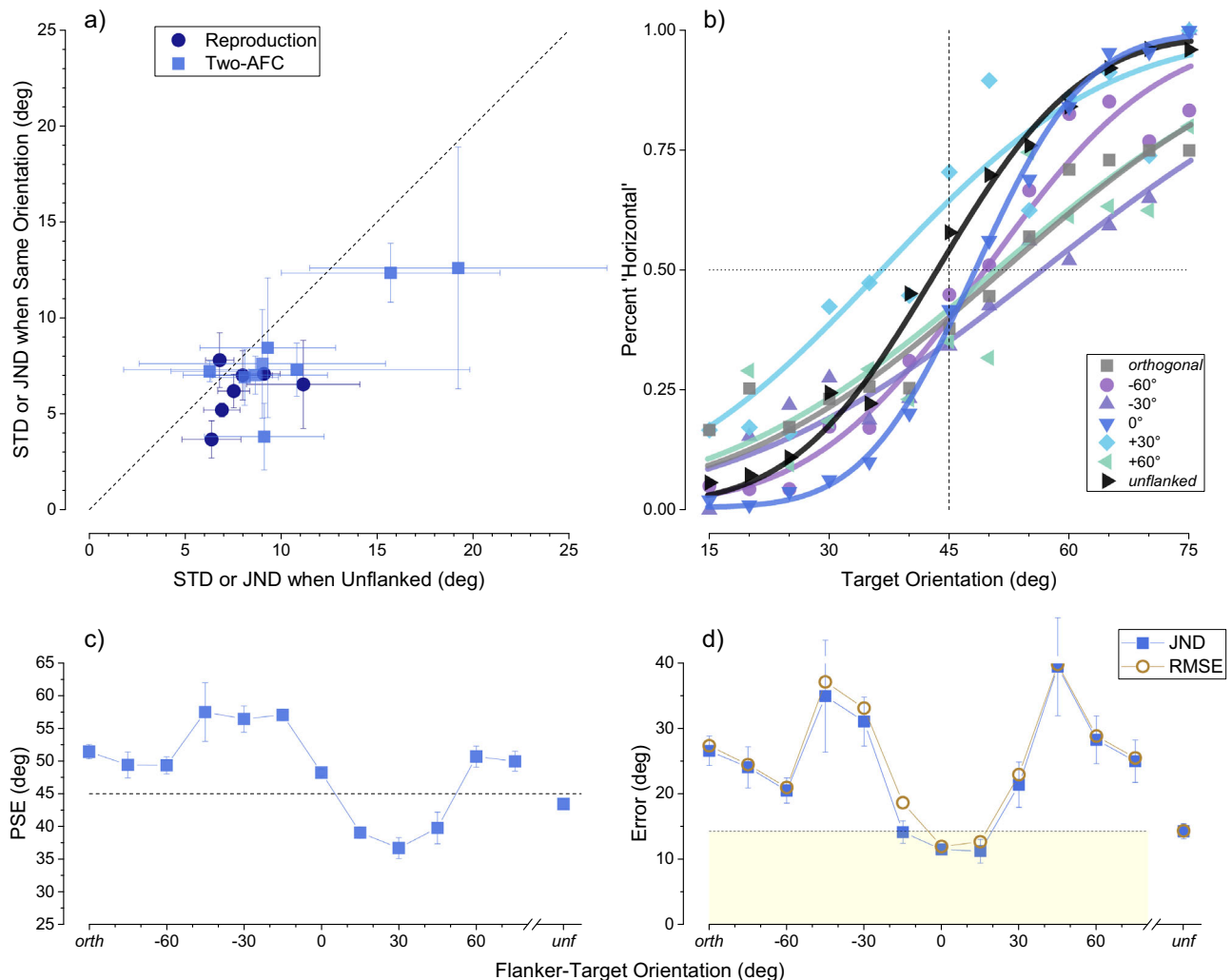
Fig. 1 | Replications with orientation judgments with low-reliability targets.
a Navy symbols show the dispersion of reproductions of orientation (i.e. standard deviation) of broad oval stimuli (like those used in the original experiment—7 new participants), blue symbols thresholds from 2AFC judgments (9 participants). Thresholds (std or jnd) for targets flanked with ovals of the same orientation are plotted against those for unflanked targets. That most points lie below the equality line shows that the thresholds for the flanked targets were lower than for the unflanked. Reproduction precision was calculated as the standard deviation of the reproductions at four possible orientations (±35° and ±55°), with each participant completing at least 28 trials per condition. For 2AFC judgments precision was derived from the standard deviations of gaussian fits to psychometric functions like those of panel **b** (at least 110 trials per condition). **b** Psychometric curves for the aggregate participant (pooling data from all 9 participants), plotting the proportion of trials judged more clockwise than the 45° standard against the orientation of the

