Comprehensive characterization of human color discrimination thresholds

Fangfang Hong[1], Ruby Bouhassira[1], Jason Chow[2], Craig Sanders[2], Michael Shvartsman[3], Phillip Guan[2,*], Alex H. Williams[4,5,*], and David H. Brainard[1,*]

This manuscript was compiled on July 16, 2025

Discrimination thresholds reveal the limits of human perception; scientists have studied them since the time of Fechner in the 1800s. Forced-choice psychophysical methods combined with the method of constant stimuli or parametric adaptive trial-placement procedures are well-suited for measuring one-dimensional psychometric functions. However, extending these methods to characterize psychometric fields in higher-dimensional stimulus spaces, such as three-dimensional color space, poses a significant challenge. Here, we introduce a novel Wishart Process Psychophysical Model (WPPM) that leverages the smooth variation of threshold across stimulus space. We demonstrate the use of the WPPM in conjunction with a non-parametric adaptive trial-placement procedure by characterizing the full psychophysical field for color discrimination in the isoluminant plane. Each participant (N = 8) completed between 6,000 and 6,466 three-alternative forced-choice (3AFC) oddity color discrimination trials. The WPPM was fit to these trials. Importantly, once fit, the WPPM allows readout of discrimination performance between any pair of stimuli, providing a comprehensive characterization of the psychometric field. In addition, the WPPM readouts were validated for each participant by comparison with 25 probe psychometric functions. These were measured with an additional 6,000 trials per participant that were held out from the WPPM fit. The dataset offers a foundational resource for developing perceptual color metrics and for benchmarking mechanistic models of color processing. This approach is broadly generalizable to other perceptual domains beyond color.

Color discrimination threshold | Wishart random process | Non-parametric adaptive sampling | Isoluminant plane

Significance Statement

Color discrimination thresholds—the smallest detectable color changes—provide a benchmark for models of color vision, guide theories of color's perceptual representation, and inform the design of display technologies. Despite their importance, comprehensive characterization of color discrimination thresholds has been considered intractable due to the extensive time required from participants. In this study, we leverage advances in semi-parametric modeling and a non-parametric adaptive trial-placement procedure to fully characterize color discrimination in the isoluminant plane. Our results enable readouts of discrimination performance for any pair of stimuli across color space and provide a novel foundational dataset. Moreover, our framework generalizes beyond color, offering a powerful approach for characterizing perception in any domain where performance varies smoothly across the underlying stimulus space.

Measurements of discrimination threshold—the smallest detectable stimulus change—are foundational for understanding biological vision. Threshold measurements support inferences about the neural mechanisms mediating performance (1, 2), guide the design of displays and specification of perceptual tolerances (3, 4), allow quantitative evaluation of diseases (5–8), inform models of suprathreshold perceptual representations (9–11), and allow perceptual effects to be incorporated into the study of cognitive processes (12–14). Modern psychophysical methods (15, 16) provide rigorous quantification of thresholds, and the theory of signal detection (17–19) provides a mature framework for relating thresholds to the precision of the underlying representation.

Despite the central role of perceptual thresholds, comprehensive characterization of thresholds has largely been limited to individual stimulus dimensions. For example, pedestal functions characterize contrast discrimination threshold across varying baseline contrasts (20). To generalize threshold characterization beyond a single dimension, we introduce the concept of the *psychometric field*: a multidimensional function that specifies the probability of a particular perceptual response as a joint function of both a reference and a comparison stimulus. In contrast to the psychometric function, which describes response probability as a function of variation around a fixed reference, the psychometric field captures how discrimination performance varies across all combinations of reference and comparison stimuli in a stimulus space. As the dimensionality of the psychometric field increases, the number of trials needed to tile the field grows exponentially—a psychophysical curse of dimensionality.

In this study, we focus on human color discrimination thresholds. Despite their significances and applications described above, fully characterizing human color discrimination—even on a single planar slice—has long been considered impracticable (21). This is because, although the stimulus space itself is two-dimensional, the underlying psychometric field is four-dimensional, as both the reference and comparison stimuli vary along two color dimensions. Mapping this field requires estimating discrimination performance across a densely sampled set

Author affiliations: [1]Department of Psychology, University of Pennsylvania; [2]Reality Lab Research, Meta; [3]FAIR, Meta; [4]Center for Neural Science, New York University; [5]Center for Computational Neuroscience, Flatiron Institute

Conceptualization: F.H., P.G., A.H.W., D.H.B.; Formal analysis: F.H.; Funding acquisition: D.H.B.; Investigation: F.H., R.B.; Methodology: F.H., J.C., C.S., M.S., P.G., A.H.W., D.H.B.; Project administration: R.B., D.H.B.; Resources: P.G., A.H.W., D.H.B.; Software: F.H., A.H.W., J.C., C.S., M.S., P.G., D.H.B; Supervision: P.G., A.H.W., D.H.B.; Validation: F.H., P.G., A.H.W., D.H.B.; Visualization: F.H.; Writing - original draft: F.H., A.H.W., D.H.B.; Writing - review & editing: F.H., M.S., P.G., A.H.W., D.H.B.

The authors declared no competing interests.

* These authors contributed equally to this work

[1]E-mail: fh862@sas.upenn.edu

of reference stimuli, with multiple comparison stimuli tested at each. The number of required trials quickly becomes intractable using conventional methods such as the method of constant stimuli (MOCS). While adaptive trial-placement procedures (22, 23) can greatly improve sampling efficiency, they typically rely on certain parametric forms. In many cases—including ours—such form is not known in advance.

Here, we show that it is possible to obtain a comprehensive characterization of the color discrimination psychometric field in the isoluminant plane. We achieved this by efficiently sampling reference-comparison stimulus pairs obtained using a non-parametric adaptive trial-placement procedure (24, 25), and then fitting the data *post hoc* with a semi-parametric model that leverages the smooth variation of psychophysical performance. We collected full data sets from 8 individual participants, and for each participant, we validated the accuracy of the model readouts against independent threshold measurements from held-out validation trials. Importantly, from the model fit, we can read out the psychometric function along any chromatic direction around any reference stimulus in the plane and thus determine the discrimination threshold in that direction. Our study provides a foundational dataset that may be used to test computational and neural models of color discrimination, benchmark color metrics, and develop models of supra-threshold color discrimination.

1. Results

1.1. Overview. The Results section is organized as follows. We begin with a brief overview of the experimental stimuli and task (Section 1.2), followed by a summary of how our model characterizes the full psychometric field (Section 1.3) and a description of the non-parametric adaptive trial-placement procedure used to collect the data (Section 1.4). Having described these essential methods, we then present our core results (Section 1.5) and evaluate the validity of our model (Section 1.6). Finally, we compare our findings with previous measurements from the color discrimination literature (Section 1.7). Additional technical details are provided in Materials and Methods and the **Supplement**.

1.2. Task and stimuli. Participants (N = 8) performed a 3AFC oddity task. On each trial, three blobby stimuli were shown in a triangular spatial arrangement—two identical reference stimuli and one comparison stimulus with a different surface color (Fig. 1A). The comparison stimulus was pseudo-randomly assigned to one of the three positions within the triangular arrangement. Participants were asked to identify the odd one out. Stimuli were rendered using the Unity graphics engine, and color was controlled by varying the specified surface reflectance using RGB (red, green, blue) coordinates, with other scene aspects held constant. We used naturalistic stimuli to increase the relevance of our results for understanding color vision in the real world. Hedjar and colleagues (26) provide a comparison of color discrimination using stimuli similar to ours versus traditional flat spatially uniform patches.

We made spectral calibration measurements (27) to establish the relationship between RGB and the light emitted from the display. These measurements allowed us to represent the stimuli in terms of the excitations of the human L, M, and S cones elicited by the stimuli, and more generally in any standard color space (28–30). For this study, stimuli were constrained to lie in the isoluminant plane passing through the monitor's gray point and bounded by its gamut. This plane was then affine-transformed into a square ranging between -1 and 1 along each axis (Fig. 1B; **Supplement S1**). We refer the space in which the transformed plane is as the *model space* because it is directly related to the way we formulated our semi-parametric model, and also served as a convenient representation for the non-parametric adaptive trial-placement procedure we used (see details in Section 3.8.2).

1.3. The Wishart Process Psychophysical Model (WPPM). As an overview of our modeling approach, we fit the color discrimination responses (coded as 'correct' or 'incorrect') with a novel model, the WPPM—a Bayesian probabilistic model that combines an observer model (specified through a likelihood function) with an expectation of smoothness in the internal noise limiting color discrimination (specified through a prior distribution). Once fit to the data, the WPPM yields a continuously varying field of covariance matrices that characterize the internal noise in the perceptual representation of color stimuli (Fig. 1C-D). These covariance matrices, in turn, determine the entire psychometric field.

More specifically, we designed the observer model within the WPPM to formalize the intuition that the stimulus perceived as most distant from the other two is identified as the "odd one out". The internal representation of each stimulus is assumed to be noisy and modeled as a multivariate Gaussian with the same dimensionality as the stimulus space. We assume the mean of each distribution is given by the corresponding stimulus' location in the model space. In contrast, we allow the covariance matrices to vary across the model space to account for differences in the encoding precision of the color stimuli. On each trial, the observer model has access to one sample from the distribution of each of the three stimuli—two identical reference stimuli and one comparison. The observer model computes the pairwise squared Mahalanobis distance between each pair of noisy samples, using the weighted average of the covariance matrices of the reference and comparison (Fig. 1E). By using Mahalanobis distance to make decisions (instead of, for example, Euclidean distance), the observer accounts for the expected noise structure. The two stimuli whose pairwise distance is smallest are identified as the references, and the remaining stimulus as the comparison (the "odd one out"). Because there is no simple closed-form solution for this decision rule (31), we used Monte Carlo simulation to approximate the percent-correct performance (Section 3.8.1).

We expect the internal noise that limits color discrimination to vary smoothly across the model space—that is, small changes in the reference stimulus should produce only small changes in the corresponding internal noise. The WPPM reflects this expectation by placing a finite-basis Wishart process prior over the continuous field of covariance matrices (32). Intuitively, the Wishart process prior introduces a regularization term to the model—it penalizes overly complex or erratic patterns of the covariance matrix field by giving preference to simple, smoothly varying covariance matrix fields. As described in Methods (Section 3.8.3), a single hyperparameter, ϵ, controls the strength of this regularization (Fig. 1C-D).

