eLife

Reviewed Preprint
v1 • November 25, 2025
Not revised

Reviewed Preprint
v2 • May 20, 2026
Revised by authors

For correspondence:
fh862@sas.upenn.edu

† These authors contributed equally to this work

Competing interests: No competing interests declared

Funding: See page 22

Reviewing editor: Krystel R Huxlin, University of Rochester, United States

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,†], David H Brainard[1,†]

[1]Department of Psychology, University of Pennsylvania, Philadelphia, United States • [2]Reality Lab Research, Meta, Menlo Park, United States • [3]FAIR, Meta, Menlo Park, United States • [4]Center for Neural Science, New York University, New York, United States • [5]Center for Computational Neuroscience, Flatiron Institute, New York, United States

eLife Assessment

This **important** study describes a novel Bayesian psychophysical approach that efficiently measures how well humans can discriminate between colors across the entire isoluminant plane. The evidence was considered **compelling**, as it included successful model validation against hold-out data and published datasets. This approach could prove to be of use to color vision scientists, as well as to those who employ computational psychophysics and attempt to model perceptual stimulus fields with smooth variations over coordinate spaces.

https://doi.org/10.7554/eLife.108943.2.sa3

Abstract

Color discrimination thresholds—the smallest detectable color differences—provide a benchmark for models of color vision, enable quantitative evaluation of eye diseases, and inform the design of display technologies. Despite their importance, a comprehensive characterization of these thresholds has long been considered intractable due to the psychophysical curse of dimensionality. Here, we address this challenge using a novel semi-parametric Wishart Process Psychophysical Model (WPPM), which leverages the feature that the internal noise limiting color discrimination varies smoothly across stimulus space. The model was fit to data collected with a non-parametric adaptive trial-placement procedure, enabling efficient stimulus selection. Together, through the combination of adaptive trial placement and *post hoc* WPPM fitting, we achieved a comprehensive characterization of color discrimination in the isoluminant plane with only ~6,000 trials per participant (N = 8). Once fit, the WPPM allows readouts of discrimination performance for any stimulus pair. We validated these readouts against 25 probe psychometric functions, measured with an additional 6,000 trials per participant held out from model fitting. In conclusion, our study provides a foundational dataset for color vision, and our approach generalizes beyond color to any domain in which the internal noise limiting performance varies smoothly across stimulus space, offering a powerful and efficient method for comprehensively characterizing various perceptual discrimination thresholds.

Introduction

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 (Hecht et al., 1942; Campbell and Robson, 1968), guide the design of displays and specification of perceptual tolerances (MacAdam, 1942; de Lange Dzn, 1958), allow quantitative evaluation of eye diseases (Aspinall et al., 1983; Johnson et al., 2011; Niwa et al., 2014; Vemala et al., 2017), inform models of supra-

threshold perceptual representations (Fechner, 1860; Hillis and Brainard, 2007; Zhou et al., 2024), and allow perceptual effects to be incorporated into the study of cognitive processes (Palmer et al., 1993; Najemnik and Geisler, 2005; Olkkonen et al., 2014). Modern psychophysical methods (Knoblauch and Maloney, 2012; Prins et al., 2016) provide rigorous quantification of thresholds, and the theory of signal detection (Green et al., 1966; Ashby and Soto, 2015; Hautus et al., 2021) provides a mature framework for relating thresholds to the precision of the underlying representation.

Despite the central role of perceptual thresholds, characterization of thresholds has largely been limited to single stimulus dimensions. For example, pedestal functions characterize contrast discrimination thresholds across varying baseline contrasts (Foley and Legge, 1981). 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 significance and applications described above, fully characterizing human color discrimination—even on a single planar slice—has long been considered impractical (Schrödinger, 1920). 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 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 can greatly improve sampling efficiency (Lesmes et al., 2010; Watson, 2017), they typically rely on certain parametric forms. In many cases—including ours— such forms are 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 (Owen et al., 2021; Letham et al., 2022), and then fitting the data *post hoc* with a semiparametric model that leverages the feature that the internal noise limiting color discrimination varies smoothly across stimulus space. We collected full datasets 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 can be used to test computational and neural models of color discrimination, benchmark color metrics, and develop models that can predict supra-threshold color discrimination performance.

Results

Overview

The Results section is organized as follows. We begin with a brief overview of the experimental stimuli and task (Task and stimuli), followed by a summary of how our model characterizes the full psychometric field (The Wishart Process Psychophysical Model (WPPM)) and a description of the non-parametric adaptive trial-placement procedure used to collect the data (Adaptively sampled trials). Having described these essential methods, we then present our core results (WPPM threshold estimates) and evaluate the validity of our model (Validation of the WPPM).

Finally, we compare our findings with previous measurements from the color discrimination literature (Comparison with previous measurements). Additional technical details are provided in Methods and Materials and Appendix 1 - Appendix 12.

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 (Figure 1A). The comparison stimulus was pseudo-randomly assigned to one of the three positions. 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 (Hedjar et al., 2025) provide a comparison of color discrimination using stimuli similar to ours versus traditional flat spatially uniform patches.

We made spectral calibration measurements (Brainard et al., 2002) 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 (Brainard, 1996, 2003; Brainard and Stockman, 2010). 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 (Figure 1B; Appendix 1). We refer to 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.

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 (Figure 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 the 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 stimuli. Because discrimination thresholds depend on the relative size of signal change and internal noise, an alternative formulation could instead attribute threshold variation across stimulus location to nonlinearities in signal encoding while assuming constant internal noise, or to a mixture of nonlinearities and stimulus-varying noise (Zhou et al., 2024). Our formulation should be understood as characterizing the signal-to-noise properties that limit discrimination and not a commitment to a particular interpretation of how these properties arise.

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 (Figure 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

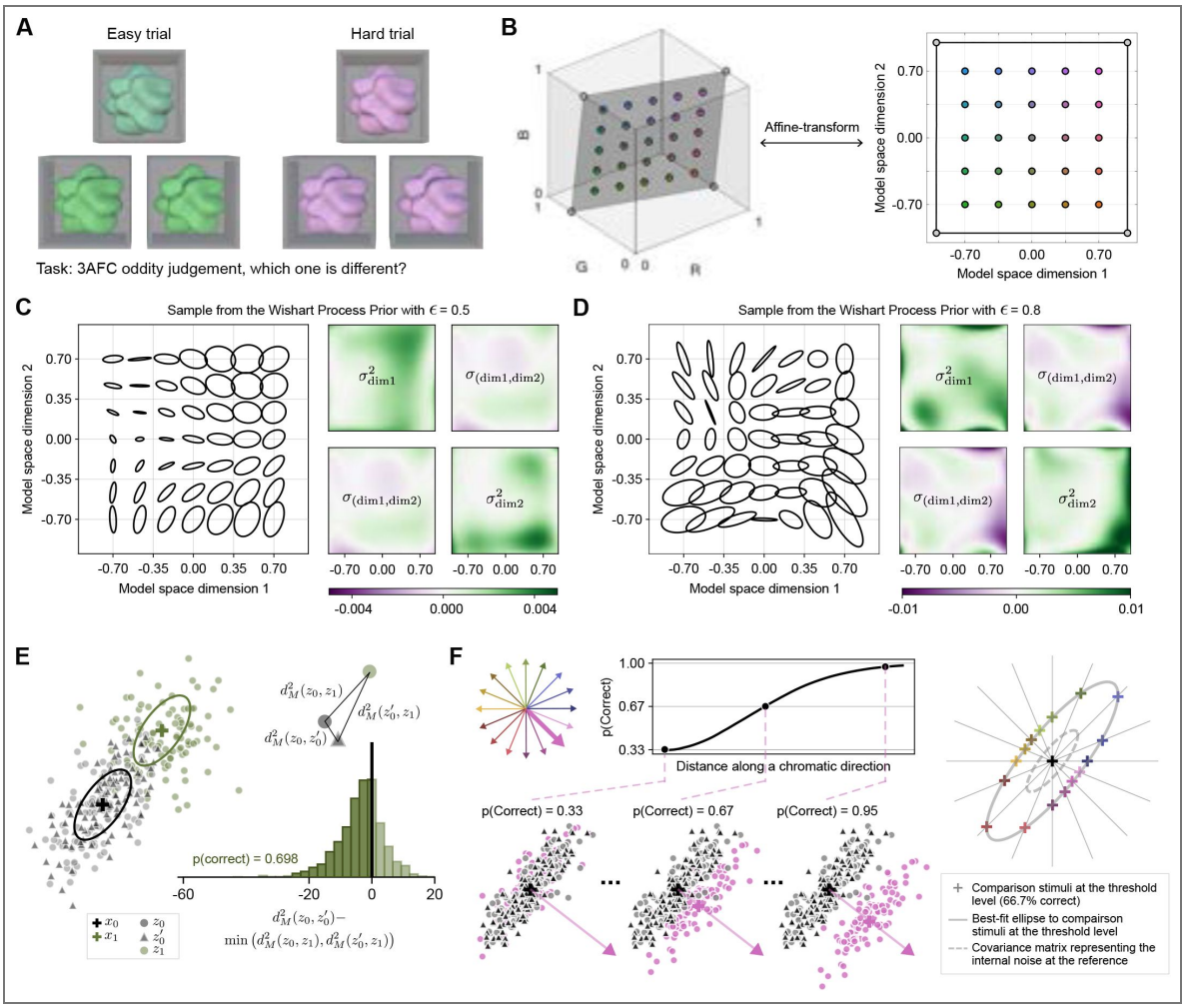

Figure 1. Task, stimuli and the WPPM.

(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 are constrained to lie in the isoluminant plane that passes through the display's gray point. Data are 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 is generated by sampling from a finite-basis Wishart random process with a smooth prior (ϵ = 0.5 and γ = 0.0003; Prior over the weight matrix). Although the field is illustrated on a 7 × 7 grid, it specifies a covariance matrix $\left[\sigma^2_{\text{dim}\,1}, \sigma_{(\text{dim}\,1,\text{dim}\,2)}; \sigma_{(\text{dim}\,1,\text{dim}\,2)}, \sigma^2_{\text{dim}\,2}\right]$ for every stimulus in the plane, as shown in the heat maps. (D) Example of a less smoothly varying covariance matrix field. This field is obtained by drawing from a finite-basis Wishart random process with a less smooth prior (ϵ = 0.8 and γ = 0.0003). (E) Observer model. For each stimulus triplet $\left[x_0, x_0^{'}, x_1\right]$, internal representations $\left[z_0, z_0^{'}, z_1\right]$ 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.

one out"). Because there is no simple closed-form solution for this decision rule (Mullen and Ennis, 1991), we used Monte Carlo simulation to approximate the percent-correct performance (Observer model).

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 (Wilson and Ghahramani, 2011). Intuitively, the Wishart process prior introduces a regularization term to the model—it penalizes rapid variation in the covariance matrix field. The strength of smoothness is controlled by two hyperparameters of the model, ϵ and γ (Figure 1C-D; Prior over the weight matrix).

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, that is, the sum of the log prior density and log likelihood function (Model fitting).

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 (Figure 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 (Figure 1F, right panel). Given our assumption that internal noise follows a multivariate Gaussian distribution, these threshold distances form approximately elliptical contours, which we fit with ellipses for visualization. This approach is consistent with prior work showing that ellipses provide a good approximation of color discrimination thresholds (MacAdam, 1942; Brown and MacAdam, 1949; Noorlander et al., 1981, 1983; Poirson and Wandell, 1990; Krauskopf and Gegenfurtner, 1992; Knoblauch and Maloney, 1996; Danilova and Mollon, 2025), despite some reported deviations (Newton and Eskew, 2003; Shepard et al., 2016, 2017). 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.

Adaptively sampled trials

Reference and comparison stimuli for each trial were selected using AEPsych (Owen et al., 2021), an open-source package for adaptive psychophysics. For the adaptive sampling model, we used a probit-Bernoulli Gaussian Process (GP) model (Williams and Rasmussen, 2006) 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 (Sobol, 1967) to provide an adequate initialization for the GP; 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 (Letham et al., 2022) (Figure 2A, S1, S2).