target. Data are fitted by cumulative gaussian error functions. Black symbols and curve show data for the target alone; plum, violet, blue, celeste, and teal the flankers −60°, −30°, 0°, +30°, and +60° away from the target, gray orthogonal flankers. **c** Point of subjective equality (PSE) for the 2AFC judgments (aggregate participant) as a function of flanker–target orientation difference from −90° to +75° where negative values refer to flankers oriented counterclockwise of the target. PSE at 45° (dashed) indicated a veridical observer. Error bars are SEM calculated via bootstrap. **d** Two measures of error as a function of relative flanker orientation in the same dataset of (**c**): just noticeable differences (JND) in blue derived from psychometric functions, root mean square error (RMSE) in brown obtained by Pythagorean sum of biasing errors (PSE−45°) and scatter errors (JND), again for the aggregate participant. Horizontal dashed line indicates the JND in the unflanked condition and the yellow region indicates where performance is better than unflanked.

Having replicated our results with orientation crowding using two different techniques, we asked whether improved precision in crowding may generalize to different stimulus features, such as color. Nine participants judged, again in 2AFC, whether a patch of variable color seemed pinker or greener than a memorized standard (see ref. 5 for full details). The resulting psychometric functions of a typical participant are shown in Fig. 2a, for the various conditions: the target alone (black symbols), gray flankers (gray symbols), or with color varying to be a constant distance (from −72° to +72°) from the target in DKL color space[9] (shown by the colored symbols and fitted curves). The resulting psychometric curves are systematically displaced away from the standard and are also steeper, implying lower precision thresholds. Gray flankers provide no such advantage. The average

results of all participants are summarized in Fig. 2c, showing the systematic shifts in PSE as flanker color varies, and Fig. 2d, showing the decrease in precision thresholds for flankers similar in color to the standard.