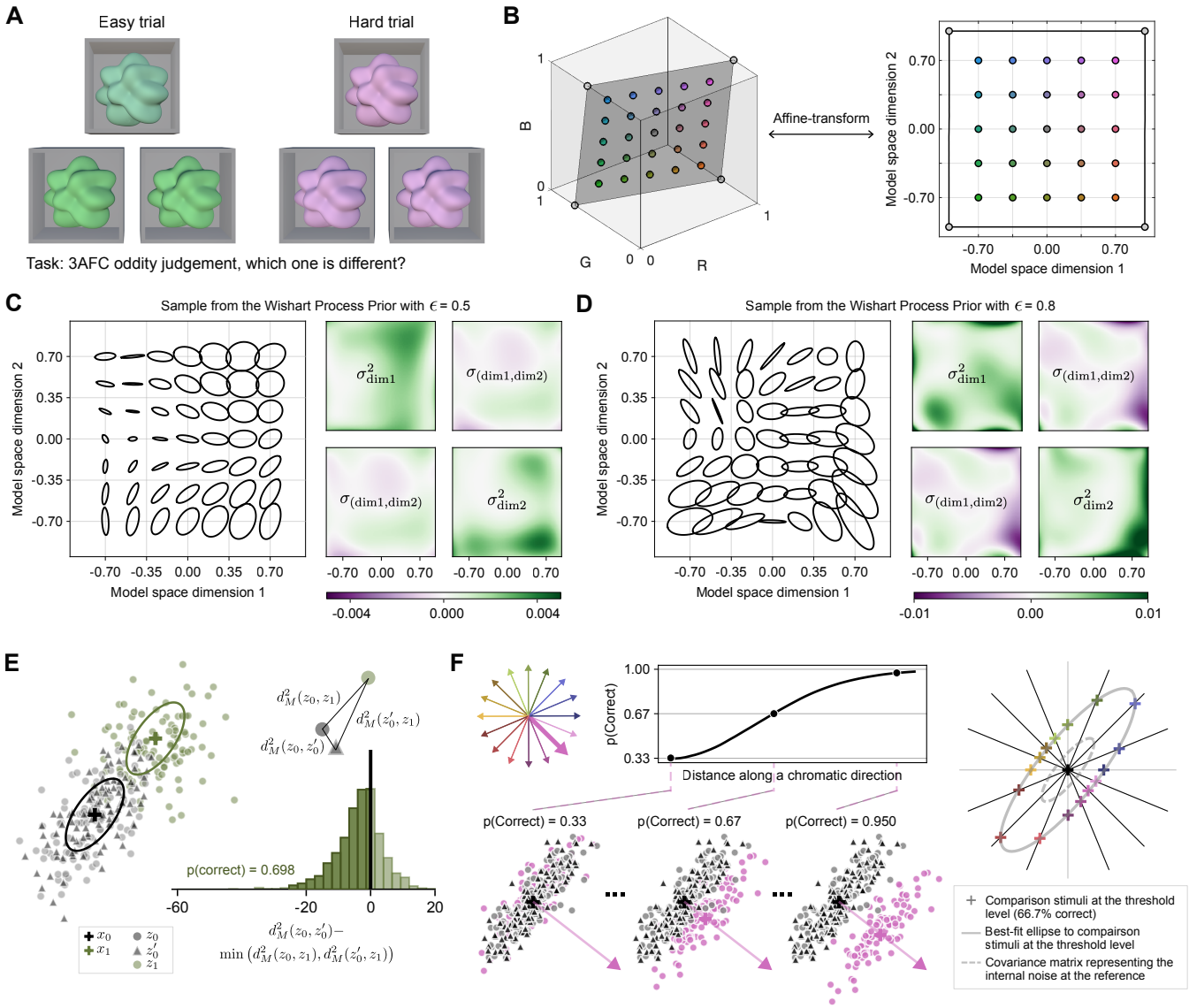

Fig. 1. Task, stimuli and the WPPM. (A) 3AFC oddity task. On each trial, participants viewed a triplet of stimuli—two identical references and one different comparison—and identified the odd one out. (B) Stimuli were constrained to lie in the isoluminant plane the display's gray point. Data were represented and fit in a transformation of this plane which we refer to as *model space*. The grid of dots illustrates the transformation between the plane in the RGB and model space. (C) Example of a smoothly varying covariance matrix field produced by the WPPM. The field was generated by sampling from a finite-basis Wishart random process with a smooth prior ($\epsilon = 0.5$; see Section 3.8.3). Although the field is illustrated on a 7×7 grid, it specifies a covariance matrix $[\sigma^2_{\text{dim1}}, \sigma_{(\text{dim1},\text{dim2})}; \sigma_{(\text{dim1},\text{dim2})}, \sigma^2_{\text{dim2}}]$ for every stimulus in the plane, as shown in the heat maps. (D) Example of a less smoothly varying covariance matrix field. This field was obtained by drawing from a finite-basis Wishart random process with a less smooth prior ($\epsilon = 0.8$). (E) Observer model. For each stimulus triplet $[x_0, x_0', x_1]$, internal representations $[z_0, z_0', z_1]$ are drawn from multivariate Gaussian distributions centered at the reference stimulus with noise characterized by the corresponding covariance matrices. The model determines whether the observer correctly identifies the odd stimulus by comparing the squared Mahalanobis distances d^2_M between the representation pairs. (F) Derivation of elliptical threshold contour. One-dimensional psychometric functions are approximated using Monte Carlo simulations (10,000 samples per stimulus pair shown for illustration; 2,000 used during model fitting). For each selected chromatic direction, we derive the threshold distance corresponding to 66.7% correct. An ellipse is then fit to the threshold distances to describe the discrimination threshold contour. Notably, while the threshold contour and the noise ellipse shown at one standard deviation have well-matched shapes, the threshold contour is larger because it corresponds to 66.7% correct.

To fit the model to each participant's data, we found the *maximum a posteriori* (MAP) estimates of the WPPM's parameters, using gradient-based numerical optimization of the log posterior density (i.e., the sum of the log prior density and log likelihood function; Section 3.8.4).

The best-fit model parameters, together with the observer model, allow us to read out percent-correct performance for any pair of reference and comparison stimuli. In particular, to read out a one-dimensional psychometric function, we select a reference stimulus and use the observer model to approximate performance as the comparison stimulus varies along a line (Fig. 1F, left panels). The threshold distance along the line is defined as the distance that yields 66.7% correct. By repeating this process across many directions, we derive a set of threshold distances around the reference (Fig. 1F, right panel). Given our assumption that internal noise follows a multivariate Gaussian distribution, these threshold distances form approximately elliptical contours, and we fit ellipses to them for data representation. This approach is consistent with prior work showing that ellipses provide a

good approximation of color discrimination thresholds (3, 33–39), despite some reported deviations (40–42). Notably, while we show threshold contours corresponding to 66.7% correct for visualization, once fit, the WPPM allows us to read out the full psychometric function for any reference and chromatic direction—effectively mapping the entire psychometric field. Given that the psychometric field is derived from the underlying field of covariance matrices that characterize internal noise, the smoothness constraint imposed on the covariance matrices naturally propagates to the threshold contours and the field itself.

1.4. Adaptively sampled trials. Reference and comparison stimuli for each trial were selected using AEPsych (24), an open-source package for adaptive psychophysics. For the adaptive sampling model, we used a probit-Bernoulli Gaussian Process (GP) model (43) with a radial basis function (RBF) kernel. As with the WPPM, the GP assumes smooth variation in performance across the model space due to the RBF kernel, but unlike the WPPM, it does not impose any specific parametric form on the internal noise (or thresholds). The semi-parametric constraint—multivariate Gaussian-shaped internal noise—was introduced only when fitting the WPPM. For this reason, we describe the adaptive trial-placement procedure as non-parametric (relative to the WPPM)—acknowledging that while it incorporates some parametric assumptions, they are less restrictive than those of the WPPM. This non-parametric approach ensures that our data collection was not biased by assuming the correctness of the WPPM prior to validation.

Each participant completed 6,000 AEPsych-driven trials: the first 900 were generated using quasi-random Sobol' sampling (44) to provide a good initialization for the GP (**Supplement S2**); for the remaining 5,100 trials, the GP was updated continuously based on participants' responses, and each trial was adaptively selected to be most informative for estimating the thresholds targeted at 66.7% correct (25). See Fig. 2A for an illustration of the procedure, and Section 3.6 for more details.

Adaptive sampling with AEPsych requires solving two optimization problems: one for updating the GP model (43) and another for selecting the next trial using the Expected Absolute Volume Change (EAVC) acquisition function (25). To reduce computational time, we updated the GP model only every 20 trials. Sometimes, however, either the fitting or the trial selection process did not complete in time for the upcoming stimulus presentation. To avoid perturbing the participants' rhythm, in these cases we slotted in pre-generated fallback trials (**Supplement S3**). The number of fallback trials varied across participants, ranging from 0 to 466. These trials were included along with the 6000 AEPsych-driven trials when fitting the WPPM.

1.5. WPPM threshold estimates. For each participant, we fit the WPPM to the 6,000 AEPsych-driven trials, along with any additional fallback trials. To visualize the fits, we read out the elliptical threshold contours around a grid of reference stimuli (Fig. 2B for a representative participant). The threshold contours revealed three key regularities: (1) thresholds were lowest for references near the achromatic point defined by the background behind the blobby stimuli, (2) thresholds increased with the distance of the reference from the achromatic point, and (3) the major axes of the elliptical threshold contours were consistently radially oriented with respect to the achromatic point. These regularities were consistent with previous results in the color discrimination literature, as explained further in Section 1.7.

The data were notably consistent across participants, with all three regularities observed in every individual (Fig. 2C; **Supplement S2** for individual plots for the remaining participants). When the variability across participants was assessed as Euclidean distance in the model space representation, it was lowest near the achromatic point where sensitivity was highest and it increased with threshold magnitude. This trend has been observed in other perceptual discrimination tasks (45–47).

1.6. Validation of the WPPM. To validate the WPPM estimates, we interleaved 6,000 validation trials throughout the experiment. These trials were held out from WPPM model fitting. For each participant, we used Sobol' sampling to select 25 reference stimuli and associated chromatic directions, with a unique draw per participant. Along each sampled chromatic direction, we used MOCS to sample 12 comparison levels (Fig. 2D): 11 were evenly spaced, and one was selected to provide easily discriminable catch trials (**Supplement S4.3**). The comparison levels were selected based on a pilot dataset to account for variability in thresholds across different reference stimuli and chromatic directions (see Section 3.6 for details). Notably, we intentionally avoided densely sampling around a small number of references to minimize differential perceptual learning between the trials used for fitting the WPPM and those reserved for validation (48).

For each of the 25 validation references, we fit a Weibull psychometric function to the 240 MOCS trials collected along the sampled chromatic direction and identified the comparison stimulus corresponding to 66.7% correct (see two examples in Fig. 2F). We then used the WPPM fit (constrained using the non-overlapping set of adaptively-sampled trials) to extract elliptical contours corresponding to the 66.7% threshold level for each validation reference (Fig. 2E). Finally, to directly compare these two methods, we read out the WPPM threshold along the MOCS chromatic direction for each reference. The confidence intervals of the WPPM estimates overlapped with those from the Weibull fits in 24 out of 25 conditions for participant CH (Fig. 2E) and in 21 to 25 conditions across other participants (**Supplement S4.1**). These results demonstrate strong agreement between thresholds derived from the WPPM psychometric field and the 25 discrete MOCS validation thresholds. This strong agreement indicates that the weak Wishart process prior we imposed to favor smoothly varying threshold contours did not lead to substantial over-smoothing, as the validation thresholds were estimated independently without any such constraint.