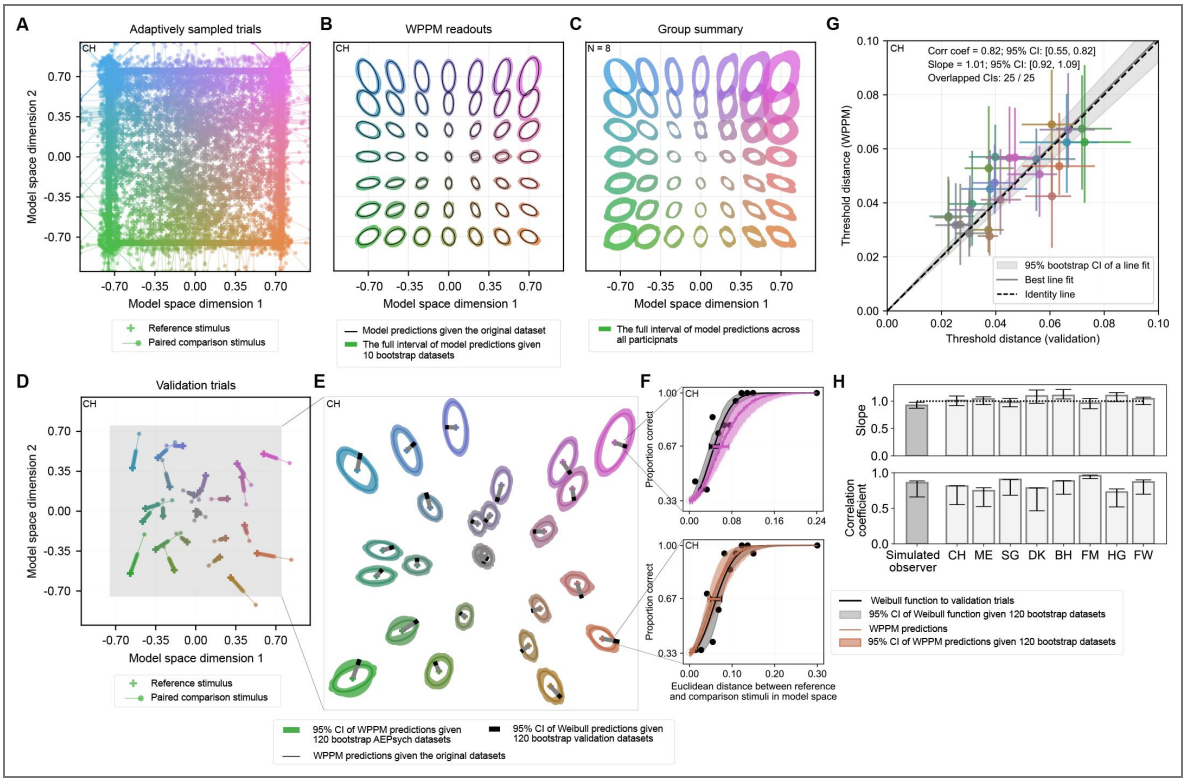

Figure 2. Threshold results and validation.

(A) Adaptively sampled trials. AEPsych-driven stimulus pairs are adaptively sampled for estimating thresholds across the psychometric field. Of the 6,000 trials, the first 900 are Sobol'-sampled; the remaining 5,100 (shown here) are adaptively selected using the EAVC acquisition function, based on a non-parametric GP model that is updated every 20 trials. (B) Discrimination threshold contours (66.7% correct) read out from the WPPM on a grid of reference stimuli for a representative participant, based on fits to the 6,000 AEPsych trials. (C) Group summary of WPPM readouts (N = 8), evaluated on the same grid of reference stimuli. (D) Validation trials for the same participant. The validation conditions (reference stimuli and chromatic directions along which the comparison stimulus varies) are randomly generated for each participant (see Appendix 4 for validation conditions used for the remaining participants). (E) Comparison of thresholds. Ellipses represent discrimination threshold contours read out from the WPPM fit (same fit as in panel B), evaluated at the 25 reference stimuli used in the validation trials. Gray lines: the validation directions; black bars: the 95% bootstrapped confidence intervals for the corresponding validation thresholds. (F) Comparison of psychometric functions. Only two validation conditions are shown for illustration (see Appendix 4.1 for all 25 conditions for each participant). (G) Linear regression of thresholds predicted by the WPPM against validation thresholds for the same participant. Horizontal and vertical error bars represent 95% confidence intervals for the validation thresholds and the WPPM predictions, respectively. (H) Summary of regression slopes and correlation coefficients for all participants. Error bars: 95% confidence intervals. As a benchmark, the same analysis is performed on a dataset simulated using a ground truth WPPM instance that approximates CIELab Δ*E*94 (Appendix 5).

Adaptive sampling with AEPsych requires solving two optimization problems: one for updating the GP model (Williams and Rasmussen, 2006) and another for selecting the next trial using the Expected Absolute Volume Change (EAVC) acquisition function (Letham et al., 2022). 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 (Appendix 3). The number of fallback trials varied across participants, ranging from 0 to 466 Figure S1). These trials were included along with the 6000 AEPsych-driven trials when fitting the WPPM.

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 (Figure 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 tend to be radially oriented with respect to the achromatic point. These regularities are consistent with previous results in the color discrimination literature, as explained further in Comparison with previous measurements.

The data were broadly consistent across participants, in the sense that the three regularities noted above are observed in the individual participant data (Appendix 2.1). In the model space representation, individual variability was lowest near the achromatic point where sensitivity was highest and increased with the distance between the reference and the achromatic point (Figure 2C). Specifically, this variation was quite large in the upper right quadrant of the model space, where ellipse orientations varied considerably. This variation in orientation was also apparent when examining the data in other colorimetric representations (Appendix 7–Appendix 8). Increases in interparticipant variability with increasing thresholds have been observed in other perceptual discrimination tasks (Girshick et al., 2011; Aguilar et al., 2017; Hong et al., 2021).

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: 11 were evenly spaced, and one was selected to provide easily discriminable catch trials (Figure 2D). The comparison levels were selected based on a pilot dataset to account for variability in thresholds across different reference stimuli and chromatic directions (see Design 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 (Horiuchi and Nagai, 2024).

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 Figure 2F). We then used the WPPM fit (constrained by non-overlapping trials) to extract elliptical contours corresponding to the 66.7% threshold level for each validation reference (Figure 2E). To compare the WPPM and validation thresholds, we read out the WPPM threshold along the MOCS chromatic direction for each reference. The 95% bootstrapped confidence intervals for the WPPM estimates overlapped with those from the Weibull fits in all 25 conditions for participant CH (Figure 2E) and in 22 to 25 conditions across other participants (Appendix 4.1). These results demonstrate a high degree of agreement between thresholds derived from the WPPM psychometric field and the 25 discrete MOCS validation thresholds. This agreement indicates that the Wishart process prior we imposed did not lead to substantial over-smoothing, as the validation thresholds were estimated

independently without a smoothness constraint. Also notable is that the size of the 95% bootstrapped confidence intervals for the WPPM and validation thresholds was similar (Figure 2F, G; Appendix 4.1).

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 (Figure 2G). The results further support agreement between the two sets of estimates (mean correlation coefficient = 0.84, range = 0.73–0.96). For 7 out of 8 participants, the regression slope was not significantly different from 1 (mean slope = 1.04, range = 0.96–1.10) (Figure 2H; see Appendix 4.1 for comparisons for each participant). To assess whether there were more subtle sources of bias not captured by the regression slope, we analyzed the residuals—the differences 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 relationship: the model slightly overestimated thresholds when validation thresholds were low and underestimated them when validation thresholds were high (slope = −0.176, $t(198) = -5.727$, $p < 0.001$, $R^2 = 0.142$). However, the magnitude of this bias was small (Appendix 4.2).

As an additional benchmark, we simulated trials and responses from a ground-truth WPPM instance chosen to approximate the CIELab $\Delta E94$ metric, and fitted the model to the simulated data (Appendix 5). This allowed us to assess the ability of the WPPM to recover simulated ground truth, which is not possible with human data. The readout threshold ellipses based on the WPPM fit are in good agreement with the ground truth (Figure S16C). We then conducted the same validation analyses on the simulated data as described above. The thresholds read out from the WPPM fit agreed with 23 of the 25 validation thresholds, based on overlapping confidence intervals. A linear regression yielded a correlation coefficient of 0.86 and a slope of 0.92—well within the confidence intervals observed in participants' data (Figure 2H, left bar). Residual analysis revealed a negative correlation between residuals and the magnitude of the ground-truth validation thresholds (Appendix 5.6; Figure S14), consistent with trends observed in human participants. With the simulated data, however, we can interpret the magnitude of this bias in the context of the agreement with ground truth and conclude that it is small (Figure S16C). Access to ground truth also provides us with additional ways to visualize patterns in the bias (Appendix 5.7).

Taken together, these results validate the accuracy of the WPPM and highlight the remarkable efficiency of our approach. With 6,000 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 a 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.

Comparison with previous measurements

The WPPM is equivariant under affine transformations of color space (Appendix 1.4), 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, 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 (MacAdam, 1942) (Appendix 6). 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 (Crozier and Holway, 1937). 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 (Figure 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 (Wandell, 1985) and their state of adaptation differed considerably across reference stimuli (Krauskopf and Gegenfurtner, 1992). 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 responses enabled us to characterize discrimination performance across all in-gamut reference–comparison pairs in the isoluminant plane.

In a more recent study, Danilova and Mollon 2025 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 (MacLeod and Boynton, 1979) used in their study (Appendix 7). Despite methodological and stimulus 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 toward the adapting point. We observe closer agreement in the absolute sizes of the threshold contours than when comparing with MacAdam's data (Figure 3B; their ellipses were magnified by 4×, ours by 1.5×). An interesting commonality between our data and those of Danilova and Mollon 2025 is the rotation of the ellipses at the adapting point relative to the axes of the MacLeod-Boynton space, a rotation seen in all of our participants (Figure S21). See their discussion of this rotation for possible mechanistic interpretations.

In the next comparison, we turned to the study by Krauskopf and Gegenfurtner 1992, 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 generally became increasingly elongated along the axis pointing toward the adapting point. Both features were observed in our data albeit with some interparticipant variability (Figure 3C; Appendix 8). Our measurements differ from theirs, however, in the orientation of the ellipse at the adapting chromaticity: in our data, the ellipses are rotated with respect to the DKL axes for all our participants.

Lastly, we compared our results with iso-distance contours obtained with different versions of the CIELab ΔE color difference metrics (CIE, 2004). Although ΔE metrics were developed to describe supra-threshold perceptual color differences for a stimulus configuration that differs from ours, comparisons with threshold-level measurements are of interest—particularly because of the widespread use of ΔE to equate perceptual differences in studies of cognitive processes (Winawer and Witthoft, 2023; Garside et al., 2025). 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 $\Delta E76$, which remains widely used, bear little resemblance to our threshold contours (Figure 3D, left panel; Appendix 9). The large deviations we observed between $\Delta E76$ and our data provide further caution against the practice of using $\Delta E76$ to predict perceptual color difference. In contrast, the more recent $\Delta E94$ and $\Delta E00$ metrics provided a much

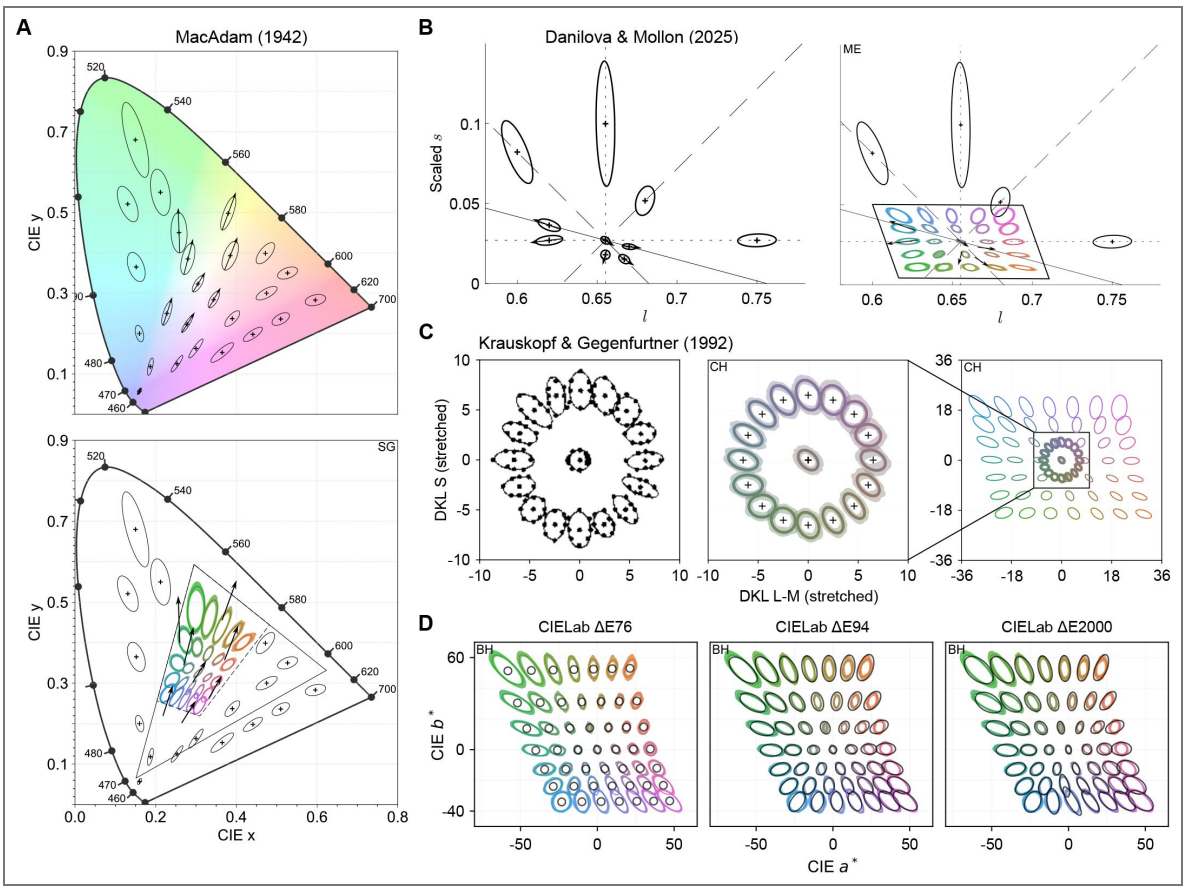

Figure 3. Comparison of color discrimination thresholds with previous measurements.