Total root mean squared error (RMSE) comprises two orthogonal components, average accuracy and precision: accuracy is given by PSE relative to the standard (plotted on the abscissa of Fig. 2e) and precision by the standard deviation of the psychometric functions (on the ordinate). RMSE is the Pythagorean sum of the accuracy and precision, corresponding to the distance from the origin. RMSE is lowest for the flankers that are most similar to the target, and highest for the neutral gray flankers (gray symbol). The no-flanker condition (black symbols) has higher RMSEs than the two of the central color-flanker conditions
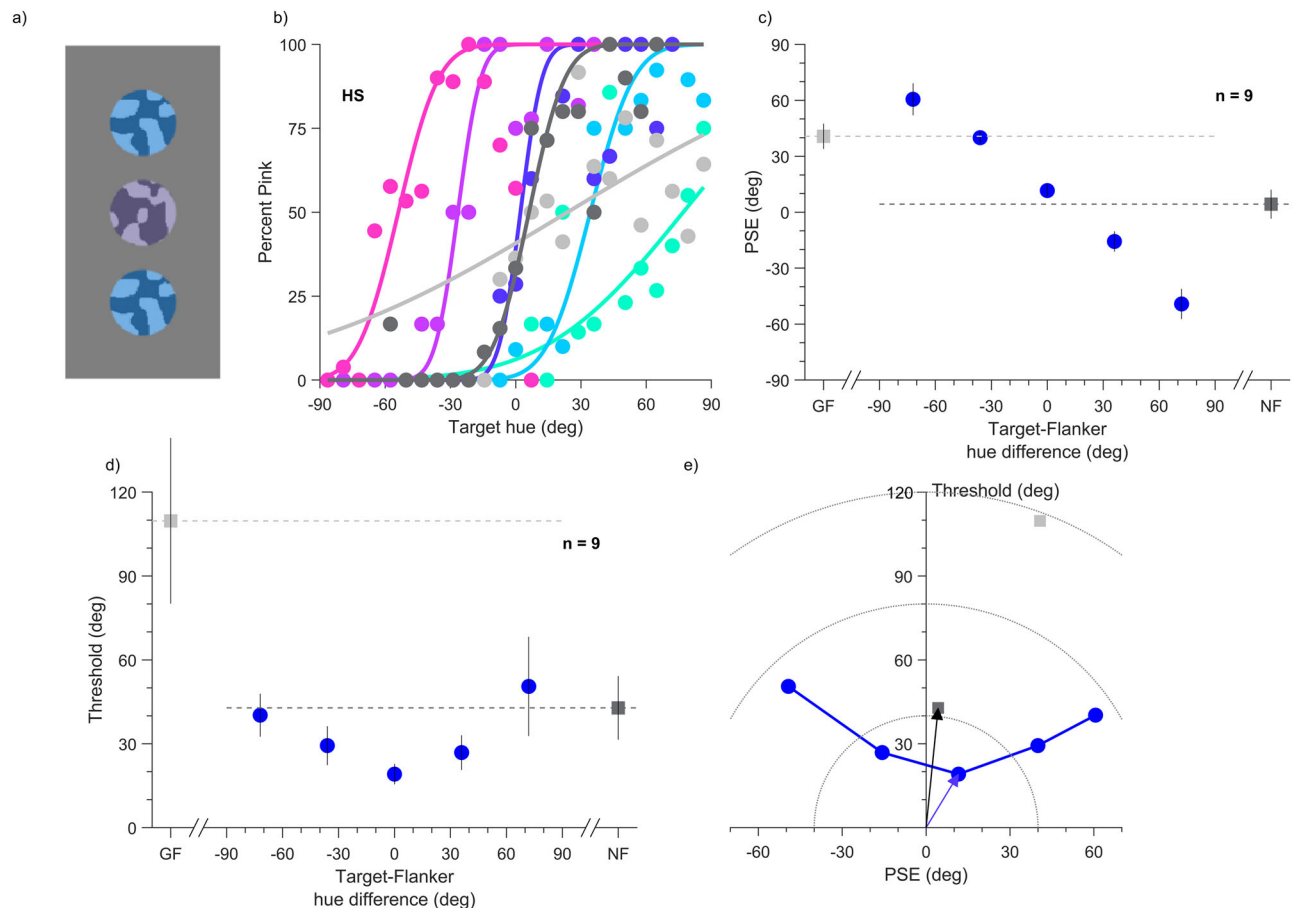
Fig. 2 | Summary of results of color discrimination with colored flankers[5].
a Example of the low-saturation target stimulus with the hue of the standard (263°
in DKL color-space[9]), with flankers −36° relative to the standard. **b** Psychometric
curves for a representative participant, plotting the proportion of trials judged
"pinker than the standard" against the hue of the target, expressed in as angle in
DKL color-space relative to the standard. Data are fitted by cumulative gaussian
error functions. Black symbols and curve show data for the target alone, gray for
gray flankers, and green, blue, navy, mauve, and pink for the colored flankers,
respectively −72°, −36°, 0°, +36° and +72° away from the target. **c** Average points of
subjective equality (PSEs) as a function of flanker hue, relative to the violet standard
of 263°. Black symbol shows the unflanked condition and gray symbol shows the
gray-flanker condition. Error bars are SEM. **d** Average precision thresholds (defined
as hue difference between the 0.5 and 0.84 points of the psychometric functions)
as a function of flanker hue. Error bars are SEM. **e** The data of panels **c** and **d**, plotted
as threshold (precision) against PSE (accuracy). Total RMSE is given by the distance
from the origin, shown by the arrows for two conditions (equal flankers and
unflanked).

(−36° and 0°). The results are entirely consistent with our orientation
dataset, not only reinforcing the original claims but extending them to
other visual tasks.

At first blush, the results reported here may seem at odds with a
recent study by Greenwood and Parsons[10], who also measure preci-
sion as the width of the psychometric curves, but find that flankers
broaden the curves. However, there were major differences in the
experiment. We varied flanker hue as a fixed difference from the
target, while their flankers had a constant hue (more akin to our gray-
flanker condition). As we show in detail in ref. 5, this difference
explains completely the different results in the two experiments:
replotted to mimic Greenwood and Parsons' paradigm, our results are
very similar to theirs.