To quantify the agreement, we performed a linear (slope only) regression between the thresholds read out from the WPPM fit and those obtained using the validation trials (Fig. 2G). The results further support agreement between the two sets of estimates (mean correlation coefficient = 0.82, range = 0.73 - 0.97). For 7 out of 8 participants, the regression slope was not significantly different from 1 (mean slope = 1.04, range = 0.97 - 1.13), indicating no systematic bias in the WPPM fits with respect to scale (Fig. 2H; see **Supplement**

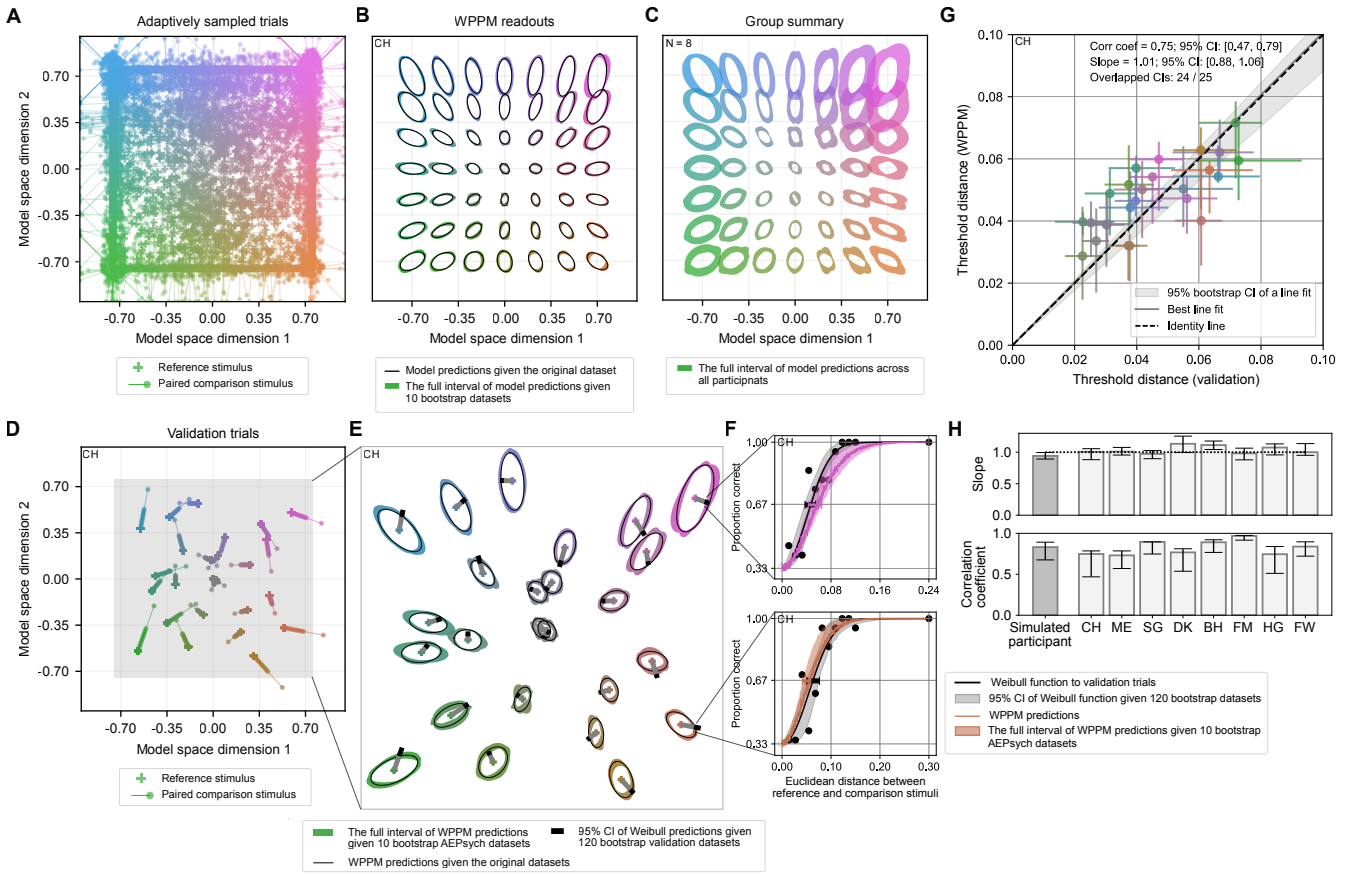

Fig. 2. Threshold results and validation. (A) Adaptively sampled trials. AEPsych-driven stimulus pairs that were most informative for estimating the entire psychometric field. Of the 6,000 trials, the first 900 were Sobol'-sampled; the remaining 5,100 (shown) were adaptively selected based on a non-parametric GP model that was updated every 20 trials and the EAVC acquisition function. (B) Discrimination threshold contours read out from the WPPM fit (66.7% correct) for a representative participant (CH), based on fits to the 6000 AEPsych trials and the fallback trials (**Supplement S2**). (C) Group summary of WPPM readouts (N = 8). Summary of regression slopes and correlation coefficients for all participants. Error bars: 95% confidence intervals. As a benchmark, the same analysis was performed on simulated data using CIELab ΔE 94 as ground truth. (D) Validation trials for the same participant. reference stimuli and chromatic directions were Sobol'-sampled uniquely for each participant. (E) Comparison of thresholds. Ellipses represent discrimination threshold contour read out from the WPPM fit (same fit as in (B)), evaluated at the 25 reference stimuli used in the validation trials. The black bars at the end of each gray line show the 95% bootstrapped confidence interval for the corresponding threshold. (F) Comparison of psychometric functions. Black lines represent the Weibull functions fit to the validation trials (black points), with 95% bootstrapped confidence intervals (gray regions). Colored lines represent the psychometric functions from the WPPM fit, with the full range of 10 bootstraps shown as colored shaded regions. (G) Linear regression of thresholds read out from the WPPM fit against validation thresholds. Horizontal and vertical error bars represent 95% confidence intervals for the validation thresholds from 120 bootstraps and the full range from 10 bootstraps of the WPPM fits, respectively. (H) Summary of regression slopes and correlation coefficients for all participants. Error bars: 95% confidence intervals. As a benchmark, the same analysis was performed on simulated data using CIELab ΔE 94 as ground truth.

S4.1 for individual participant fits). To assess whether there were more subtle sources of bias not captured by the regression slope, we analyzed the residuals—the discrepancies between the WPPM and validation thresholds. While we found no evidence that residuals depended on the orientation or shape of threshold contours read out from the WPPM fit, we did observe one small but statistically significant relation: the model slightly overestimated thresholds when validation thresholds were low and underestimated them when validation thresholds were high (slope $= -0.151$, $t(198) = -4.632$, $p < 0.001$, $R^2 = 0.098$). However, the magnitude of this bias in residuals was small (**Supplement S4.2**).

As an additional benchmark, we simulated trials and responses based on a ground-truth WPPM instance chosen to approximate the predictions from the CIELab ΔE 94 metric (**Supplement S5.1-5.5**). Using this simulated dataset, we repeated the same validation analyses described above. The thresholds read out from the WPPM fit agreed with 23 out of 25 validation thresholds, based on overlapping confidence intervals. A linear regression revealed a slope of 0.94 and a correlation coefficient of 0.83—well within the range observed in human data (Fig. 2H, left bar). Residual analysis showed a negative correlation between the residuals and the magnitude of the ground-truth validation thresholds (**Supplement S5.6**), consistent with trends observed in human participants. Finally, we assessed the accuracy of the WPPM against simulation ground truth, across a dense grid of reference stimuli in the model space. Although there were some deviations (**Supplement S5.7**), they were small relative to the inter-participant variation in thresholds (compare Figs. 2C and S15).

Taken together, these results validate the accuracy of the WPPM and highlight the remarkable efficiency of our approach. With 6000 trials, conventional psychophysical methods only allowed us to estimate percent-correct performance along one chromatic direction across 25 references.

In contrast, our new approach—combining non-parametric, adaptive trial placement with *post hoc* fitting of the semi-parametric WPPM—allowed us to map the entire psychometric field, providing the percent-correct performance for any reference-comparison stimulus pair in the isoluminant plane, using the same number of trials.

1.7. Comparison with previous measurements. The WPPM is equivariant under affine transformations of color space (**Supplement S1.4**), allowing threshold contours derived in our model space to be transformed into other colorimetric representations. This flexibility enables direct comparisons with color discrimination thresholds reported in the literature. At the outset, we emphasize that the size and shape of threshold contours depend on ancillary experimental factors, including task design, stimulus spatial and temporal properties, and participants' state of adaptation. Given these differences, we do not expect quantitative agreement across studies. Nonetheless, such comparisons help set our findings in the context of the literature. To illustrate, we present several such comparisons below, with the first three shown in the colorimetric representations used in the original studies.

We first compared the overall pattern of threshold variation in the isoluminant plane with measurements made by MacAdam using the method of adjustment (3) (see **Supplement S6** for details). In his seminal work, the ellipses do not represent discrimination thresholds *per se*, but rather the standard deviation of color matches for each reference stimulus. Nevertheless, we consider his measurements to be comparable to ours, based on the linking assumption that discrimination thresholds are proportional to the internal noise that governs the variability of the appearance-based matches (49). We observed a similar global structure in how the orientation and scale of the ellipses vary with reference stimulus. As expected, the absolute sizes of the threshold contours differ between studies (Fig. 3A; MacAdam ellipses magnified by 10× while ours magnified by 2×). In addition to differences in task and stimulus spatial and temporal structure, it is worth noting that in MacAdam's experiment, participants controlled the stimulus duration themselves (50) and their state of adaptation differed considerably across reference stimuli (37). Despite these differences, the general correspondence between the datasets is apparent. It is also noteworthy that MacAdam's results are based on 25,000 adjustments at a limited number of reference locations, whereas our ∼6,000 forced-choice trials enabled us to characterize discrimination performance across all in-gamut reference–comparison pairs.

In a more recent study, Danilova and Mollon (39) measured threshold contours across a relatively broad region of the isoluminant plane, with sparse sampling of reference stimuli. The experimental paradigm in their study closely resembled ours: both used a fixed adapting point—D65 in their case and the monitor gray point in ours—and employed an oddity task to estimate discrimination thresholds. They used a 4AFC design combined with an adaptive staircase procedure, whereas we used a 3AFC version. To compare our data with theirs, we transformed our discrimination threshold contours read out from the WPPM fit into the same scaled MacLeod–Boynton space (51) used in their study (**Supplement S7**). Despite minor methodological differences, our results replicated the overall pattern of variation in ellipse orientation and size across the color space. In particular, thresholds were smallest near the adapting point, increased with distance from it, and the ellipses generally pointed towards the adapting point. Given the similarity in experimental design, it is not surprising that we observe closer agreement in the absolute sizes of the threshold contours than in comparison to MacAdam's data (Fig. 3B; their ellipses were magnified by 4×, ours by 1.5×).