Across all panels, black contours represent thresholds from prior studies, whereas colored contours represent the 66.7% discrimination thresholds estimated in our study. Colored shaded regions indicate 95% confidence intervals computed from 120 bootstrapped datasets. (A) MacAdam 1942.Top panel: MacAdam's original threshold ellipses, magnified 10× for visualization. Bottom panel: Threshold contours measured from one participant in our study and transformed from the model space into the CIE 1931 chromaticity diagram. Reference stimuli are sampled from a 5 × 5 grid spanning [−0.7, 0.7] along each dimension of the model space. To reduce visual clutter, MacAdam ellipses falling within the gamut of our isoluminant plane (parallelogram) are shown only by arrows indicating their major axes. For visual comparability, our ellipses are magnified 2× to approximately match the scale of MacAdam's data. The triangle indicates our monitor's gamut. (B) Danilova and Mollon 2025. Left panel: Original threshold contours (79.4% correct) from their study, magnified by 4×. Right panel: Threshold contours from one participant in our study, transformed from the model space into the scaled MacLeod–Boynton space used in their study. Reference points are sampled on a 5 × 5 grid ranging from −0.7 to 0.7. As in panel A, to reduce visual clutter, their ellipses that fall within the gamut of our isoluminant plane (parallelogram) are shown as black arrows indicating only their major axes. For visual comparability, our ellipses are magnified 1.5×. (C) Krauskopf and Gegenfurtner 1992. Left panel: 79.4% threshold contours (Fig. 14 from their study, reproduced under Creative Commons CC BY-NC-ND 4.0). Right panel: 66.7% threshold contours from one participant in our study, transformed into DKL space with the axes scaled for each participant to equate thresholds along the L-M and S axes at the adapting chromaticity, as was done in their study. All contours are shown at their original sizes in this scaled representation. (D) CIELab Δ*E*76, Δ*E*94, and Δ*E*00. Threshold is defined as Δ*E* = 2.5, chosen to approximately match the scale of our measured thresholds, which are shown at their original sizes. See Appendix 6 - Appendix 9 for additional details.

closer match (Figure 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 ΔE metrics. An important feature of our data is that it enables such comparison with any perceptual metric across the isoluminant plane.

Discussion

A data-efficient approach for characterizing color discrimination thresholds

In this study, we demonstrated a data-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 (Owen et al., 2021; Letham et al., 2022). 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 (MacAdam, 1942; Krauskopf and Gegenfurtner, 1992; Danilova and Mollon, 2025). Moreover, our measurements provide a more comprehensive characterization, in that the WPPM allows direct readout of a 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 and collected preliminary data that indicate it will be feasible to fully characterize the color discrimination psychometric field across the three-dimensional gamut of our display (Hong et al., 2026), a goal that has been previously described as "hopelessly difficult" (Schrödinger, 1920).

Prior specification

A key assumption of the WPPM is that internal noise varies smoothly across the stimulus space. This smoothness assumption is implemented through a prior on the variance of the weights applied to the model's basis functions (Prior over the weight matrix). The smoothness prior plays a nontrivial role in the final WPPM estimates and therefore the choice of prior required careful selection.

Our cross-validation analyses indicate that the smoothness hyperparameters that characterize the prior we used in the main analyses fall within a regime that balances over-smoothing against excessive uncertainty in the estimates (Appendix 10.1). When the smoothness embodied in the prior is too strong, the model produces overly uniform threshold estimates that fail to capture structure in the data. When the prior smoothness is too weak, the estimates become more variable. Consistent with this bias–variance tradeoff, agreement between WPPM and validation thresholds starts off low with strong smoothness, increases as the smoothness constraint is

relaxed, and declines when smoothness becomes too weak (Appendix 10.3). These two analyses narrow the range of sensible hyperparameter values and provide support for the hyperparameter values (ϵ = 0.4; γ = 0.0003) adopted in the main analyses.

As a general matter, determining appropriate prior hyperparameter values can be challenging for interpreting data using Bayesian models. This problem is at the heart of empirical Bayesian approaches to data analyses, in which prior hyperparameters are estimated from the data, and then the data are analyzed with these estimates (Efron, 2012). A full empirical Bayesian approach of this sort currently exceeds what we can compute under reasonable time constraints. In our experience, evaluating model performance across a range of the hyperparameters using cross-validation helps identify the region that balances over-smoothing against excessive uncertainty in the estimates, while the inclusion of validation trials helps identify the regime that maximizes agreement between WPPM predictions and validation thresholds.

The number of basis functions included in the model constitutes an additional modeling choice. To evaluate this choice, we examined the fitted weights as a function of the order of the Chebyshev polynomials and found that they decay to near zero at the highest polynomial order used in the model. This indicates that including additional basis functions would not materially affect the inferred psychometric field, given our hyperparameter values (Figure S4).

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 (Craik, 1938; Brown, 1952; Hurvich and Hurvich-Jameson, 1961; Pointer, 1974; Loomis and Berger, 1979; Krauskopf and Gegenfurtner, 1992). In addition, threshold contours were oriented toward the achromatic center, in agreement with previous findings (Krauskopf and Gegenfurtner, 1992; Gegenfurtner, 2025; Danilova and Mollon, 2025). Although the WPPM characterizes the data in terms of stimulus-dependent noise, the fitted psychometric field can be used to evaluate mechanistic models that posit specific transformations between stimuli and their internal representations.

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 stimulus 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 the lengths of the major and minor ellipse axes, but predict that the orientation of these axes will be the same for all reference stimuli.

Cone-opponent models that posit noise and nonlinearities at multiple stages of processing, possibly with an over-complete cone-opponent representation to capture parallel channels along the visual pathways, 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 Wyszecki 1982; Wandell 1995; Chen et al. 2000; Eskew Jr 2009; Stockman et al. 2010; Hansen and Gegenfurtner 2013; Shevell and Martin 2017. Notably, mechanistic models are often tested using additional manipulations such as adaptation and noise-masking; our approach can be extended to incorporate manipulation of such factors (Zhang et al., 2026), as well as of stimulus spatial and temporal structure and retinal location.

We studied a relatively young cohort of eight participants and found broadly consistent patterns across individuals, while also observing individual differences. Individual differences have provided valuable insights into the mechanisms of color vision (Bosten, 2022) and are also of interest for understanding how much any given individual is likely to differ from an average characterization. A successful mechanistic model should allow investigation of whether our observed individual differences can be attributed to individual variation in biological factors

known to influence color vision, for example pre-retinal absorption, photopigment spectral sensitivity, and the ratio of L to M cones in the mosaic (Neitz and Jacobs, 1986; Brainard et al., 2000; Kremers et al., 2000; Carroll et al., 2002; Hofer et al., 2005; Bosten, 2022; Rezeanu et al., 2023).

Extensions of the WPPM framework

To enable studies involving larger and more diverse populations, further improvements in the efficiency of our approach are likely achievable. In the present study, we used a non-parametric adaptive trial-placement procedure to avoid biasing data collection by assuming the correctness of the WPPM. Given the validation of the WPPM presented here, future studies could instead incorporate adaptive trial-placement strategies tailored to the model, thereby improving efficiency. Alternatively, one could leverage the current dataset to develop stronger priors that capture the regularities we observe and use these priors to guide more efficient trial placement. Stronger priors could also increase the quality of the estimates available from a fixed set of trials, although care should be taken to ensure that the prior does not overly constrain the estimates. Another approach is to develop mechanistic models with relatively few parameters, which could be estimated efficiently using parametric adaptive sampling (Watson, 2017). Finally, a complementary strategy is to increase the rate at which participants provide information about thresholds through more efficient psychophysical experimental paradigms (Agosti et al., 2026; Barnett et al., 2025; Burge and Cormack, 2024).

Another aspect of the WPPM framework that can be adapted for different applications is the mapping between stimulus space and model space, as well as the choice of basis functions. In the present work, we defined the model space to be bounded within [−1, 1] for mathematical convenience, as this is the domain on which our chosen basis functions—the 2D Chebyshev polynomials— are defined. These basis functions could in principle be replaced by alternatives such as Zernike (Zernike, 1934; Thibos et al., 2000) or sinusoidal (Fourier) basis functions (Stein and Shakarchi, 2011), which may be better suited for stimulus domains with disc-shaped geometries or without clear boundaries. The number of basis functions can also be adjusted based on prior knowledge about the expected smoothness of the psychometric field for a given stimulus domain. More generally, other approaches to leveraging the smoothness of psychophysical performance and physiological response have been developed (Gravesen, 2015; Waz et al., 2025; Rad and Paninski, 2010; Savin and Tkacik, 2016), including recent work on color discrimination (Koenderink et al., 2026).

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 Brainard 2003; Robertson et al. 1977). Our measurements directly establish a locally Euclidean metric for threshold-level differences. While threshold behavior is well described as locally Euclidean, supra-threshold judgments have been shown to violate the assumptions of a globally Euclidean geometry (Wuerger et al., 1995; Ennis and Zaidi, 2019). 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 (Wuerger et al., 1995), and that a stimulus equidistant from two endpoints should be perceived as equally similar to both (Ennis and Zaidi, 2019).

An alternative framework, originally proposed by Fechner (Fechner, 1860) and explored subsequently (Schrödinger, 1920; Macadam, 1979; Wyszecki, 1982; Zaidi, 2001; Koenderink, 2010; Bujack et al., 2022; Roberti, 2024; Stark et al., 2025), 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 *geodesic hypothesis* requires knowledge of how internal noise (or thresholds) 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 geodesic hypothesis within this slice of color space, as well as elaborations of this hypothesis (Bujack et al., 2022; Stark et al., 2025). 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 geodesic 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 (Hong et al., 2026), which, when completed, will allow subsequent investigations to overcome this limitation.

It is possible that the geodesic 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 supra-threshold color difference perception.

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 (Reisbeck and Gegenfurtner, 1999; Champion and Freeman, 2010), auditory speed discrimination (Freeman et al., 2014; Carlile and Leung, 2016; Bertonati et al., 2021), motion-in-depth (Wardle and Alais, 2013), and numerosity perception (Cicchini et al., 2016, 2019, 2023). These parallels highlight the broader relevance of our framework and suggest that combining the non-parametric adaptive trial-placement procedure with the WPPM could be a powerful strategy for characterizing perceptual limits across diverse domains.

Methods and Materials

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 preregistering an initial hyperparameter choice of $\epsilon = 0.5$ for the main analysis. After data collection completed, we performed the hyperparameter sweeps (Appendix 10). These led to our final choice of $\epsilon = 0.4$ and $\gamma = 0.0003$.

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. 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.

Apparatus

Stimuli were presented using an Alienware computer (Aurora R11) running Windows 10 Enterprise, equipped with Intel Core– 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.

Stimulus

The visual scene (Figure S33A) 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 (Figure 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.

Calibration and color depth

We used a SpectraScan PR-670 to measure the monitor's primaries and gamma function as rendered through Unity (Appendix 11.1). 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 12-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 (Appendix 11.2).

Design

We restricted our stimuli to lie within the isoluminant plane that passes through the monitor's gray point (i.e., R = G = B = 0.5). To define the boundaries of this plane, we identified the limits of RGB values that remained within the monitor's gamut. These boundary points formed a parallelogram in RGB space. We then computed an affine transformation that maps this parallelogram onto a square bounded within [−1, 1] (Appendix 1). The forward and inverse transformations enabled conversion between RGB and the 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 (Sobol, 1967), 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, 0.75] along each model dimension. Each element of Δ was constrained between [−0.25, 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 (Letham et al., 2022), 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 within [−0.6, 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 × 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 (Appendix 3): 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, pregenerated 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.

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 AEPsych-driven trials, 500 predetermined validation trials, and a small number 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.

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 percent-correct performance for a given pair of reference and comparison stimuli by modeling both the noisy internal representations 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, along with the factors that control its smoothness. Third, we describe the weak prior imposed on the covariance matrix field to favor smooth variation. Fourth, we explain how, given a 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.

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,

$$z_0 \sim \mathcal{N}\left(x_0, \Sigma\left(x_0\right)\right) \tag{1}$$

$$z_0' \sim \mathcal{N}\left(x_0, \Sigma\left(x_0\right)\right) \tag{2}$$

$$z_1 \sim \mathcal{N}\left(x_0 + \Delta, \Sigma\left(x_0 + \Delta\right)\right) \tag{3}$$

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

$$d_M^2\left(z_0, z_0^{'}\right) - \min\left(d_M^2\left(z_0, z_1\right), d_M^2\left(z_0^{'}, z_1\right)\right) < 0, \tag{4}$$

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

$$d_M^2\left(z_0, z_0^{'}\right) = \left(z_0 - z_0^{'}\right)^{\top}\mathbf{S}^{-1}\left(z_0 - z_0^{'}\right) \tag{5}$$

$$d_M^2\left(z_0, z_1\right) = \left(z_0 - z_1\right)^{\top}\mathbf{S}^{-1}\left(z_0 - z_1\right) \tag{6}$$

$$d_M^2\left(z_0^{'}, z_1\right) = \left(z_0^{'} - z_1\right)^{\top}\mathbf{S}^{-1}\left(z_0^{'} - z_1\right), \tag{7}$$

where **S** is the weighted average of the covariance across the reference and the comparison stimuli, that is,

$$\mathbf{S} = \frac{2}{3}\cdot\Sigma\left(x_0\right) + \frac{1}{3}\cdot\Sigma\left(x_0 + \Delta\right). \tag{8}$$

This decision rule is consistent with an observer that uses distances between internal representations to judge stimulus similarity (Churchland, 1986). We approximated the percent-correct performance using (N=2,000) Monte Carlo simulations (Figure 1E) as the closed-form analytical solution is complicated to derive (Ennis and Mullen, 2014). In each Monte Carlo simulation, we draw samples according to Equation 1 - Equation 3 and the outcome is marked as correct if the condition in Equation 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 Model fitting.