Ozkirli et al. [1] suggest that their data and, more in general, these
Bayesian-like effects in crowding can be interpreted in the light of a
probabilistic substitution of target and flankers: on a fraction of the
trials the flankers "substitute" the target, so the average orientation will
be between the two. If this mechanism were augmented by a
mechanism that assesses the similarity of target and flankers, indeed it
can make very close predictions to a Bayesian observer. There is,
however, a crucial aspect that this theory cannot capture: when the
reliability of targets and flankers is reversed, the probability of

substituting the target's orientation with the flanker's orientation
should be unchanged. But all the datasets on both orientation and
color[5], as well as Ozkirli et al.'s orientation data, show that there is
much less influence from the flanker when the target is reliable. This
squares perfectly with a Bayesian framework, but it is difficult to
reconcile with a substitution theory of crowding unless further rules
are added. Similarly, the substitution theories find difficulty in
explaining the improved performance when target and flankers are
identical, one of the conditions we detail elsewhere[5]: substitution
changes nothing.

It is worth noting that our participants do not behave exactly like
an ideal participant. Our original publication reported behavior in the
predicted direction but only half as efficient as predicted. Similar
conclusions can be drawn from the current datasets. However, with the
new orientation data reported here, and the color experiment, we find
that some flanker conditions lead to a worsening of performance,
which is incompatible with a strict Bayesian approach. Thus, it is
possible that the effects are the results of a flexible integrator, broadly
compliant with Bayesian design rules: beneficial in key conditions, but
not rigidly fixed.

We have proposed an alternative explanation to crowding[2], at
odds with the commonly accepted view that crowding results from

fundamental limitations or bottlenecks in processing[11]: but not at all incompatible with the ideas of Herzog and his colleagues[12,13]. We suggest that crowding is an unwelcome by-product of otherwise efficient mechanisms that exploit spatial redundancies of the visual scene, in an analogous way that they exploit temporal redundancies in serial dependence[6,8]. While this is clearly disadvantageous for many modern tasks such as reading, the improvement in precision may be more beneficial in other natural tasks, such as texture discrimination. Our data, in both orientation and color discrimination, provide strong support for this alternate approach. That very similar results hold for both orientation and color judgments suggests that reliability-based spatial integration may be a general property of peripheral vision.

Methods

Participants
Sixteen participants were recruited for the experiments, 9 females and 7 males. Two were authors; all the others were naïve to the purposes of the experiment. All participants had normal color vision (assessed by the Ishihara color blindness test) and normal or corrected-to-normal vision. The participants were aged from 21 to 46 years at the time of measurement. Experimental procedures are in line with the declaration of Helsinki and approved by the local ethics committee (Commissione per l'Etica della Ricerca, University of Florence, July 7, 2020, no. 111). Written informed consent was obtained from each participant, which included consent to process, preserve, and publish the data in an anonymous form.

Orientation reproduction
This replication study followed closely that of Cicchini et al.[2] employing the same stimuli and setup, albeit with a new set of participants. In brief, participants had to reproduce the orientation of a peripheral target oval defined by a set of 12 small dots presented briefly on the screen at 26° horizontal eccentricity. The target was flanked by two similar ovals, which, however, had a different aspect ratio. To speed-up data collection in this replication series, we only tested the condition low-reliability target (Target aspect ratio 1.4, Flanker aspect ratio 2.8), and we focused only on the Identical-flankers condition and the no-flanker conditions. In an attempt to replicate our original data but also to respond to the criticism of Ozkirli et al.[1] we employed in all conditions the same target orientations. To prevent response stereotyping, we intermingled orientations of 35–55° clockwise and counterclockwise from the vertical. About 130 reproductions were collected for each participant.

2 AFC orientation judgments
This new experiment used the same stimuli and setup of the orientation reproduction experiment (i.e. low-reliability target), but the paradigm and task were different. Participants again watched a briefly presented peripheral oval surrounded by two other ovals, but they were requested to judge simply as to whether the target oval was more vertical or horizontal than the 45° diagonal. Target orientation could vary adaptively between 5° and 85° from the vertical (see Fig. 1b for a sample psychometric function). Several flanker conditions were employed from −75° to +90° difference from the target along with the unflanked condition; all intermingled within the same session. Judgments of horizontal targets were plotted as a function of target orientation to yield psychometric curves whose PSE indicates a bias and whose JND is a correlate of sensory precision. About 800 trials were collected for each participant.