In the next comparison, we turned to the study by Krauskopf and Gegenfurtner (37), whose measurements were concentrated within a small region near the achromatic point. Their experiment used a fixed adapting point and a 4AFC oddity task, with individual thresholds estimated using a three-down-one-up staircase procedure. To enable direct comparison, we read out threshold contours from our model at the same set of reference stimuli they used. Their results revealed two key features: (1) the threshold contour was smallest at the adapting point, and (2) as the reference moved away from it, the contours became increasingly elongated along the axis pointing toward the adapting point. Both features were observed in our data albeit with some inter-participant variability (Fig. 3C; **Supplement S8**). However, because their measurements were limited to reference stimuli near the adapting point, their data did not reveal how thresholds vary with increasing distance from it. In contrast, our results demonstrate that these patterns extend across a much broader region of the isoluminant plane.

Lastly, we compared our results with iso-distance contours obtained with different versions of the CIELab ΔE color difference metrics (52). For these comparisons, we plotted the CIELab contours in our model space. Although ΔE metrics were developed to describe supra-threshold perceptual color differences and are based on data integrated from multiple sources, comparisons with threshold-level measurements remain of interest—particularly given the widespread use of ΔE to equate perceptual differences in studies of cognitive processes (53, 54). To derive a threshold contour for any given reference stimulus, we identified the comparison stimuli corresponding to $\Delta E = 2.5$ across multiple chromatic directions and fit an ellipse. While the choice of $\Delta E = 2.5$ is arbitrary, it primarily affects the overall size of the contour rather than its shape. The comparison reveals that the iso-distance contours of the original CIELab ΔE 76, which remains widely used, bear little resemblance to our threshold contours (Fig. 3D, left panel; **Supplement S9**). The large deviations we observed between ΔE 76 and our data provide further caution against the practice of using ΔE 76 to predict perceptual color difference. In contrast, the more recent ΔE 94 and ΔE 2000 metrics provided a much closer match (Fig. 3D, center and right), with only modest deviations from our measurements. These deviations may arise from differences between threshold and supra-threshold perceptual judgments, as well as from discrepancies in experimental conditions between our study and those used to constrain the parameters of the CIELab metrics. An important feature of our data is that it enables such comparison with any perceptual metric across the isoluminant plane.

2. Discussion

2.1. A data-efficient approach for characterizing color discrimination thresholds. In this study, we demonstrated a data-

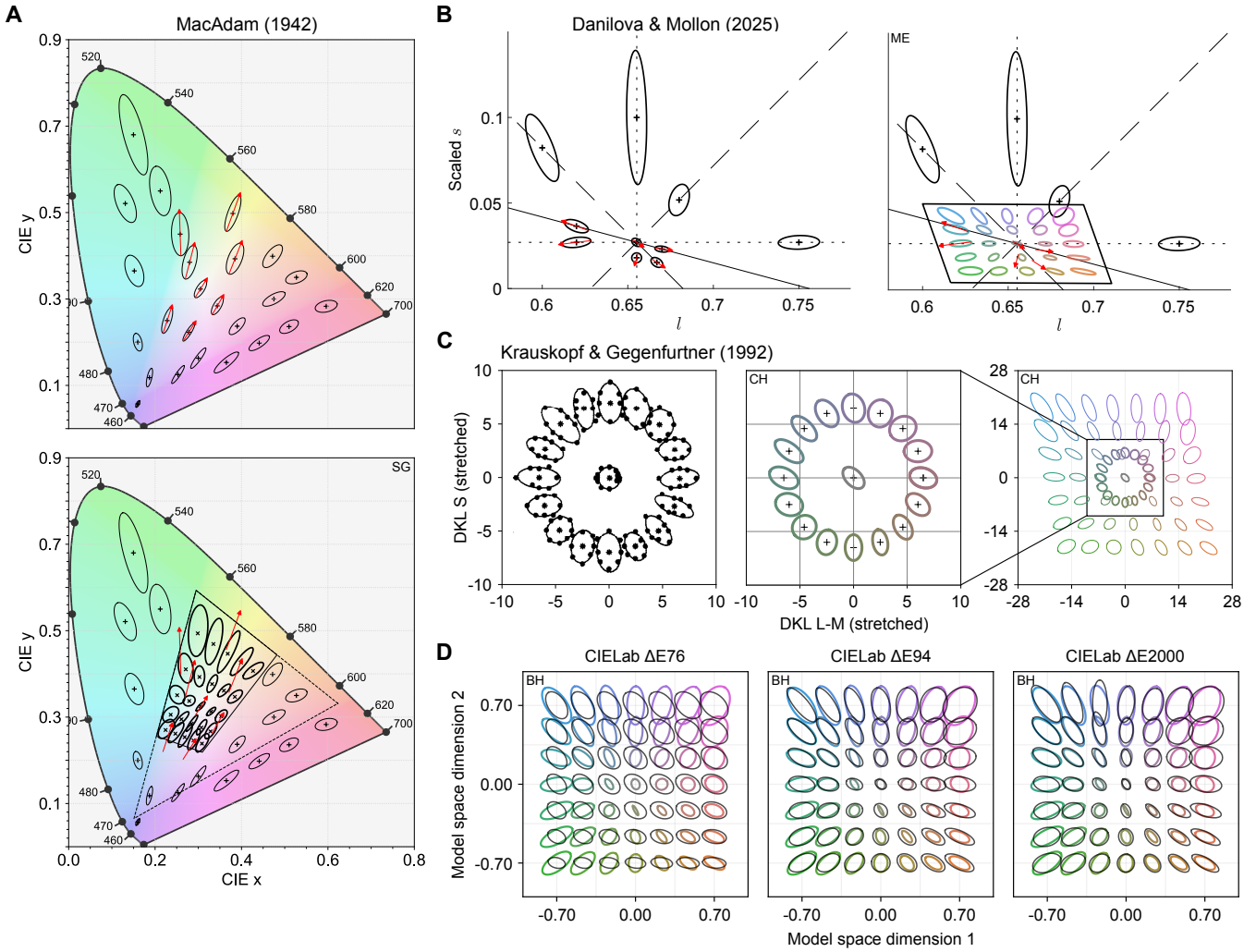

Fig. 3. Comparison of color discrimination thresholds with previous measurements. (A) MacAdam ellipses (3). Top: MacAdam's original threshold contours, magnified by $10\times$ for visualization. Bottom: Threshold contours from one participant in our study, transformed from the model space into CIE 1931 chromaticity space. reference stimuli were sampled from a 5×5 grid evenly spaced from –0.75 to 0.75 along each dimension of the model space. To reduce visual clutter, MacAdam ellipses that fall within the gamut of our isoluminant plane (parallelogram) are shown as red arrows indicating only their major axes. For visual comparability, our ellipses are magnified $2\times$ to approximate the size of those in MacAdam's data. Triangle: gamut of our monitor. (B) Danilova and Mollon (39). Left: Original threshold contours (79.4% correct) from their study, magnified by $4\times$. Right: Threshold contours from one participant in our study (colored ellipses), transformed from the model space into a scaled MacLeod–Boynton space. Reference points were sampled on a 5×5 grid ranging from –0.7 to 0.7. As in (A), to reduce visual clutter, their ellipses that fall within the gamut of our isoluminant plane (parallelogram) are shown as red arrows indicating only their major axes. For visual comparability, our ellipses are magnified by $1.5\times$. (C) Krauskopf and Gegenfurtner (37). Left: Original threshold contours (79.4% correct converged by a three-down-one-up staircase) from their study (Fig. 14 from (37); reproduced under Creative Commons CC BY-NC-ND 4.0, https://creativecommons.org/licenses/by-nc-nd/4.0/). Right: Threshold contours from one participant in our study, transformed into a stretched DKL space. All contours are shown at their original sizes. (D) CIELab ΔE 76, ΔE 94, and ΔE 2000. ΔE values were converted into percent correct using a Weibull psychometric function, and threshold is defined as the $\Delta E = 2.5$. Colored lines represent the measured thresholds from one participant, shown at their original sizes. For visual comparability, the predicted threshold contours from each CIELab metric (black lines) were magnified by factors of 5, 2.5, and 2.5, for ΔE 76, ΔE 94, ΔE 2000 metrics, respectively, to approximately match the scale of the measured thresholds in our study. See **Supplement S6-9** for additional details.

efficient approach for achieving a comprehensive characterization of human color discrimination thresholds. Participants performed a 3AFC oddity task and completed 6,000 trials that were specifically targeted near threshold via a non-parametric adaptive trial-placement procedure (24, 25). We then developed and fit a novel WPPM to these adaptively sampled trials (along with a small number of fallback trials). The WPPM defines a continuous mapping from each reference stimulus to its associated internal noise, characterized by a covariance matrix. This mapping, in turn, enables predictions of discrimination performance for any pair of reference and comparison stimuli—effectively mapping out the full four-dimensional psychometric field. To evaluate model validity, we interleaved 6,000 additional validation trials to estimate 25 probe psychometric functions. The results revealed that thresholds read out from the WPPM closely matched those derived from the validation trials, supporting the model's accuracy. Thus, by combining the non-parametric adaptive trial-placement procedure with *post hoc* fitting of the semi-parametric WPPM, we achieved an unprecedentedly comprehensive characterization of color discrimination in the isoluminant plane.

Our measurements align qualitatively with previous studies that used either sparse sampling or targeted at a

small region of a color space (3, 37, 39). Moreover, our measurements provide a more comprehensive characterization, in that the WPPM allows direct readout of threshold contour at any reference stimulus without the need for additional measurement. Additionally, for studies examining how thresholds vary with factors such as stimulus size, presentation duration, or adaptation state, our approach offers a scalable and data-efficient approach for measuring how these factors affect the psychometric field. Finally, we have performed simulations that indicate it will be feasible to fully characterize the color psychometric discrimination field across the three-dimensional gamut of our display, a goal that was has been previously described as "hopelessly difficult" (21).

2.2. Implications for the mechanisms of color perception. Consistent with a well-established body of evidence, we found that thresholds were smallest near the achromatic reference, reflecting heightened sensitivity at the adapting point (37, 55–59). In addition, threshold contours were oriented toward the achromatic center, in agreement with previous findings (37, 39, 60). The observation that the size and orientation of the elliptical threshold contours vary with the reference stimulus rules out mechanistic models that posit a linear transformation of cone excitations into three post-receptoral channels followed by fixed additive noise. Such models predict identical ellipses across the space. Moreover, the observation that the orientation of the elliptical threshold contours changes across reference stimuli also rules out mechanistic models in which a linear transformation of cone excitations is followed by limiting noise applied independently to each of the three channels. These models allow variation in size and shape but still predict identical principal axes for all threshold contours.

Cone-opponent models that posit noise and nonlinearities at multiple stages of processing, possibly with an overcomplete cone-opponent representation, may be able to account for the observed data, as may models that invoke higher-order mechanisms (e.g. mechanisms narrowly tuned for hue). For more on relevant ideas see refs 61–67. Although mechanistic models are often tested using additional manipulations such as adaptation and noise-masking, our dataset provides a new and comprehensive benchmark for evaluating and distinguishing competing accounts. In future work, our approach could be extended by incorporating such manipulations to further differentiating those models.