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 $(\sigma^2_{\mathrm{dim}\,1}, \sigma^2_{\mathrm{dim}\,2})$ and their covariance ($\sigma_{\mathrm{dim1,dim2}}$) (Figure 1C-D).

The covariance matrix field is constructed using one-dimensional Chebyshev polynomial basis functions (Chebyshev, 1853). We chose Chebyshev polynomials because they allow for the expression of smoothness over a bounded interval without imposing periodic boundary conditions. Let$x$ = [x_{dim1}, x_{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}:

$$T_0\left(x_{\mathrm{dim}\,1}\right) = 1 \tag{9}$$

$$T_1\left(x_{\mathrm{dim}\,1}\right) = x_{\mathrm{dim}\,1}, \tag{10}$$

$$T_{i+1}\left(x_{\mathrm{dim}\,1}\right) = 2x_{\mathrm{dim}\,1}\cdot T_i\left(x_{\mathrm{dim}\,1}\right) - T_{i-1}\left(x_{\mathrm{dim}\,1}\right), \tag{11}$$

where x_{dim1}, $T_i(x_{\mathrm{dim1}}) \in \mathbb{R}^n$, 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:

$$\phi_{i,j}(x) = T_i\left(x_{\mathrm{dim}\,1}\right)\cdot T_j\left(x_{\mathrm{dim}\,2}\right), \tag{12}$$

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., $i, j \in \{0, 1, ..., 4\}$, resulting in a total of $5 \times 5 = 25$ 2D basis functions (Figure 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, $\mathbf{W} \in \mathbb{R}^{5\times5\times2\times3}$, where the first two dimensions index the Chebyshev basis functions along each model space dimension ($i, j \in \{0, 1, ..., 4\}$), 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 $\mathbf{U}_{k,l} \in \mathbb{R}^{n \times m}$ (Figure 4, second panel) as the following,

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

This weighted sum overcomplete representation was then combined with its own transpose to yield a positive semi-definite covariance matrix (Figure 4, third panel), $\Sigma(x) \in \mathbb{R}^{2\times2}$ for x at any discretized point in the model space, that is,

$$\Sigma(x) = \begin{bmatrix} \sigma^2_{\dim 1} & \sigma_{\dim 1,\dim 2} \\ \sigma_{\dim 1,\dim 2} & \sigma^2_{\dim 2} \end{bmatrix} = \mathbf{U}_{k,l}(x) \cdot \mathbf{U}_{k,l}(x)^T. \tag{14}$$

Notably, in our implementation, rather than matching the dimensionality of the intermediate representation U to that of Σ, we adopt an overcomplete parameterization motivated primarily by practical considerations. When we restricted the dimensionality indexed by l to 2, the optimization occasionally became ill-conditioned, leading to singular or unstable solutions. Expanding l to 3 substantially improved numerical stability and made the fitting procedure more robust. Increasing l beyond 3, however, would introduce additional degrees of freedom. We therefore selected $l = 3$ as a compromise between model flexibility and numerical stability. Regardless of whether the representation is square or overcomplete, the resulting matrices are symmetric and positive semidefinite.

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 (Figure 1C-D).

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,

$$W_{i,j,k,l} \sim \mathcal{N}(0, \eta_{i+j}), \tag{15}$$

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,

$$\eta_{i+j} = \gamma \cdot \epsilon^{i+j}. \tag{16}$$

The hyperparameter γ controls the overall amplitude of the variance. The hyperparameter ϵ controls the rate at which the prior variance decays with increasing polynomial order. A higher value of γ or ϵ results in a prior that favors more rapidly varying covariance matrix fields, while a lower value favors smoother fields. By setting $\gamma = 0.0003$ and $\epsilon = 0.4$, we adopted a prior that favors relatively smooth variation across the space (Figure 4, fourth panel).

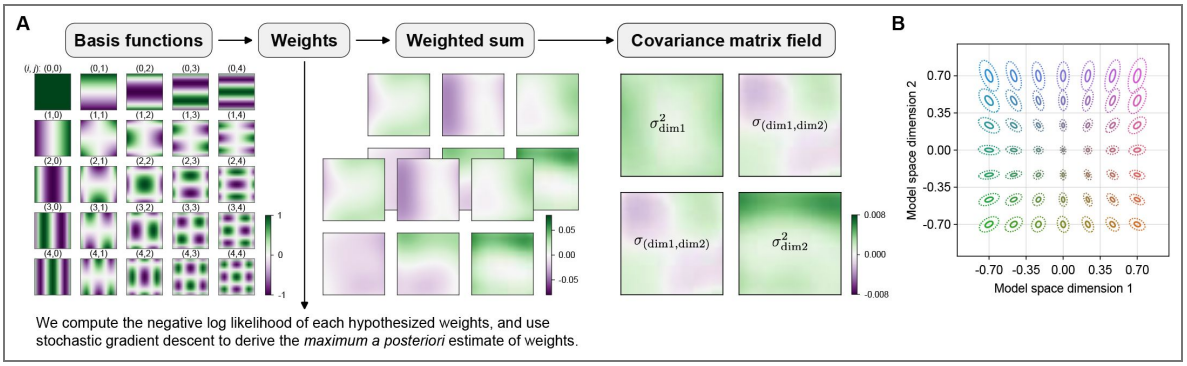

Figure 4. The finite-basis Wishart Process Psychophysical Model (WPPM).

(A) Model overview. In our implementation, we use a set of 5 × 5 two-dimensional Chebyshev polynomial basis functions, denoted $\Phi_{i,j}(x)$, where $i, j \in \{0, 1, ..., 4\}$. These basis functions are summed using a learnable weight matrix **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 $\mathbf{U}_{k,l}$ is 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 $(\sigma^2_{\text{dim}\,1}, \sigma^2_{\text{dim}\,2})$ and their covariance ($\sigma_{\text{dim1,dim2}}$). For this example covariance matrix field, the weights used to generate the field correspond to the best-fitting parameters for participant CH (see *Figure S3* for all participants). (B) Model readouts. Internal noise can be read out anywhere in the model space, illustrated here on a grid of 7 × 7 reference stimuli (solid lines), from which threshold contours (dashed lines) can be derived.

Model fitting

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

$$p_r\left(y_1,\ldots,y_r \mid \mathbf{W}\right)=\sum_{r=1}^{R}\left(y_r\cdot\log\left(p_r\right)+\left(1-y_r\right)\cdot\log\left(1-p_r\right)\right), \tag{17}$$

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:

$$p_r=\mathbf{Pr}\left[d_M^2\left(z_0,z_0^{'}\right)<\min\left(d_M^2\left(z_0,z_1\right),d_M^2\left(z_0^{'},z_1\right)\right)\mid\mathbf{W}\right]. \tag{18}$$

Note that on the r^{th} 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 (Equation 17) and the prior (Equation 16) to calculate the posterior probability of $\mathbf{W}$. As there is no simple closed form expression for p_r, 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 Appendix 12).

Notably, the factorization in Equation 14 is not unique; multiple choices of U and consequently of W can yield the same covariance matrix. This non-uniqueness reflects the overcomplete parameterization and does not affect the uniqueness of the resulting covariance matrices or the corresponding threshold readouts of the model.

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 (Figure 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.

Data analysis

Color calibration analyses were performed using MATLAB 2023b. We computed inverse gamma lookup tables from the measured gamma functions (Appendix 11) and derived transformation matrices to convert values from the model space to RGB space (Appendix 1). Stimulus presentation, including gamma correction, was implemented in Unity, coded in C#.

All analyses were conducted in Python 3.11 using a variety of open-source packages. Model fitting was implemented primarily using JAX (Bradbury et al., 2018). Behavioral data were separated into AEPsych-driven and 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 120 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 120 bootstrapped datasets.

To compute a 95% bootstrap confidence interval on the threshold contours, we first computed the summed Normalized Bures Similarity (NBS) score (Muzellec and Cuturi, 2018) between the predictions from the model fit to each bootstrapped dataset and those from the model fit to the original dataset, evaluated on a finely sampled grid of reference stimuli (-0.85 to 0.85 with 103 uniformly spaced steps). Higher scores indicate greater similarity to the predictions from the original dataset. We then sorted the model fits by their summed NBS scores and retained the top 114 (95% of 120) fits. The confidence interval bounds were defined by the union and intersection of the threshold contours, subsequently computed for any reference stimulus using this fixed set of retained model fits.

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. Threshold was 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 agreement between the thresholds predicted by the WPPM and those estimated from the validation trials, we performed linear regression (constrained to pass through the origin) between the two sets of predictions using the original dataset, as well as for each of the 120 paired bootstrapped datasets. We then sorted the resulting slopes and correlation coefficients and computed 95% confidence intervals separately for each. Additionally, we computed the number of conditions for which the 95% bootstrap confidence intervals of the WPPM-predicted thresholds and the validation thresholds overlapped as an additional measure of agreement.

Data availability

Data (*https://osf.io/k27js/overview*) and code (*https://github.com/fh862/ellipsoids_eLife2025.git*) are publicly available.

Acknowledgements

We thank Nicolas P. Cottaris for his assistance with calibration, our colleagues at the UPenn Vision Labs, Laurence Maloney, and Karl Gegenfurtner for helpful feedback.

Additional information

Author contribution

F.H.: Conceptualization; Formal analysis; Investigation; Methodology; Software; Validation; Visualization; Writing – original draft; Writing – review & editing.

R.B.: Investigation; Project administration.

J.C.: Methodology; Software; Writing – review & editing.

C.S.: Methodology; Software.

M.S.: Methodology; Software; Writing – review & editing.

P.G.: Conceptualization; Methodology; Resources; Software; Supervision; Validation; Writing – review & editing.

A.H.W.: Conceptualization; Methodology; Resources; Software; Supervision; Validation; Writing – original draft; Writing – review & editing.

D.H.B.: Conceptualization; Funding acquisition; Methodology; Project administration; Resources; Software; Supervision; Validation; Writing – original draft; Writing – review & editing.

Funding

Funder	Grant reference number
Meta	

Author ORCID iDs

Fangfang Hong: https://orcid.org/0000-0003-1890-1977
David H Brainard: https://orcid.org/0000-0001-9827-543X

Appendix 1 Transformation between the DKL, RGB, and model spaces

This section summarizes the colorimetric transformations between the RGB space of our monitor, the 2D model space, and standard color spaces.

The DKL space provides a representation of the isoluminant plane with the adapting point at the origin (Derrington et al., 1984; Brainard, 1996). We defined our DKL space with respect to the CIE physiologically-relevant 2-degree cone fundamentals and corresponding photopic luminosity function. We began in DKL space, with adapting point defined by the cone excitations elicited by the displayed background so that the space's isoluminant plane included the background. In this plane, we then densely sampled chromatic directions spanning 360° around the origin. For each direction we marched outward from the origin to find the edge of the monitor's gamut in that direction (details explained in Appendix 1.1). Repeating across directions, we obtained a set of gamut boundary points that define a quadrilateral in the isoluminant plane. We then identified the four vertices and recorded their coordinates in both DKL and RGB spaces (Table S1). Using these vertices, we derived a projective transformation matrix (homography) that maps coordinates from the DKL space to the model space (see Appendix 1.2 and Table S2). While the homography provides a general solution applicable to any quadrilateral, our derived matrix revealed that an affine transformation provided an accurate approximation. We used this affine transformation and its inverse to convert back and forth between linear RGB values and the model space (see Appendix 1.3 and Table S2).

Appendix 1.1: Search for the boundary points within the monitor's gamut

We selected 1,000 angles θ that span the isoluminant plane uniformly in the DKL representation. For each angle, we defined a chromatic direction vector in DKL space as $\mathbf{d}_{\mathrm{DKL}} = [\cos(\theta), \sin(\theta), 0]^{\top}$, where the first two elements correspond to the L–M and S axes of the DKL space, and the third element is set to zero to constrain the direction to the isoluminant plane (i.e., no change in luminance). For each direction, we then determined the farthest point along that vector that remained within the monitor's gamut in linear RGB space. Specifically, we first converted $\mathbf{d}_{\mathrm{DKL}}$ to LMS cone excitations, denoted $\mathbf{e}_{\mathrm{LMS}}$, as follows:

$$\mathbf{e}_{\mathrm{LMS}} = M_{\text{cone contrast}\,\rightarrow\mathrm{LMS}} \cdot M_{\Delta\,\mathrm{LMS}\rightarrow\,\text{cone contrast}} \cdot M_{\mathrm{DKL}\rightarrow\Delta\,\mathrm{LMS}} \cdot \mathbf{d}_{\mathrm{DKL}}. \tag{S1}$$

Notably, because the actual LMS cone excitations include contributions from ambient light, denoted as $\mathbf{e}_{\mathrm{LMS,\ ambient}}$, we subtracted it to isolate the portion due to the RGB stimulus, denoted as $\mathbf{e}_{\mathrm{LMS,\ stimulus}}$, that is,

$$\mathbf{e}_{\mathrm{LMS,\ stimulus}} = \mathbf{e}_{\mathrm{LMS}} - \mathbf{e}_{\mathrm{LMS,\ ambient}}. \tag{S2}$$

Next, we converted this isolated stimulus response into RGB space:

$$\mathbf{d}_{\mathrm{RGB}} = M_{\mathrm{LMS}\to\mathrm{RGB}} \cdot \mathbf{e}_{\mathrm{LMS,\,stimulus}} - M_{\mathrm{LMS}\to\mathrm{RGB}} \cdot \mathbf{e}_{\mathrm{LMS,\,background}} \tag{S3}$$

$$= M_{\mathrm{LMS}\to\mathrm{RGB}} \cdot \mathbf{e}_{\mathrm{LMS,\,stimulus}} - [0.5, 0.5, 0.5]^{\top}. \tag{S4}$$

Finally, we marched outward along the direction $\mathbf{d}_{\mathrm{RGB}}$ until the RGB values reached the edge of the RGB cube. The values at this boundary were recorded as a limiting point along that chromatic direction. Repeating this procedure across all sampled directions yielded the full boundary of the isoluminant plane constrained by the monitor's gamut. From this boundary set, we then identified the four corner vertices (Table S1).