2 AFC color judgments
This experiment is described in full elsewhere[5] and is summarized here briefly. Participants judged the color of a peripheral (18° vertical eccentricity) circular cowhide patch which could have any color hue ranging from green to purple (176° to 349° hue in DKL color space),

and had to judge it with respect to a previously learned periwinkle standard (263° hue), reporting whether it appeared greener or pinker than the standard. The target patch could be flanked by similar cowhide patterns, which could differ in hue from the target by −72°, −36° (greener than the standard), 0° (identical), +36°, and +72° (pinker than the standard); or could be gray or absent altogether. Here we show data of the low purity target condition where the target had a purity of 0.12 and the flankers 0.34. A minimum of 720 trials were collected for every participant.

Reporting summary
Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability
Raw trial-by-trial data for the current research are available on the Zenodo Platform. Data for the orientation judgments is published here[14] https://doi.org/10.5281/zenodo.13361493, whereas the full dataset for the color experiments is published here[15]: https://doi.org/10.5281/zenodo.10782525.

Code availability
The original Matlab code used for the three experiments will be provided upon request to the corresponding author.

References
1. Ozkirli, A., Pascucci, D. & Herzog, M. H. Failure to replicate a superiority effect in crowding. *Nat. Commun.* https://doi.org/10.1038/s41467-025-56762-5 (2025).
2. Cicchini, G. M., D'Errico, G. & Burr, D. C. Crowding results from optimal integration of visual targets with contextual information. *Nat. Commun.* **13**, 5741 (2022).
3. Cicchini, G. M., Arrighi, R., Cecchetti, L., Giusti, M. & Burr, D. C. Optimal encoding of interval timing in expert percussionists. *J. Neurosci.* **32**, 1056–1060 (2012).
4. Jazayeri, M. & Shadlen, M. N. Temporal context calibrates interval timing. *Nat. Neurosci.* **13**, 1020–1026 (2010).
5. Cicchini, G. M., D'Errico, G. & Burr, D. C. Color crowding considered as adaptive spatial integration. *J. Vis.* **24**, 9 (2024).
6. Cicchini, G. M., Mikellidou, K. & Burr, D. C. Serial dependence in perception. *Annu. Rev. Psychol.* https://doi.org/10.1146/annurev-psych-021523-104939 (2023).
7. Fischer, J. & Whitney, D. Serial dependence in visual perception. *Nat. Neurosci.* **17**, 738–743 (2014).
8. Pascucci, D. et al. Serial dependence in visual perception: a review. *J. Vis.* **23**, 9 (2023).
9. Derrington, A. M., Krauskopf, J. & Lennie, P. Chromatic mechanisms in lateral geniculate nucleus of macaque. *J. Physiol.* **357**, 241–265 (1984).
10. Greenwood, J. A. & Parsons, M. J. Dissociable effects of visual crowding on the perception of color and motion. *Proc. Natl Acad. Sci. USA* **117**, 8196–8202 (2020).
11. Levi, D. M. Crowding—an essential bottleneck for object recognition: a mini-review. *Vis. Res.* **48**, 635–654 (2008).
12. Herzog, M. H., Sayim, B., Manassi, M. & Chicherov, V. What crowds in crowding? *J. Vis.* **16**, 25 (2016).
13. Tiurina, N. A., Markov, Y. A., Choung, O. H., Herzog, M. H. & Pascucci, D. Unlocking crowding by ensemble statistics. *Curr. Biol.* **32**, 4975–4981.e4973 (2022).
14. Cicchini, G. M. Dataset for 2AFC experiment in correspondence Ozkirli-Cicchini. *Zenodo* https://doi.org/10.5281/zenodo.13361494 (2024).
15. Cicchini, G. M. & D'Errico, G. Dataset for colour crowding explained as adaptive spatial integration. *Zenodo* https://doi.org/10.5281/zenodo.10782525 (2024).

Author contributions

G.M.C., G.D.E., and D.C.B. drafted the initial version of the manuscript and planned experiments. G.M.C. and G.D.E. programmed the experiments, collected the data, analyzed the results, and produced the final figures. G.M.C., G.D.E., and D.C.B. reviewed the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at
https://doi.org/10.1038/s41467-025-56763-4.

Correspondence and requests for materials should be addressed to David Charles Burr.

Peer review information *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work.

Reprints and permissions information is available at
http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.