2.3. Toward a metric of supra-threshold color difference. A longstanding and fundamental question in vision science is whether it is possible to develop a perceptual metric that accurately predicts both threshold-level and supra-threshold judgments of color difference. For example, considerable effort has gone into attempts to find color representations where the perceptual color difference between two color stimuli is predicted by the Euclidean distance of their coordinates in the representation (e.g., the original 1976 CIELab and CIELuv ΔE metrics; see refs 29, 68). Our measurements directly establish a locally Euclidean metric for threshold-level differences. While threshold behavior is well-described as locally Euclidean, however, supra-threshold judgments have been shown to violate the assumptions of a globally Euclidean geometry (69, 70). In particular, perceptual similarity judgments at larger distances often fail to satisfy key Euclidean properties such as the expectation that variability in judgments should increase with Euclidean distance (69), and that a stimulus equidistant from two endpoints should be perceived as equally similar to both (70).

An alternative framework, originally proposed by Fechner (9) and explored subsequently (21, 61, 71–76), suggests that supra-threshold differences may be understood as the accumulation of small threshold-level differences along a path between stimuli. In this framework, color space is taken to be a Riemannian manifold—a space that is locally Euclidean but may be globally curved. The perceptual distance between two colors is hypothesized to correspond to the geodesic—the shortest path between them in terms of accumulated thresholds. This distance is computed by integrating local thresholds along all possible paths between the two points and selecting the path with the smallest total. In our observer model, this integration is effectively equivalent (up to a constant) to summing internal noise along the path.

Testing this *Riemannian hypothesis* requires knowledge of how internal noise (or threshold) varies across color space, as this determines the geodesics. Our measurements provide the necessary knowledge for the isoluminant plane, enabling direct empirical tests of the Riemannian hypothesis within this slice of color space, as well as elaborations of this hypothesis (74, 76). The results of such tests may depend on the particular experimental paradigms used to assess supra-threshold perceptual differences.

Because there is no guarantee that the geodesics between two stimuli in the isoluminant plane are themselves confined to this plane within the full three-dimensional color space, testing the Riemannian hypothesis in this plane based on our current data would be considered provisional. Nonetheless, such tests would provide valuable exploration of the perceptual geometry revealed by our measurements. As noted above, our approach makes it feasible to extend the measurements to the full three-dimensional color space, which, when completed, will allow subsequent investigations to overcome this limitation.

It is possible that the Riemannian hypothesis—and more generally the idea that threshold-level judgments can predict supra-threshold judgments—will fail. Nonetheless, we view understanding whether, when and how such failures occur as central to guiding development of a successful account of color difference perception.

2.4. Individual differences and implications. We studied a relatively young cohort of eight participants and found good consistency across individuals, with only modest individual differences. Such differences have, in general, provided useful insights into the neural mechanisms of color vision and are also of interest for understanding how much any given person may differ from an average characterization. To enable studies involving larger and more diverse populations, further improvements in the efficiency of our approach are likely feasible. In the present study, we deliberately used a non-parametric adaptive trial-placement procedure to avoid biasing data collection by assuming the accuracy of the WPPM. Now that the WPPM has been validated, future studies could incorporate adaptive trial-placement strategies tailored to this model or develop priors informed by the current dataset to further improve efficiency. Thus, our

results have the potential to support investigations into how variation in the psychometric field of color discrimination relates to underlying biological factors such as pre-retinal absorption, photopigment spectral sensitivity, the ratio of L to M cones in the mosaic, and other color vision measures such as unique hues. Individual differences in these factors have been extensively studied (77–85).

2.5. Beyond color discrimination. Our approach is generalizable to a wide range of perceptual tasks. A key insight that makes comprehensive characterization of human color discrimination thresholds feasible is the assumption—shared by both our model and the models implemented in AEPsych—that internal noise, and thus thresholds, vary smoothly across stimulus space. This smoothness assumption is not unique to color perception; it applies broadly to other domains. Indeed, smoothly varying elliptical or ellipsoidal thresholds have been reported in studies of motion perception (86, 87), auditory speed discrimination (88–90), motion-in-depth (91), and numerosity perception (92–94). These parallels highlight the broader relevance of our framework and suggest that combining the non-parametric adaptive trial-placement prodcedure with the WPPM could be a powerful strategy for characterizing perceptual limits across diverse domains.

3. Materials and Methods

3.1. Preregistration. This study was preregistered at a public repository. As described in the preregistration document, exploratory analyses were conducted on data from one participant (CH) prior to finalizing the analysis plans for other participants.

3.2. Participants. Eight participants (six female, aged 22–30 years; seven right-handed), were recruited for the study. Six were paid volunteers who were naive to the purpose of the study and were compensated for their participation. The remaining two were experimenters and participated without additional compensation. All participants had normal or corrected-to-normal vision (20/40 or better in each eye, assessed using a Snellen eye chart) and normal color vision (assessed using Ishihara plates). The study was approved by the Institutional Review Board at University of Pennsylvania, and written informed consent was obtained from all participants prior to the experiment.

3.3. Apparatus. Stimuli were presented using an Alienware computer (Aurora R11) running Windows 10 Enterprise, equipped with Intel Core^{TM} i7-10700K processor and NVIDIA GeForce RTX 3080 GPU. The display was a DELL U2723QE monitor (59.8cm width, 33.6 cm height, 3840 × 2160 resolution, 60 Hz refresh rate). The monitor was positioned 189 cm from the chinrest, subtending a visual angle of 18.0 × 10.2 degrees of visual angle (dva). Monitor color and luminance measurements were obtained with a Klein K-10A colorimeter and a SpectraScan PR-670 radiometer. The display resolution was approximately 200 pixels/dva, above the typical human foveal resolution limit.

The Alienware computer was used solely for stimulus presentation, whereas adaptive sampling of the stimuli was performed on a separate custom-built PC with a high-performance Gigabyte motherboard (X299X aorus master), an NVIDIA GeForce RTX3070 GPU and a 12-core Intel i9-10920X processor. This computer also ran Windows Enterprise 10. The two computers communicated via a shared network disk, using a custom protocol based on text files that both computers could read and write.

A USB speaker (3 Watts output power, 20k Hz frequency response) was used for playing auditory feedback, and a gamepad controller (Logitech Gamepad F310) was used for registering trial-by-trial responses.

3.4. Stimulus. The visual scene (Fig. **S25**A) was constructed in Unity (v2022.3.24f1) and rendered using its standard shader. The scene consisted of three identical blobby 3D objects, each created in Blender (v4.0) with a matte, non-reflective surface. On each trial, the surface color of the blobby objects was varied by adjusting their RGB values in Unity. The three blobby objects (2.5 × 2.5 dva; 203,900 pixels each) were arranged in a triangular configuration (Fig. 1A). Each blobby object was centered and floating inside its own cubic room (3.3 × 3.3 dva; $x = 0.302, y = 0.322, Y = 66.1\ cd/m^2$). Each room, along with the blobby stimulus inside it, was illuminated exclusively by an achromatic spotlight positioned in front of the object and set to maximum intensity ($R = G = B = 1$). The three rooms were presented against a spatially uniform gray background (18.0 × 10.2 dva; $x = 0.306$, $y = 0.326$, $Y = 116.8\ cd/m^2$). The centers of the blobby objects were 3.7 dva apart.

3.5. Calibration and color depth. We used a SpectraScan PR-670 to measure the monitor's primaries and gamma function as rendered through Unity (**Supplement S10.1**). These measurements directly characterized the relationship between the specified RGB values for the blobby stimuli and the light emitted from the display. The same calibration was repeated for all three blobby stimuli, confirming consistent color behavior across screen locations. Based on these results, a single gamma correction—derived from the bottom-right stimulus—was applied to all three objects during the experiment. This correction was validated by remeasuring the output with gamma correction applied, showing good alignment with the predicted identity line. To confirm stability over time, we repeated the calibration one month into data collection and observed negligible changes.

Additionally, we used a Klein K-10A colorimeter to verify that the system achieved sufficient color depth. For this check, a single blobby stimulus was presented at the center of the screen, rather than in the full triangular arrangement. Measurements confirmed that Unity and our video chain were able to produce at least 10-bit color precision via its native 8-bit output and implicit spatial dithering that occurred across the surface of the blobby object through the rendering process (**Supplement S10.2**).

3.6. Design. We restricted our stimuli to lie within the isoluminant plane that passes through the monitor's gray point (i.e., the point where each RGB primary was set to half of its maximum value). To define the boundaries of this plane, we identified the limits of RGB values that remained within the monitor's color gamut. These boundary points formed a parallelogram in RGB space. We then computed an affine transformation that maps this parallelogram onto a square bounded between −1 and 1 (**Supplement S1**). The forward and inverse transformations enabled conversion between RGB and model space: stimuli were rendered in RGB space, while trial placement and model fitting were performed in the model space.

We used AEPsych (v0.7) to sample a total of 6,000 reference–comparison stimulus pairs. The first 900 trials were generated using Sobol' sampling (44), a "space-filling" design based on a low-discrepancy quasi-random sequence. The remaining 5,100 trials were adaptively selected to efficiently estimate thresholds across the entire psychometric field. Each stimulus pair was defined in the 2D model space. As a result, the psychometric field comprised four variables: two specifying the reference stimulus, $x_0 \in \mathbb{R}^2$, and the other two specifying a difference vector, $\Delta \in \mathbb{R}^2$, which was added to the reference to define the comparison stimulus $x_1 = x_0 + \Delta$. Reference values were constrained between –0.75 and 0.75 along each model dimension. Each element of Δ was constrained between –0.25 and 0.25 to ensure that all comparison stimuli remained within the $[-1, 1]^2$ bounds of the model space. During the initial 900 Sobol'-sampled trials, the difference vector Δ was scaled by one of three factors (1/4, 2/4, or 3/4) before being added to the reference stimulus. This controlled the distance between the reference and comparison stimuli, effectively modulating task difficulty. These scaling factors were evenly distributed and pseudo-randomized across trials. For the remaining 5,100 trials, all four variables were adaptively selected using AEPsych's optimization procedure. Specifically, the underlying GP model was updated every 20 trials, and new trials were selected using the Expected Absolute Volume Change (EAVC) acquisition function (25), targeting the 66.7% threshold level across the entire psychometric field.

In addition to the 6,000 AEPsych-driven trials, we interleaved an additional 6,000 validation trials sampled using MOCS. Each participant was tested on 25 reference stimuli: one was fixed at the achromatic point and the remaining 24 were Sobol'-sampled within the isoluminant plane bounded between –0.6 and 0.6 along each model dimension. For each reference, a chromatic direction was Sobol'-sampled between 0° and 360°. Each validation condition consisted of 12 stimulus levels: 11 equally spaced along the sampled direction and one easily discriminable level, with each level repeated 20 times. These levels were determined based on a pilot dataset described in the preregistration documents.