Corner	DKL_{L-M}	DKL_{S}	DKL_{Lum}	L	M	S	R	G	B	W_{dim1}	W_{dim2}
1	-0.123	-0.812	0	0.145	0.147	0.016	0.000	0.733	0.000	-1	-1
2	0.175	-0.830	0	0.164	0.111	0.014	1.000	0.407	0.000	1	-1
3	-0.175	0.830	0	0.142	0.154	0.152	0.000	0.593	1.000	-1	1
4	0.123	0.812	0	0.160	0.117	0.150	1.000	0.267	1.000	1	1

Table S1. Corner vertices in the DKL, LMS, RGB, and model spaces.

Appendix 1.2: An affine transformation matrix that maps DKL to model space

These vertices, denoted as $\mathbf{v}$, were then used to derive a projective transformation matrix $M_{\mathrm{DKL}\to\mathrm{W}}$ such that for each vertex pair, we have:

$$\mathbf{v}_{\mathrm{W}} = M_{\mathrm{DKL}\to\mathrm{W}} \cdot \mathbf{v}_{\mathrm{DKL}}, \tag{S5}$$

where $\mathbf{v}_{\mathrm{W}} = [v_{\mathrm{W,\,dim1}}, v_{\mathrm{W,\,dim2}}, 1]^{\top}$ is the homogeneous coordinate of a vertex in model space, and $\mathbf{v}_{\mathrm{DKL}} = [v_{\mathrm{DKL,}\,L\text{-}M}, v_{\mathrm{DKL,}\,S}, 0]^{\top}$ is the corresponding homogeneous coordinate in DKL space. By plugging in the vertices, we solved the matrix $M_{\mathrm{DKL}\to\mathrm{W}}$ as the following,

$$M_{\mathrm{DKL}\to\mathrm{W}} = \begin{bmatrix} - & \mathbf{v}_{\mathrm{W,dim\,1}} & - \\ - & \mathbf{v}_{\mathrm{W,dim\,2}} & - \\ - & \mathbf{1} & - \end{bmatrix} \cdot \begin{bmatrix} - & \mathbf{v}_{\mathrm{DKL,\,L-M}} & - \\ - & \mathbf{v}_{\mathrm{DKL,\,s}} & - \\ - & \mathbf{0} & - \end{bmatrix}^{\dagger}, \tag{S6}$$

where † denotes the pseudoinverse. Note that the last row of $M_{\mathrm{DKL}\to\mathrm{W}}$ is [0, 0, 1], indicating that the transformation is affine (Table S2). Although this affine formulation would be sufficient, we initially computed the full projective transformation matrix for generality, since it was uncertain that the DKL vertices would form a parallelogram. In this particular case, both methods yielded equivalent results.

$M_{\mathrm{DKL}\rightarrow\mathrm{W}}$			$M_{\mathrm{RGB}\rightarrow\mathrm{W}}$		
6.724	0.213	0.000	1.556	−1.364	−0.192
0.076	1.221	0.000	−0.444	−1.364	1.808
0.000	0.000	1.000	0.444	1.364	0.192

Table S2. **Transformation matrices between DKL, RGB and model spaces.**

Appendix 1.3: An affine transformation matrix that maps RGB to model space

Given that the transformations from DKL to LMS, LMS to RGB, and DKL to model space are all affine, by the composition property of affine transformations, it follows that the transformation from RGB to model space must also be affine.

To compute the affine transformation matrix, we used corresponding corner vertices in RGB and model spaces. Specifically, we solved for the matrix $M_{\mathrm{RGB}\rightarrow\mathrm{W}}$ as:

$$M_{\mathrm{RGB}\rightarrow\mathrm{W}} = \begin{bmatrix} — & \mathbf{v}_{\mathrm{W},\dim 1} & — \\ — & \mathbf{v}_{\mathrm{W},\dim 2} & — \\ — & \mathbf{1} & — \end{bmatrix} \cdot \begin{bmatrix} — & \mathbf{v}_{\mathrm{R}} & — \\ — & \mathbf{v}_{\mathrm{G}} & — \\ — & \mathbf{v}_{\mathrm{B}} & — \end{bmatrix}^{\dagger}, \tag{S7}$$

where † denotes the pseudoinverse and we appended a row of ones to the Wishart coordinates to express them in homogeneous form.

Appendix 1.4: Affine invariance of Mahalanobis distance

We performed trial placement, model fitting, and data presentation using the model space (bounded between –1 and 1). An important feature of the WPPM is that it is equivariant with respect to affine transformations of the color space used to represent the stimuli. That is, if we transform reference and comparison stimuli to a new color space using an affine transformation, and transform the covariance field by the same affine transformation, then the observer model yields a prediction of performance that is unchanged by the transformation. This is because the Mahalanobis distance is itself unchanged by the transformation, as we show below. This is an attractive property because it avoids assigning special status to the particular color space used to represent the stimuli and covariance field.

Let Σ be the covariance matrix. The squared Mahalanobis distance between two points $\mathbf{x}_0$ and $\mathbf{x}_1$ is defined as:

$$d^2\left(\mathbf{x}_0, \mathbf{x}_1\right) = \left(\mathbf{x}_0 - \mathbf{x}_1\right)^{\top} \Sigma^{-1} \left(\mathbf{x}_0 - \mathbf{x}_1\right) \tag{S8}$$

Now consider a linear transformation $\mathbf{x}_0^{'} = A\mathbf{x}_0, \mathbf{x}_1^{'} = A\mathbf{x}_1$. The corresponding transformation of the covariance matrix is $\Sigma^{'} = A\Sigma A^{\top}$. Then the squared Mahalanobis distance in the transformed space becomes:

$$
\begin{aligned}
d^2\left(\mathbf{x}_0^{'}, \mathbf{x}_1^{'}\right) &= d^2\left(A\mathbf{x}_0, A\mathbf{x}_1\right) \\
&= \left(A\mathbf{x}_0 - A\mathbf{x}_1\right)^\top \cdot \left(A\Sigma A^\top\right)^{-1} \cdot \left(A\mathbf{x}_0 - A\mathbf{x}_1\right) \\
&= \left(\mathbf{x}_0 - \mathbf{x}_1\right)^\top \cdot A^\top \cdot \left(A\Sigma A^\top\right)^{-1} \cdot A \cdot \left(\mathbf{x}_0 - \mathbf{x}_1\right) \\
&= \left(\mathbf{x}_0 - \mathbf{x}_1\right)^\top \cdot A^\top \cdot \left(A^\top\right)^{-1} \cdot \Sigma^{-1} \cdot A^{-1} \cdot A \cdot \left(\mathbf{x}_0 - \mathbf{x}_1\right) \\
&= \left(\mathbf{x}_0 - \mathbf{x}_1\right)^\top \cdot \Sigma^{-1} \cdot \left(\mathbf{x}_0 - \mathbf{x}_1\right) \\
&= d^2\left(\mathbf{x}_0, \mathbf{x}_1\right)
\end{aligned}
\tag{S9}
$$

Thus, the Mahalanobis distance is invariant under linear transformations of the data when the covariance matrix is transformed accordingly. Since distance is also invariant to translations (i.e., independent of the choice of origin), this further implies that the Mahalanobis distance is invariant under general affine transformations.

Appendix 2 Adaptive sampling and WPPM estimates

Appendix 2.1: AEPsych-driven trials and WPPM-predicted thresholds

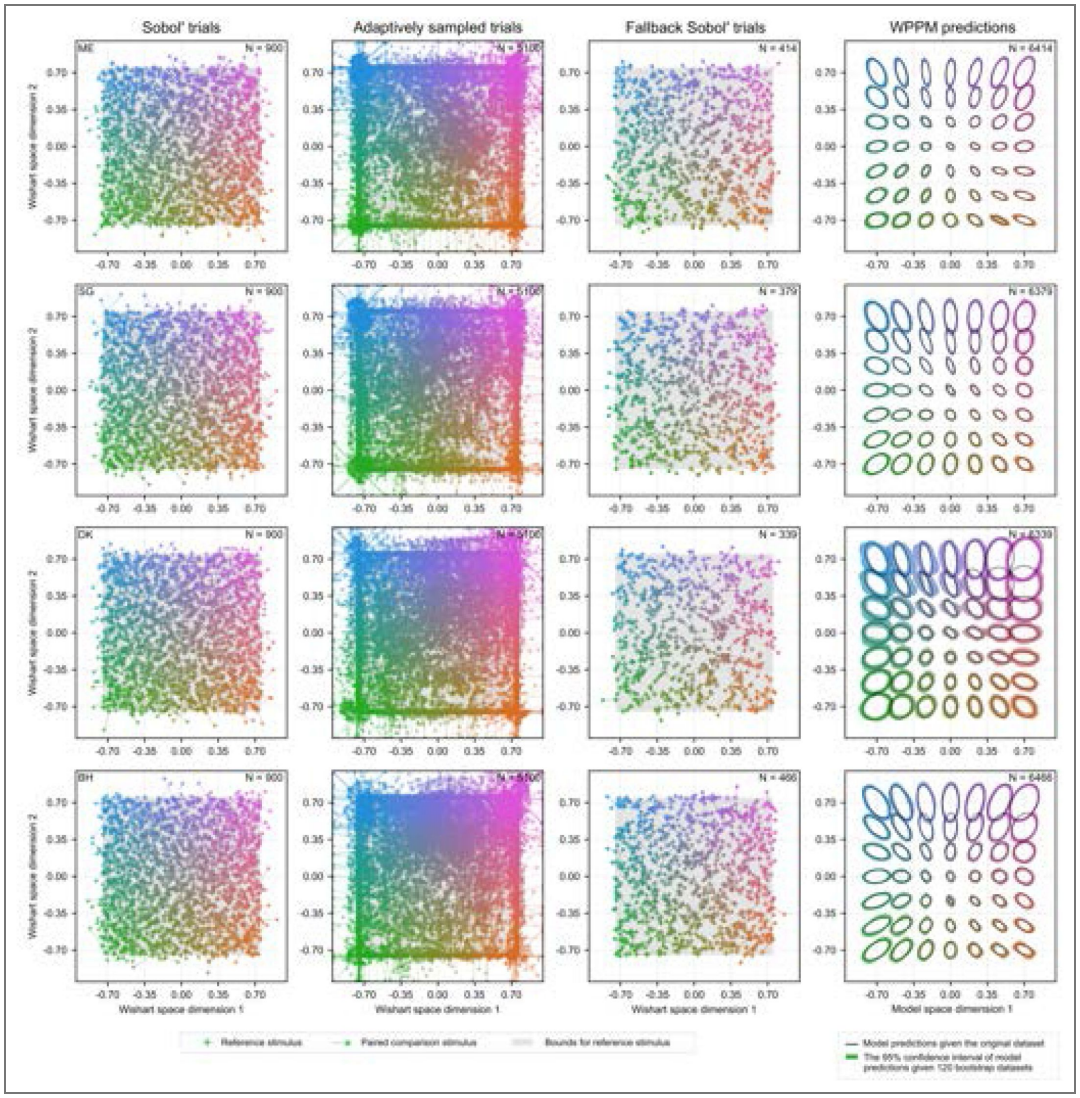

Figure S1. AEPsych-driven trials (900 Sobol'-sampled and 5,100 adaptively sampled), fallback trials, and WPPM predictions for all participants. Each row represents data from one participant.