The validation trials were pre-generated for each participant, pseudo-randomized so that every 300 validation trials contained all the unique trials (25 conditions x 12 levels). To minimize differential learning effects between AEPsych-driven and validation trials, we pre-generated a randomized sequence in which the two trial types were arranged in alternating pairs, with the order within each pair shuffled. However, because AEPsych occasionally required longer time to determine the next trial placement, this sequence could not always be followed in real time. For this reason, we implemented a fallback trial strategy (**Supplement S3**): if for any trial AEPsych did not have trial-placement computed in time, the next validation trial was inserted to keep the experiment moving. If necessary, subsequent validation trials were queued, but this was capped to a lead of four validation trials ahead of AEPsych trials. Once the cap was reached and AEPsych was still not ready, pre-generated fallback trials were presented instead. These fallback trials were Sobol'-sampled with the difference vector Δ scaled by one of three factors (2/8, 3/8, or 4/8) to manipulate task difficulty. Validation trials resumed once AEPsych caught up. Notably, the fallback trials (range: 0-466) were included alongside the 6,000 AEPsych trials when fitting the WPPM.

3.7. Procedure. Participants performed a 3AFC oddity task. Each trial began with a fixation cross presented at the center of the screen for 0.5 s, followed by a blank screen for 0.2 s. Then, three blobby stimuli appeared inside the cubic rooms for 1 s. After participants responded, a blank screen was shown for 0.2 s, followed by auditory and visual feedback indicating accuracy ("correct" with a beep or "incorrect" with a buzz). Each trial was separated by a 1.5 s inter-trial interval. Participants were instructed that they could move their eyes freely during the stimulus presentation, but should maintain fixation while the fixation cross was on the screen.

The majority of the participants (7 out of 8) completed a total of 12 sessions. Each session began with 40 practice trials to familiarize participants with the task. This was followed by 1,000 experimental trials, consisting of 500 trials determined by AEPsych, 500 predetermined validation trials and a small amount of fallback trials. The validation trials were randomized and the two trial types were fully intermixed. Participants took a break every 200 trials. Each session took approximately 1.5 hours to complete. In total, those seven participants completed between 12,256 and 12,466 trials, depending on the number of fallback trials inserted. Participant CH completed 12,000 trials across 10 sessions, without fallback trials implemented. As a result, the inter-trial interval was slightly longer for this participant, but we expected this to have a negligible effect on performance.

3.8. The Wishart Process Psychometric Model. Our implementation of the WPPM relies on two core assumptions about color perception: (1) internal noise that limits color discrimination follows a multivariate Gaussian distribution centered at the reference stimulus, with a covariance matrix that captures both the magnitude and directional structure (i.e., size and orientation) of the noise, and (2) the covariance matrix varies smoothly across the model space, without abrupt local discontinuities. In the following subsections, we describe the WPPM in five parts. First, we define the observer model, which predicts discrimination performance—quantified as the percent-correct responses—for a given pair of reference and comparison stimuli by modeling both the internal representation of the stimuli and the decision rule. Second, we describe how we use a finite-basis Wishart process to parameterize the entire field of covariance matrices across the model space, and what governs its smoothness. Third, we describe the weak prior imposed on the covariance matrix field to favor smooth variation. Fourth, we explain how, given any specification of the covariance matrix field, we compute the likelihood and thereby the posterior probability of the model and its free parameters given binary (correct or incorrect) color-discrimination responses. Finally, we show how, once the model is fit, the covariance matrix for any reference-comparison stimulus pair can be read out and combined with the observer model to predict percent-correct performance, including threshold contours around any reference stimulus.

3.8.1. Observer model. On each trial, the observer is presented with two identical reference stimuli, denoted x_0, and one comparison stimulus, denoted $x_1 = x_0 + \Delta$ where Δ represents a small offset from the reference. Our model assumes that these three stimuli are independently encoded into an internal representational space by a noisy process, which we assume to follow a multivariate Gaussian distribution. Formally,

$$z_0 \sim \mathcal{N}(x_0, \Sigma(x_0)) \quad [1]$$

$$z_0' \sim \mathcal{N}(x_0, \Sigma(x_0)) \quad [2]$$

$$z_1 \sim \mathcal{N}(x_0 + \Delta, \Sigma(x_0 + \Delta)) \quad [3]$$

where z_0, z_0', z_1 denote the internal representations derived from the two reference and the comparison stimuli, respectively. Our model posits that the observer correctly identifies z_1 as representing the comparison stimulus (i.e. the "odd-one-out") if

$$d_M^2(z_0, z_0') - \min\left(d_M^2(z_0, z_1),\ d_M^2(z_0', z_1)\right) < 0, \quad [4]$$

where $d_M^2(\cdot,\cdot)$ denotes the squared Mahalanobis distance for a selected pair of internal representations, formulated as

$$d_M^2(z_0, z_0') = (z_0 - z_0')^\top \mathbf{S}^{-1}(z_0 - z_0') \quad [5]$$

$$d_M^2(z_0, z_1) = (z_0 - z_1)^\top \mathbf{S}^{-1}(z_0 - z_1) \quad [6]$$

$$d_M^2(z_0', z_1) = (z_0' - z_1)^\top \mathbf{S}^{-1}(z_0' - z_1), \quad [7]$$

where $\mathbf{S}$ is the weighted average of the covariance across the reference and the comparison stimuli, that is,

$$\mathbf{S} = \frac{2}{3} \cdot \Sigma(x_0) + \frac{1}{3} \cdot \Sigma(x_0 + \Delta). \quad [8]$$

This decision rule is consistent with an observer that uses distances between internal representations to judge stimulus similarity (95). We approximated the percent-correct performance using (N=2,000) Monte Carlo simulations (Fig. 1E) as the closed-form analytical solution is complicated to derive (96). In each Monte Carlo simulation, we draw samples according to Eq. (1) - Eq. (3) and the outcome is marked as correct if the condition in Eq. (4) is fulfilled. The proportion of correct outcomes in the Monte Carlo simulation defines the model's predicted percent-correct performance, which is then used to evaluate the likelihood function as explained in Section 3.8.4.

3.8.2. Covariance matrix field. The WPPM specifies a covariance matrix at any selected reference stimulus across the entire isoluminant plane. Each matrix specifies the perceptual noise in terms of the variance along the two model dimensions (σ^2_{dim1}, σ^2_{dim2}) and their covariance ($\sigma_{\text{dim1,dim2}}$) (Fig. 1C-D).

The covariance matrix field is constructed using one-dimensional Chebyshev polynomial basis functions (97). We chose Chebyshev polynomials because they allow for the expression of smoothness over a bounded interval without imposing periodic boundary conditions. Let $x = [x_{\text{dim1}}, x_{\text{dim2}}]$ denote a location in the 2D model space. The basis functions are defined recursively for each model space dimension as given here for x_{dim1}:

$$T_0(x_{\text{dim1}}) = 1, \quad [9]$$

$$T_1(x_{\text{dim1}}) = x_{\text{dim1}}, \quad [10]$$

$$T_{i+1}(x_{\text{dim1}}) = 2x_{\text{dim1}} \cdot T_i(x_{\text{dim1}}) - T_{i-1}(x_{\text{dim1}}), \quad [11]$$

where $x_{\text{dim1}},\ T_i(x_{\text{dim1}}) \in \mathbb{R}^n$, and n is the number of discretized points along that stimulus dimension, which can be chosen flexibly to achieve any desired resolution. We construct two-dimensional basis functions by taking the outer product:

$$\phi_{i,j}(x) = T_i(x_{\text{dim1}}) \cdot T_j(x_{\text{dim2}}), \quad [12]$$

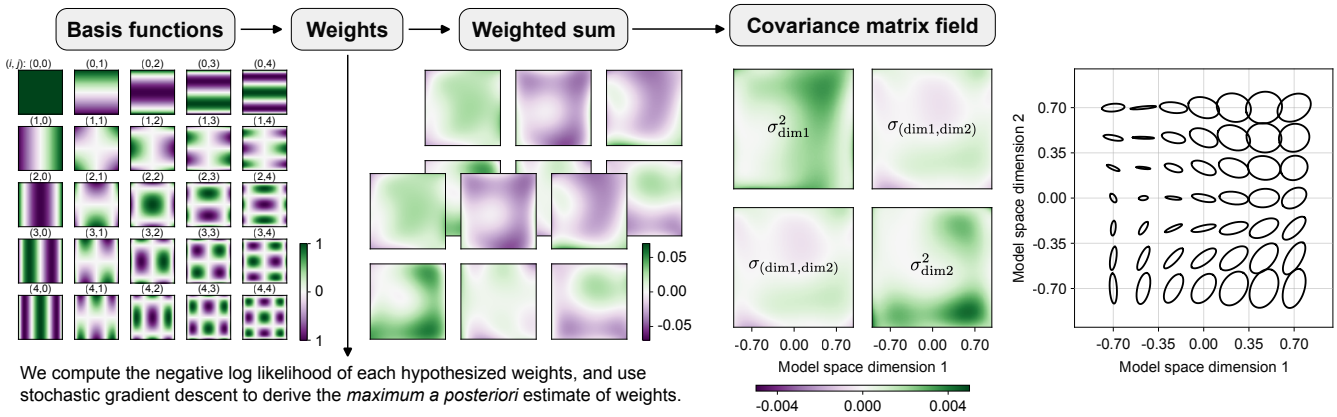

Fig. 4. The finite-basis Wishart Process Psychophysical Model (WPPM). In our implementation, we used a set of 5×5 two-dimensional Chebyshev polynomial basis functions, denoted $\phi_{i,j}(x)$, where $i, j \in \{0, 1, \ldots, 4\}$. These basis functions were linearly combined using a learnable weight matrix $\mathbf{W}$ to produce an overcomplete representation $\mathbf{U}_{k,l}(x)$, where $k \in 1, 2$ and $l \in 1, 2, 3$. The resulting representation $\mathbf{U}_{k,l}$ was then combined with its own transpose to produce a field of symmetric, positive semi-definite covariance matrices. Each matrix specifies the internal noise in terms of the variance along the two model dimensions (σ^2_{dim1}, σ^2_{dim2}) and their covariance ($\sigma_{\text{dim1,dim2}}$). For this example covariance matrix field, the weights used to generate the field were samples from the Wishart process prior with $\epsilon = 0.5$.

where $\phi_{i,j} \in \mathbb{R}^{n \times m}$, with $n = m$ representing the number of discretized points along each dimension of the model space. We limited the number of basis functions to five per dimension, i.e., $i, j \in \{0, 1, \ldots, 4\}$, resulting in a total of $5 \times 5 = 25$ 2D basis functions (Fig. 4, first panel). The polynomial order of each 2D basis function is given by $i + j$, with higher-order basis functions describing more rapidly varying patterns.