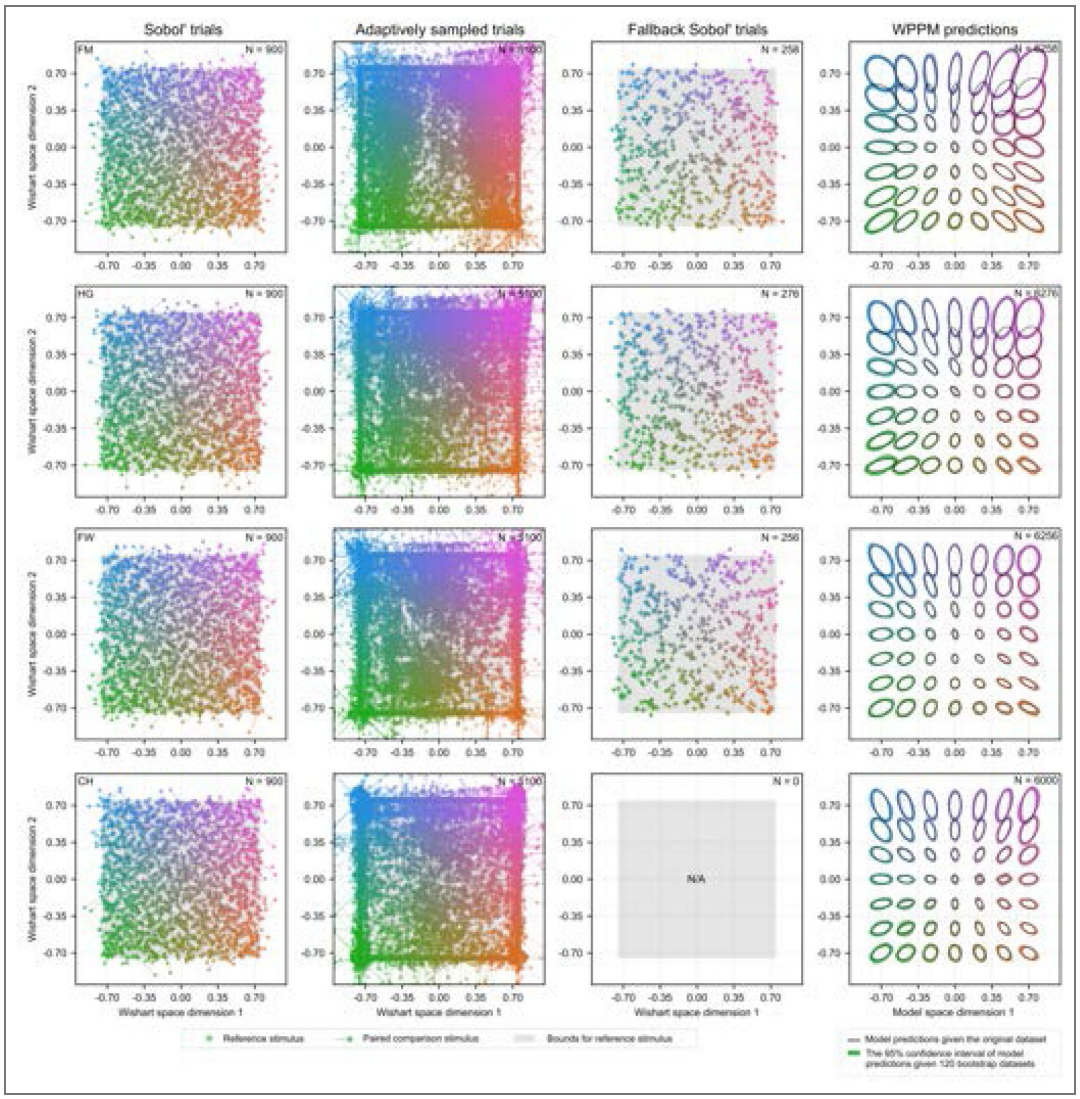

Figure S1. AEPsych-driven trials (900 Sobol'-sampled and 5,100 adaptively sampled), fallback trials, and WPPM predictions for all participants (continued). Each row represents data from one participant. Note that for participant CH, no pre-generated Sobol' trials were used, as the fallback strategy was implemented later in the study to maintain experimental continuity and reduce delays between trials.

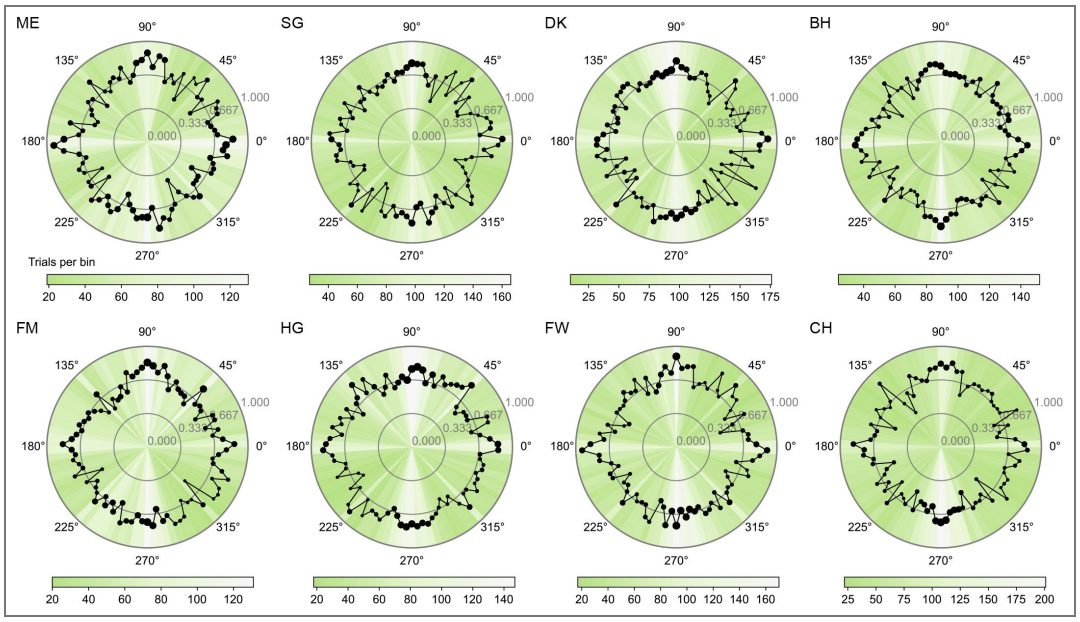

Figure S2. Percent correct as a function of the angular difference between reference and comparison stimuli for all participants. The number of trials within each bin (bin width = 4°) is encoded by both marker size and color, with larger markers and brighter colors indicating a greater number of trials.

Appendix 2.2: Efficacy of adaptive sampling

After the initial 900 Sobol'-sampled trials, the subsequent 5,100 trials were adaptively placed using AEPsych (Owen et al., 2021; Letham et al., 2022) to target the 66.7% performance threshold. To evaluate the efficiency of this adaptive sampling procedure, we binned these adaptive trials based on the angular difference between the reference and comparison stimuli and computed the proportion of correct responses within each bin. If adaptive sampling is effective, we expect performance to cluster around 66.7% correct. Consistent with this, the observed percent correct remained close to 66.7% correct across bins (Figure S2), suggesting that AEPsych provided informative trial placement.

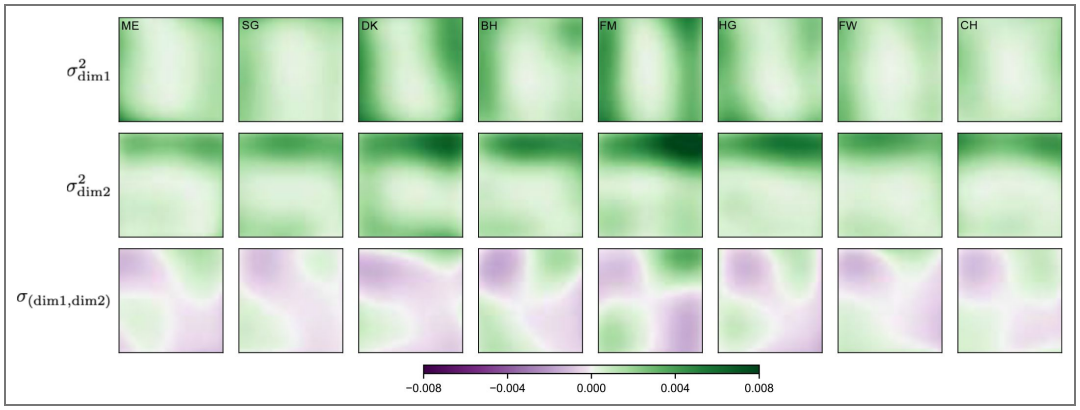

Figure S3. Covariance matrix fields obtained from the best-fitting WPPM for all participants. Rows show (top to bottom): noise variance along the first model dimension, noise variance along the second model dimension, and the covariance between the two dimensions.

Appendix 2.3: The covariance matrix field

The WPPM specifies a covariance matrix at each reference stimulus across the isoluminant plane. Each matrix characterizes internal noise in terms of the variance along the two model dimensions $(\sigma^2_{\text{dim}\,1}, \sigma^2_{\text{dim}\,2})$ and their covariance ($\sigma_{\text{dim1,dim2}}$). For each participant, the best-fitting model

produces a covariance matrix field that exhibits qualitatively similar structure across individuals (Figure S3).

Appendix 2.4: The best-fitting weight matrices

The covariance matrix field is determined by the weights applied to the Chebyshev basis functions. Across participants, the absolute magnitude of the best-fitting weights decreases with increasing basis order, indicating that higher-order components contribute less to the covariance field. At the highest order we included, the weights are close to zero. This indicates that given the prior hyperparameters, adding more basis functions would be unlikely to change the estimates we obtained.

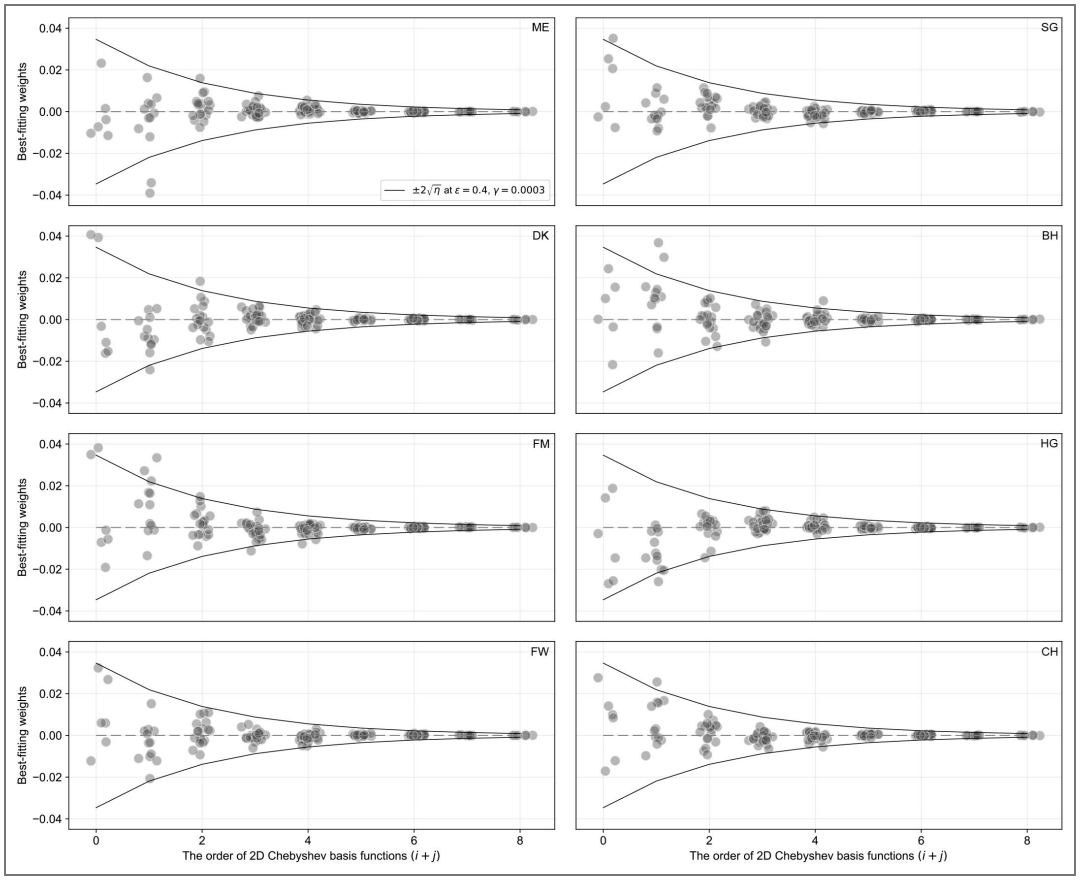

Figure S4. The best-fitting weights for all participants. The symmetric solid curves represent the prior imposed on the model. The prior is implemented by specifying the variance of the weights, η, as a function of the polynomial order of the Chebyshev basis functions and two hyperparameters ϵ and γ (Equation 16). The solid curves indicate $\pm 2\sqrt{n}$ for our hyperparameter choice, corresponding to two standard deviations of the prior distribution.

Appendix 3 Real-time trial scheduling via dual-computer coordination

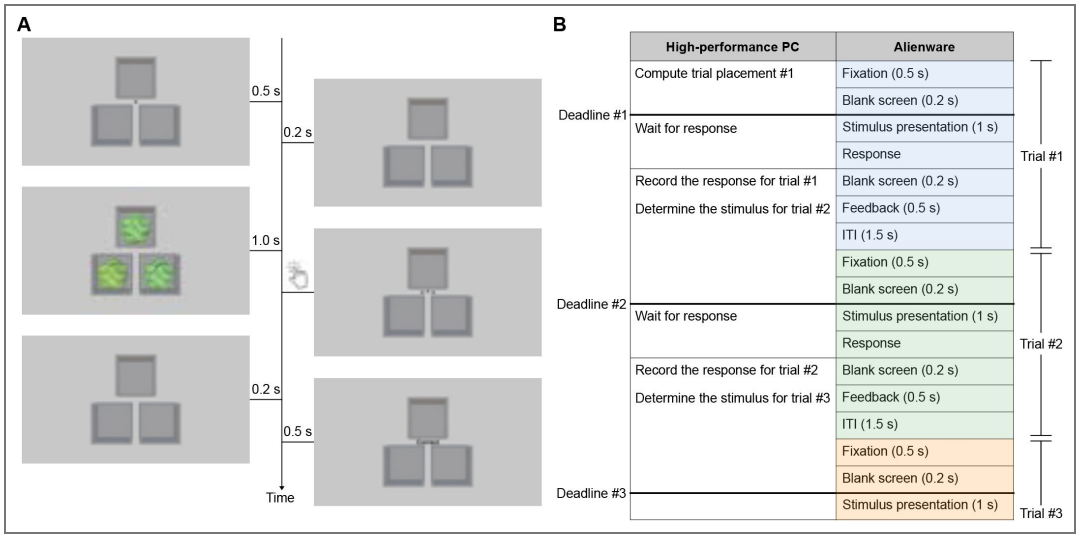

Figure S5. Task timing and real-time trial scheduling. (A) Trial sequence: a 0.5 s fixation cross was followed by a 0.2 s blank interval, then a 1 s presentation of three blobby stimuli. Participants responded at their own pace to identify the odd-one-out, after which a 0.2 s blank screen and a 0.5 s feedback were shown. The inter-trial interval (ITI) was 1.5 s. (B) A schematic representation of the trial timing and computational responsibilities of the two computers.