The basis functions were weighted by a learnable parameter matrix, $\mathbf{W} \in \mathbb{R}^{5 \times 5 \times 2 \times 3}$, where the first two dimensions index the Chebyshev basis functions along each model space dimension ($i, j \in \{0, 1, \ldots, 4\}$), and the last two dimensions index the output components ($k \in \{1, 2\}$ and $l = \{1, 2, 3\}$). The weighted basis functions are expanded into an overcomplete representation $\mathbf{U}_{k,l} \in \mathbb{R}^{n \times m}$ (Fig. 4, second panel) as the following,

$$\mathbf{U}_{k,l}(x) = \sum_{i=0}^{4} \sum_{j=0}^{4} W_{i,j,k,l} \cdot \phi_{i,j}(x). \quad [13]$$

This weighted sum overcomplete representation was then combined with its own transpose to yield a positive semi-definite covariance matrix, $\Sigma(x) \in \mathbb{R}^{2 \times 2}$ for x at any discretized point in the model space, that is,

$$\Sigma(x) = \begin{bmatrix} \sigma^2_{\text{dim1}} & \sigma_{\text{dim1,dim2}} \\ \sigma_{\text{dim1,dim2}} & \sigma^2_{\text{dim2}} \end{bmatrix} = \mathbf{U}_{k,l}(x) \cdot \mathbf{U}_{k,l}(x)^T. \quad [14]$$

This process ensures that the resulting covariance matrices are symmetric and positive semi-definite (Fig. 4, third panel). The weight matrix serves as the free parameters of the model, controlling the smoothness of the covariance matrix field. The model is highly flexible, capable of generating a wide range of covariance matrix fields, from smooth to rapidly varying fields (Fig. 1C-D).

3.8.3. Prior over the weight matrix. We imposed a weak prior over the weight matrix $\mathbf{W}$. Specifically, we assumed that each weight was distributed *a priori* as a zero-mean one-dimensional Gaussian,

$$W_{i,j,k,l} \sim \mathcal{N}(0, \eta(i + j)), \quad [15]$$

where η represents the variance of each weight and it decays exponentially with $i + j$, which denotes the polynomial order of the corresponding 2D basis function, that is,

$$\eta(i + j) = \gamma \cdot \epsilon^{(i+j)}. \quad [16]$$

The scalar γ controls the overall amplitude of the variance and was fixed at 3×10^{-4}.

The scalar ϵ controls the rate at which the prior variance decays with increasing polynomial order, and was fixed at 0.5. A higher value of ϵ results in a prior that favors more sharply varying covariance matrix fields, while a lower value favors smoother fields. By setting $\epsilon = 0.5$, we adopted a prior that favors relatively smooth variation across the space (Fig. 4 fourth panel). This value was chosen by hand based on examination of a pilot dataset and analysis of the data reported here for participant CH, available at a public repository. As preregistered, this decision was made before analyzing the validation trials from the remaining seven participants.

3.8.4. Model fitting. We computed the negative log-likelihood of any hypothesized weight matrix given the participant's binary responses y_r as follows:

$$p_r(y_1, \ldots, y_r | \mathbf{W}) = \sum_{r=1}^{R} \Big(y_r \cdot \log(p_r) + (1 - y_r) \cdot \log(1 - p_r) \Big), \quad [17]$$

where $y_r \in \{0, 1\}$ indicates whether the response on trial r was correct (1) or incorrect (0), R is the total number of trials used to fit the WPPM. The model-predicted accuracy p_r for each trial is given by:

$$p_r = \mathbf{Pr}\left[d^2_M(z_0, z'_0) < \min\left(d^2_M(z_0, z_1),\ d^2_M(z'_0, z_1) \right) | \mathbf{W} \right]. \quad [18]$$

Note that on the rth trial, z_0, z'_0, and z_1 are internal representations that depend on the reference and comparison stimuli (x_0 and x_1) for that trial. For notational simplicity, the subscript r is omitted here.

Since we imposed a prior on the covariance matrix field to reflect the expectation of smooth variation, we combined the likelihood (Eq. 17) and the prior (Eq. 16) to calculate the posterior probability of $\mathbf{W}$. As there is no simple closed form expression for p, we resorted to a numerical approximation based on Monte Carlo simulations. The numerical approximation we built was differentiable with respect to the covariance matrix field, which enabled us to use gradient descent to maximize the posterior probability of $\mathbf{W}$ (see details in **Supplement S11**).

3.8.5. Psychometric field. For any given reference stimulus, the WPPM allows readouts of percent-correct performance along any chromatic direction, which in turn allows us to construct a threshold contour. We sampled comparison stimuli along 16 chromatic directions and simulated internal representations to estimate percent-correct performance, yielding a psychometric function for each direction (Fig. 1F). The threshold distance in each direction was defined as the comparison stimulus corresponding to 66.7% correct. Collectively, these threshold distances form a contour that closely resembles an ellipse, with only minor deviations due to inhomogeneous internal noise between the reference and comparison stimuli. However, because the stimuli are locally proximal in the model space, such discrepancies are

negligible. We therefore fit an ellipse to these points as a practical approximation. As a way of visualizing the psychometric field, we plot these ellipses—each corresponding to 66.7% threshold level—at a grid of reference locations. We emphasize, however, that the WPPM provides the full four-dimensional psychometric field, enabling readouts of the psychometric function along any chromatic direction for any reference stimulus within the model space.

3.9. Data analysis. Color calibration analyses were performed using MATLAB 2023b. We computed inverse gamma lookup tables from the measured gamma functions and derived transformation matrices to convert values from the model space to RGB space (**Supplement S10**). Stimulus presentation, including gamma correction, was implemented in Unity using C#.

The experiment and model fitting were conducted in Python 3.11 and JAX (98). Behavioral data were separated into AEPsych-driven plus fallback trials on the one hand and validation trials on the other. The WPPM was fit exclusively to the AEPsych and fallback trials. To assess variability in model estimates, we performed 10 bootstrap resamplings of the AEPsych-driven trials, preserving the original ratio between Sobol', adaptively sampled, and fallback trials in each resampled dataset. The WPPM was then re-fit to each of the 10 bootstrapped datasets.

For the held-out validation trials, we computed the Euclidean distance between each reference and its paired comparison stimulus. For each of the 25 conditions, a Weibull psychometric function was fit to the binary color discrimination responses, with the guess rate fixed at 33.3% correct. Thresholds were defined as the comparison stimulus corresponding to 66.7% correct. To estimate variability, we bootstrapped each condition 120 times and computed 95% confidence intervals for the threshold estimates.

To assess the validity of the WPPM estimates, we performed linear regression (constrained to pass through the origin) between the thresholds estimated from the WPPM fits and those from the validation trials. This comparison was repeated across all 120 bootstrapped validation datasets. Notably, since only 10 Wishart bootstrapped fits were available, we replicated each one 12 times and shuffled to align with the number of validation bootstraps, rather than generating 120 separate WPPM fits, which would have been computationally expensive. This allowed us to approximate the confidence intervals for the regression slope and correlation coefficient.

Data and code availability. Data and code will be made publicly available upon acceptance for publication. All experiments, data collection, processing activities, and open sourcing were conducted at the University of Pennsylvania.

Acknowledments. We thank Nicolas P. Cottaris for his assistance with the calibration, and other colleagues at UPenn Vision labs for valuable feedback.

1. S Hecht, S Shlaer, MH Pirenne, Energy, quanta, and vision. *J. Gen. Physiol.* **25**, 819–840 (1942).
2. FW Campbell, JG Robson, Application of fourier analysis to the visibility of gratings. *The J. physiology* **197**, 551 (1968).
3. DL MacAdam, Visual sensitivities to color differences in daylight. *J. Opt. Soc. Am.* **32**, 247–274 (1942).
4. H de Lange Dzn, Research into the dynamic nature of the human fovea→ cortex systems with intermittent and modulated light. i. attenuation characteristics with white and colored light. *J. Opt. Soc. Am.* **48**, 777–784 (1958).
5. PA Aspinall, PR Kinnear, LJ Duncan, BF Clarke, Prediction of diabetic retinopathy from clinical variables and color vision data. *Diabetes Care* **6**, 144–8 (1983).
6. CA Johnson, M Wall, HS Thompson, A history of perimetry and visual field testing. *Optom. Vis. Sci.* **88**, E8–E15 (2011).
7. Y Niwa, S Muraki, F Naito, T Minamikawa, M Ohji, Evaluation of acquired color vision deficiency in glaucoma using the rabin cone contrast test. *Invest Ophthalmol Vis Sci* **55**, 6686–90 (2014).
8. R Vemala, S Sivaprasad, JL Barbur, Detection of early loss of color vision in age-related macular degeneration - with emphasis on drusen and reticular pseudodrusen. *Invest Ophthalmol Vis Sci* **58**, BIO247–BIO254 (2017).
9. GT Fechner, *Elemente der psychophysik.* (Breitkopf u. Härtel) Vol. 2, (1860).
10. JM Hillis, DH Brainard, Distinct mechanisms mediate visual detection and identification. *Curr. Biol.* **17**, 1714–1719 (2007).
11. J Zhou, LR Duong, EP Simoncelli, A unified framework for perceived magnitude and discriminability of sensory stimuli. *Proc. Natl. Acad. Sci.* **121**, e2312293121 (2024).
12. J Palmer, CT Ames, DT Lindsey, Measuring the effect of attention on simple visual search. *J. Exp. Psychol. Hum. Percept. Perform.* **19**, 108 (1993).
13. J Najemnik, WS Geisler, Optimal eye movement strategies in visual search. *Nature* **434**, 387–391 (2005).
14. M Olkkonen, PF McCarthy, SR Allred, The central tendency bias in color perception: Effects of internal and external noise. *J. vision* **14**, 5–5 (2014).
15. K Knoblauch, LT Maloney, *Modeling psychophysical data in R.* (Springer Science & Business Media) Vol. 32, (2012).
16. N Prins, , et al., *Psychophysics: a practical introduction.* (Academic Press), (2016).
17. DM Green, JA Swets, , et al., *Signal detection theory and psychophysics.* (Wiley New York) Vol. 1, (1966).
18. FG Ashby, FA Soto, Multidimensional signal detection theory. *Oxf. handbook computational mathematical psychology* pp. 13–34 (2015).
19. MJ Hautus, NA Macmillan, CD Creelman, *Detection theory: A user's guide.* (Routledge), (2021).
20. JM Foley, GE Legge, Contrast detection and near-threshold discrimination in human vision. *Vis. research* **21**, 1041–1053 (1981).
21. Ev Schrödinger, Outline of a theory of color measurement for daylight vision. *Phys. Annu.* **63**, 397–520 (1920).
22. LA Lesmes, ZL Lu, J Baek, TD Albright, Bayesian adaptive estimation of the contrast sensitivity function: the quick csf method. *J. vision* **10**, 17–17 (2010).
23. AB Watson, Quest+: A general multidimensional bayesian adaptive psychometric method. *J. Vis.* **17**, 10–10 (2017).
24. L Owen, et al., Adaptive nonparametric psychophysics. *arXiv preprint arXiv:2104.09549* (2021).
25. B Letham, P Guan, C Tymms, E Bakshy, M Shvartsman, Look-ahead acquisition functions for bernoulli level set estimation in *International Conference on Artificial Intelligence and Statistics.* (PMLR), pp. 8493–8513 (2022).
26. L Hedjar, M Toscani, KR Gegenfurtner, Importance of hue: color discrimination of three-dimensional objects and two-dimensional discs. *J. Opt. Soc. Am. A* **42**, B296–B304 (2025).
27. DH Brainard, DG Pelli, T Robson, Display characterization. *Signal Process.* **80**, 2–067 (2002).
28. D Brainard, Cone contrast and opponent modulation color spaces. *Hum. color vision* (1996).
29. DH Brainard, Color appearance and color difference specification in *The Science of Color*, ed. SK Shevell. (Elsevier), 2nd edition, pp. 191–216 (2003).
30. DH Brainard, A Stockman, Colorimetry. (McGraw Hill), (2010).
31. K Mullen, DM Ennis, A simple multivariate probabilistic model for preferential and triadic choices. *Psychometrika* **56**, 69–75 (1991).
32. AG Wilson, Z Ghahramani, Generalised wishart processes in *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, UAI'11. (AUAI Press, Arlington, Virginia, USA), p. 736–744 (2011).
33. WRJ Brown, DL MacAdam, Visual sensitivities to combined chromaticity and luminance differences. *J. Opt. Soc. Am.* **39**, 808–834 (1949).
34. C Noorlander, MJ Heuts, JJ Koenderink, Sensitivity to spatiotemporal combined luminance and chromaticity contrast. *J. Opt. Soc. Am.* **71**, 453–459 (1981).
35. C Noorlander, JJ Koenderink, RJ Den Olden, BW Edens, Sensitivity to spatiotemporal colour contrast in the peripheral visual field. *Vis. Res.* **23**, 1–11 (1983).
36. AB Poirson, BA Wandell, The ellipsoidal representation of spectral sensitivity. *Vis. research* **30**, 647–652 (1990).
37. J Krauskopf, G Karl, Color discrimination and adaptation. *Vis. research* **32**, 2165–2175 (1992).
38. K Knoblauch, LT Maloney, Testing the indeterminacy of linear color mechanisms from color discrimination data. *Vis. research* **36**, 295–306 (1996).
39. M Danilova, J Mollon, Effect of stimulus size on chromatic discrimination. *J. Opt. Soc. Am. A* **42**, B167–B177 (2025).
40. JR Newton, RT Eskew, Chromatic detection and discrimination in the periphery: a postreceptoral loss of color sensitivity. *Vis. neuroscience* **20**, 511–521 (2003).
41. TG Shepard, EA Swanson, CL McCarthy, RT Eskew, A model of selective masking in chromatic detection. *J. vision* **16**, 3–3 (2016).
42. TG Shepard, SI Lahlaf, RT Eskew, Labeling the lines: A test of a six-mechanism model of chromatic detection. *J. Vis.* **17**, 9–9 (2017).
43. CK Williams, CE Rasmussen, *Gaussian processes for machine learning.* (MIT press Cambridge, MA) Vol. 2, (2006).
44. IM Sobol, The distribution of points in a cube and the approximate evaluation of integrals. *USSR Comput. mathematics mathematical physics* **7**, 86–112 (1967).
45. AR Girshick, MS Landy, EP Simoncelli, Cardinal rules: visual orientation perception reflects knowledge of environmental statistics. *Nat. neuroscience* **14**, 926–932 (2011).
46. G Aguilar, FA Wichmann, M Maertens, Comparing sensitivity estimates from mlds and forced-choice methods in a slant-from-texture experiment. *J. Vis.* **17**, 37–37 (2017).
47. F Hong, S Badde, MS Landy, Causal inference regulates audiovisual spatial recalibration via its influence on audiovisual perception. *PLOS Comput. Biol.* **17**, 1–37 (2021).
48. S Horiuchi, T Nagai, Color discrimination repetition distorts color representations. *Sci. Reports* **14**, 9615 (2024).