For each participant, we ran 6,000 AEPsych-driven trials interleaved with an additional 6,000 validation trials. Although our initial plan was to present these 12,000 trials in a predetermined randomized sequence, we quickly realized this approach was impractical. Under such a design, AEPsych would have only the inter-trial interval (ITI) to compute the next trial placement—a window that is difficult to optimize. A long ITI risks participant fatigue or loss of attention, while a short ITI does not provide AEPsych with enough time to complete its computations. To achieve both a smooth experimental flow and adequate computation time for AEPsych, we implemented a fallback trial strategy using a dual-computer setup.

In this setup, stimulus presentation was handled by an Alienware computer, while adaptive trial placement using AEPsych ran on a separate high-performance PC. The two systems communicated via a shared network disk using a custom protocol based on text files that both computers could read and write. This decoupled design provided modular separation between code specialized for stimulus presentation and the sequence of events on each trial and code specialized for trial placement, and should allow our trial placement code to be more easily ported to different stimulus display systems.

With the dual-computer design, AEPsych had at least 2.9 s to compute the next trial after the participant's response (Figure S5). This window spanned both the post-stimulus period of the current trial (0.2 s blank, 0.5 s feedback, 1.5 s ITI) and the pre-stimulus period of the upcoming trial (0.5 s fixation and 0.2 s blank). Importantly, this computation window began only after the participant responded—since AEPsych requires the participant's response to update its model—and ended just before the stimulus presentation of the next trial, when AEPsych must deliver the RGB values for the upcoming stimuli.

The fallback trial strategy ensured continuous stimulus presentation. If AEPsych failed to return a new trial within the 2.9 s window, we defaulted to presenting the next available MOCS trial from the pre-determined randomized sequence. In such cases, AEPsych's computation continued in a different thread and attempted to meet the following decision deadline, which is approximately 7

s after the previous one, including an additional 1 s stimulus presentation and an estimated 0.2 s response time. If AEPsych again missed this deadline, the next opportunity came at around 11.1 s. This staggered scheduling ensured that trials continued smoothly while allowing AEPsych sufficient time to compute adaptive placements when possible.

A potential drawback of the fallback trial strategy is that it could disrupt the intended interleaving of adaptive and validation trials, potentially introducing differential learning effects. To mitigate this, we capped how far MOCS trials could advance relative to AEPsych trials. This cap was set at four trials. If this limit was reached and no AEPsych trial was ready, we inserted pre-generated Sobol' trials instead. These Sobol' trials were created in advance using participant- and session-specific random seeds and were separate from those selected by AEPsych.

Appendix 4 Comparison between WPPM and validation thresholds

Appendix 4.1: Validation data for all participants

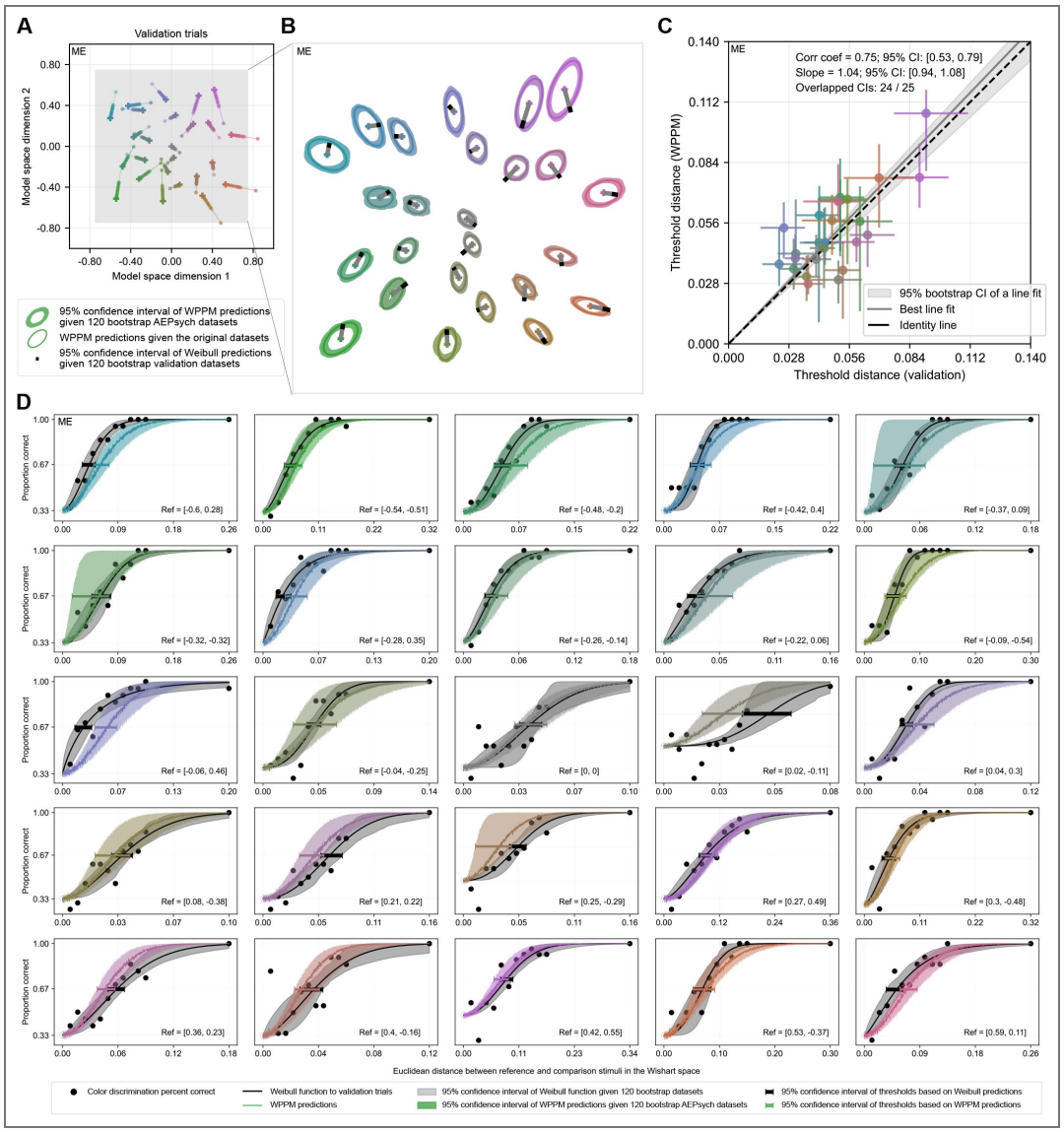

Figure S6. Validation for participant ME. Same format as Figure 2D-G in the main text.