49. WJ Crozier, AH Holway, On the law for minimal discrimination of intensities: I. *Proc. Natl. Acad. Sci.* **23**, 23–28 (1937).
50. BA Wandell, Color measurement and discrimination. *J. Opt. Soc. Am. A* **2**, 62–71 (1985).
51. DI MacLeod, RM Boynton, Chromaticity diagram showing cone excitation by stimuli of equal luminance. *J. Opt. Soc. Am.* **69**, 1183–1186 (1979).
52. C Colorimetry, Commission internationale de l'èclairage: Vienna (2004).
53. J Winawer, N Witthoft, Effects of color terms on color perception and cognition in *Encyclopedia of color science and technology*. (Springer), pp. 777–785 (2023).
54. DJ Garside, AL Chang, HM Selwyn, BR Conway, The origin of color categories. *Proc. Natl. Acad. Sci.* **122**, e2400273121 (2025).
55. K Craik, The effect of adaptation on differential brightness discrimination. *The J. Physiol.* **92**, 406 (1938).
56. W Brown, The effect of field size and chromatic surroundings on color discrimination. *J. Opt. Soc. Am.* **42**, 837–844 (1952).
57. LM Hurvich, D Hurvich-Jameson, *Opponent chromatic induction and wavelength discrimination*. (Springer), (1961).
58. MR Pointer, Color discrimination as a function of observer adaptation. *J. Opt. Soc. Am.* **64**, 750–759 (1974).
59. JM Loomis, T Berger, Effects of chromatic adaptation on color discrimination and color appearance. *Vis. Res.* **19**, 891–901 (1979).
60. KR Gegenfurtner, The verriest lecture: Color vision from pixels to objects. *J. Opt. Soc. Am. A* **42**, B313–B328 (2025).
61. G Wyszecki, Concepts and methods, quantitative data and formulae. *Color. Sci.* pp. 130–175 (1982).
62. BA Wandell, *Foundations of vision*. (Sinauer Associates), (1995).
63. CC Chen, JM Foley, DH Brainard, Detection of chromoluminance patterns on chromoluminance pedestals ii: model. *Vis. Res.* **40**, 789–803 (2000).
64. RT Eskew Jr, Higher order color mechanisms: A critical review. *Vis. research* **49**, 2686–2704 (2009).
65. A Stockman, DH Brainard, , et al., Color vision mechanisms. *OSA handbook optics* **3**, 11–1 (2010).
66. T Hansen, KR Gegenfurtner, Higher order color mechanisms: Evidence from noise-masking experiments in cone contrast space. *J. vision* **13**, 26–26 (2013).
67. SK Shevell, PR Martin, Color opponency: tutorial. *J. Opt. Soc. Am. A* **34**, 1099–1108 (2017).
68. AR Robertson, et al., Cie recommendations on uniform color spaces, color-difference equations, and metric color terms. *Color. Res. Appl* **2**, 3 (1977).
69. SM Wuerger, LT Maloney, J Krauskopf, Proximity judgments in color space: tests of a euclidean color geometry. *Vis. research* **35**, 827–835 (1995).
70. RJ Ennis, Q Zaidi, Geometrical structure of perceptual color space: Mental representations and adaptation invariance. *J. vision* **19**, 1–1 (2019).
71. DL Macadam, Judd's contributions to color metrics and evaluation of color differences. *Color. Res. & Appl.* **4**, 177–193 (1979).
72. Q Zaidi, Is there a perceptual color space? (2001).
73. JJ Koenderink, *Color for the Sciences*. (MIT press), (2010).
74. R Bujack, E Teti, J Miller, E Caffrey, TL Turton, The non-riemannian nature of perceptual color space. *Proc. Natl. Acad. Sci.* **119**, e2119753119 (2022).
75. V Roberti, Helmholtz, schrödinger, and the first non-euclidean model of perceptual color space. *Annalen der Physik* **536**, 2300536 (2024).
76. E Stark, TL Turton, R Bujack, Diminishing returns in perceptual color space-now in color. (2025).
77. J Neitz, GH Jacobs, Polymorphism of the long-wavelength cone in normal human colour vision. *Nature* **323**, 623–625 (1986).
78. MA Webster, DI MacLeod, Factors underlying individual differences in the color matches of normal observers. *J. Opt. Soc. Am. A* **5**, 1722–1735 (1988).
79. DH Brainard, et al., Functional consequences of the relative numbers of l and m cones. *J. Opt. Soc. Am. A* **17**, 607–614 (2000).
80. J Kremers, et al., L/m cone ratios in human trichromats assessed by psychophysics, electroretinography, and retinal densitometry. *J. Opt. Soc. Am. A* **17**, 517–526 (2000).
81. J Carroll, J Neitz, M Neitz, Estimates of l: M cone ratio from erg flicker photometry and genetics. *J. vision* **2**, 1–1 (2002).
82. H Hofer, J Carroll, J Neitz, M Neitz, DR Williams, Organization of the human trichromatic cone mosaic. *J. Neurosci.* **25**, 9669–9679 (2005).
83. JM Bosten, Do you see what i see? diversity in human color perception. *Annu. review vision science* **8**, 101–133 (2022).
84. KJ Emery, VJ Volbrecht, DH Peterzell, MA Webster, Fundamentally different representations of color and motion revealed by individual differences in perceptual scaling. *Proc. Natl. Acad. Sci.* **120**, e2202262120 (2023).
85. D Rezeanu, M Neitz, J Neitz, From cones to color vision: a neurobiological model that explains the unique hues. *J. Opt. Soc. Am. A* **40**, A1–A8 (2023).
86. TE Reisbeck, KR Gegenfurtner, Velocity tuned mechanisms in human motion processing. *Vis. research* **39**, 3267–3286 (1999).
87. RA Champion, TC Freeman, Discrimination contours for the perception of head-centered velocity. *J. Vis.* **10**, 14–14 (2010).
88. TC Freeman, et al., Discrimination contours for moving sounds reveal duration and distance cues dominate auditory speed perception. *PloS one* **9**, e102864 (2014).
89. S Carlile, J Leung, The perception of auditory motion. *Trends hearing* **20**, 2331216516644254 (2016).
90. G Bertonati, MB Amadeo, C Campus, M Gori, Auditory speed processing in sighted and blind individuals. *Plos one* **16**, e0257676 (2021).
91. SG Wardle, D Alais, Evidence for speed sensitivity to motion in depth from binocular cues. *J. Vis.* **13**, 17–17 (2013).
92. GM Cicchini, G Anobile, DC Burr, Spontaneous perception of numerosity in humans. *Nat. communications* **7**, 12536 (2016).
93. GM Cicchini, G Anobile, DC Burr, Spontaneous representation of numerosity in typical and dyscalculic development. *Cortex* **114**, 151–163 (2019).
94. GM Cicchini, G Anobile, DC Burr, P Marchesini, R Arrighi, The role of non-numerical information in the perception of temporal numerosity. *Front. Psychol.* **14**, 1197064 (2023).
95. PM Churchland, Some reductive strategies in cognitive neurobiology. *Mind* **95**, 279–309 (1986).
96. DM Ennis, K Mullen, A general probabilistic model for triad discrimination, preferential choice, and two-alternative identification in *Multidimensional models of perception and cognition*. (Psychology Press), pp. 115–122 (2014).
97. PL Chebyshev, *Théorie des mécanismes connus sous le nom de parallélogrammes*. (Imprimerie de l'Académie impériale des sciences), (1853).
98. J Bradbury, et al., JAX: composable transformations of Python+NumPy programs (2018).