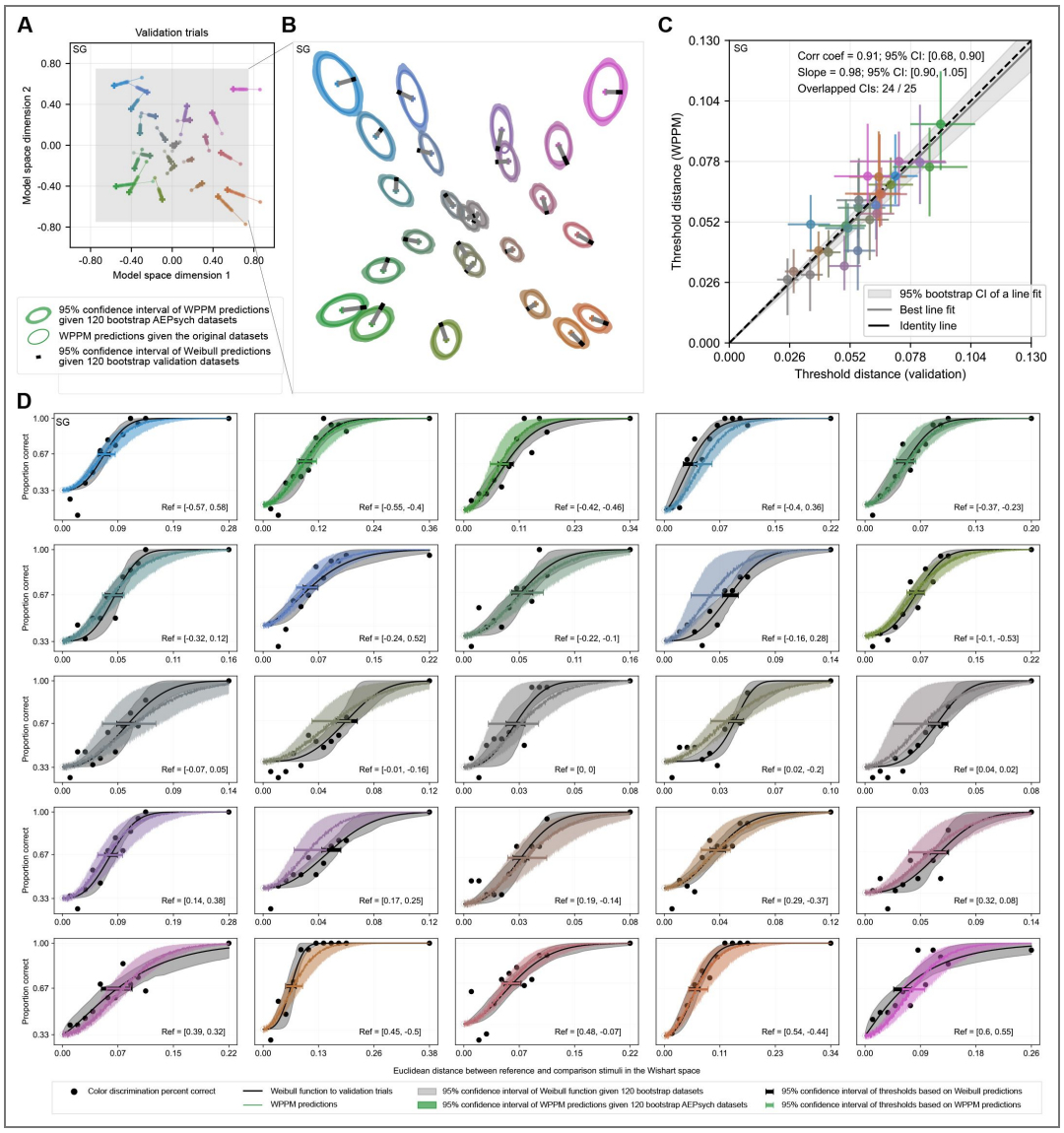

Figure S7. **Validation for participant SG.**

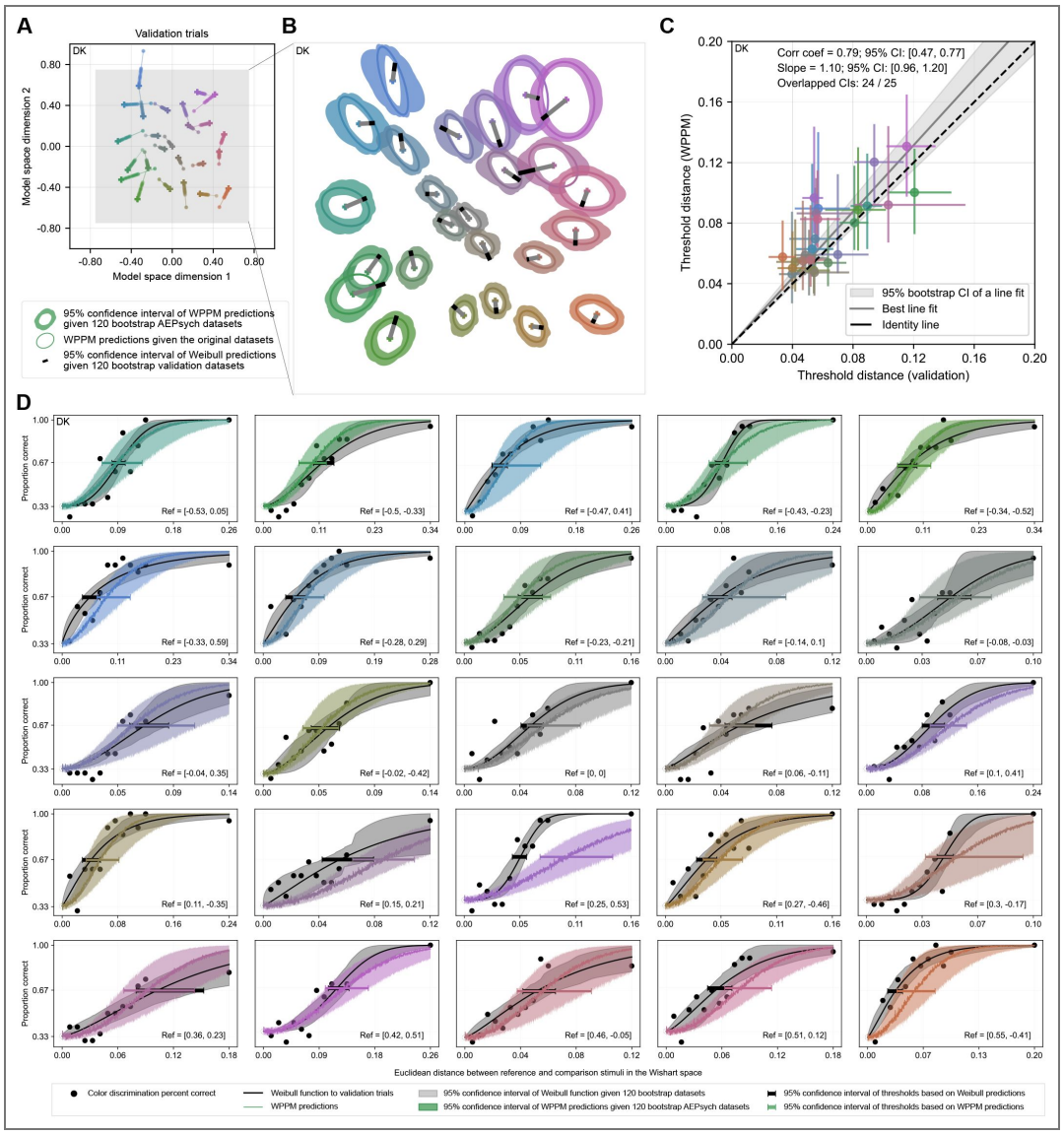

Figure S8. **Validation for participant DK.**

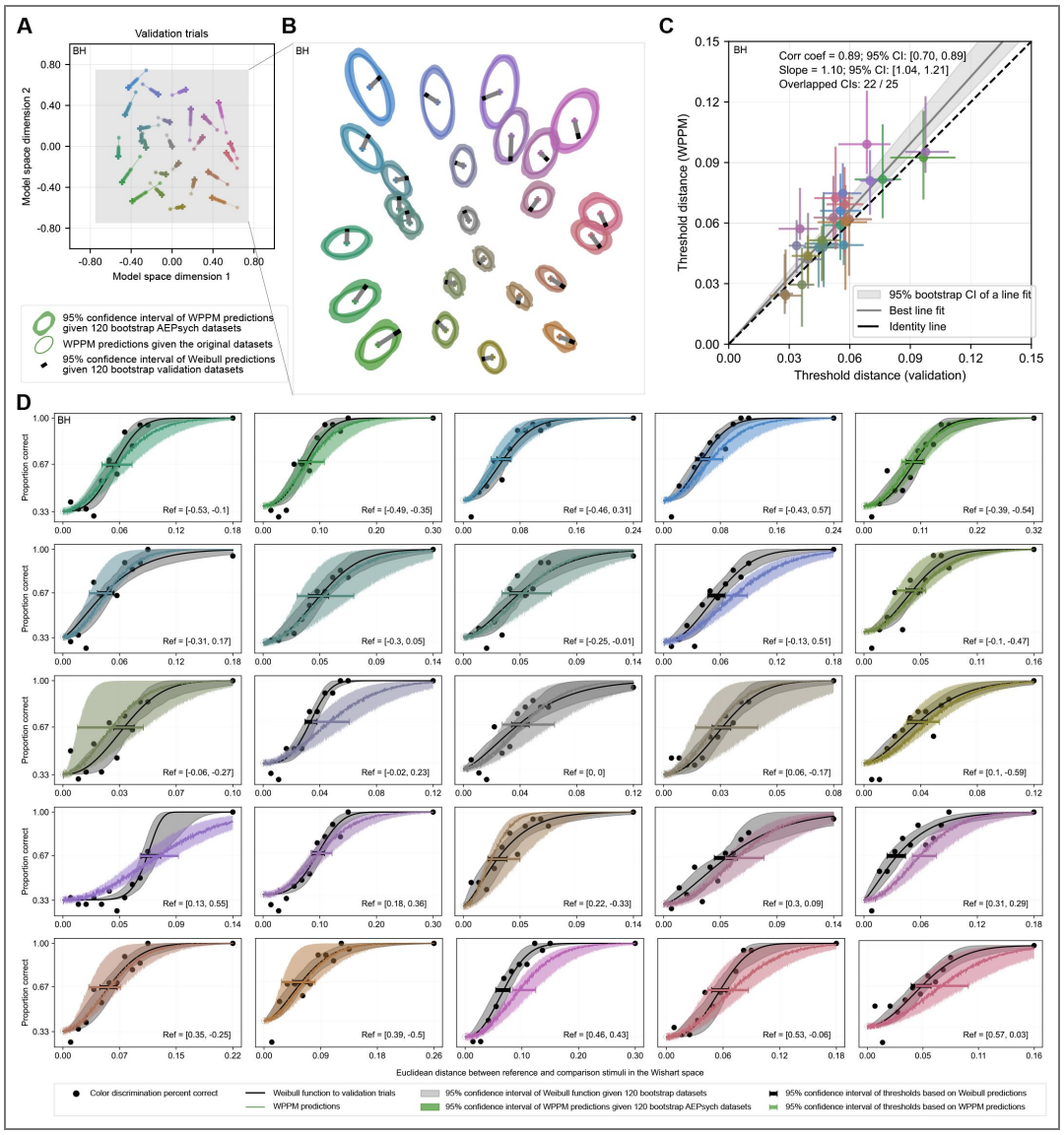

Figure S9. **Validation for participant BH.**

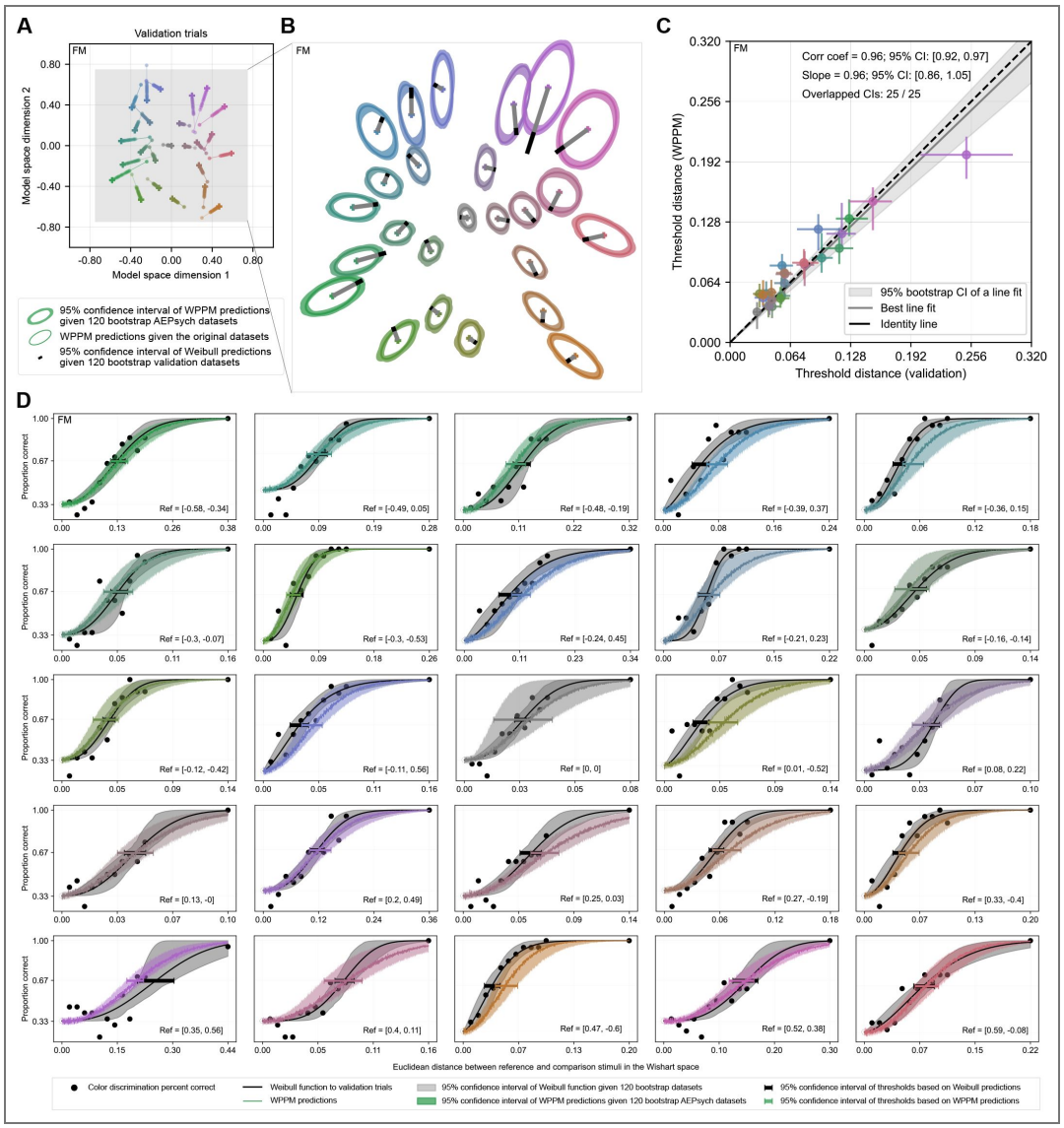

Figure S10. **Validation for participant FM.**

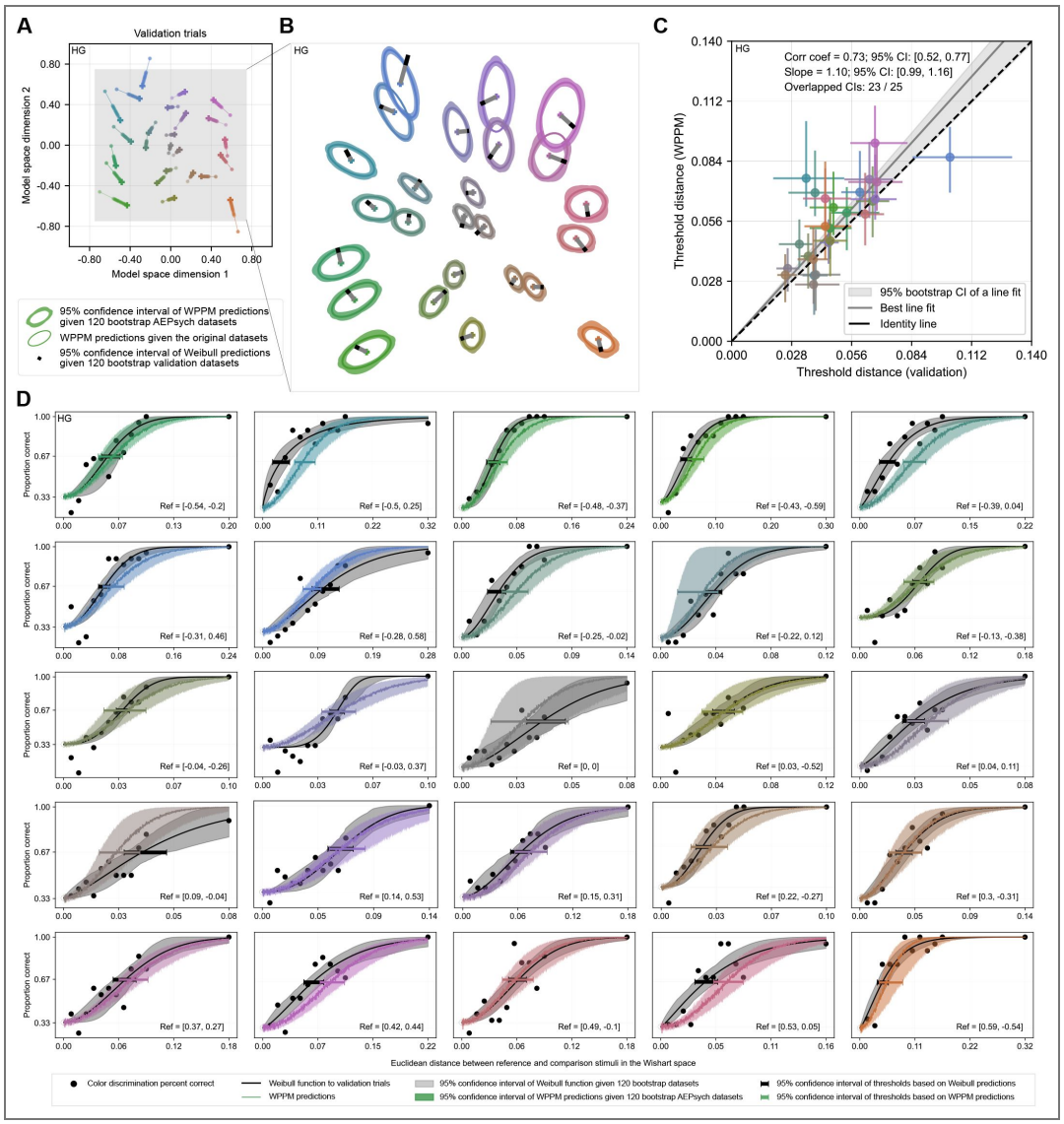

Figure S11. Validation for participant HG.

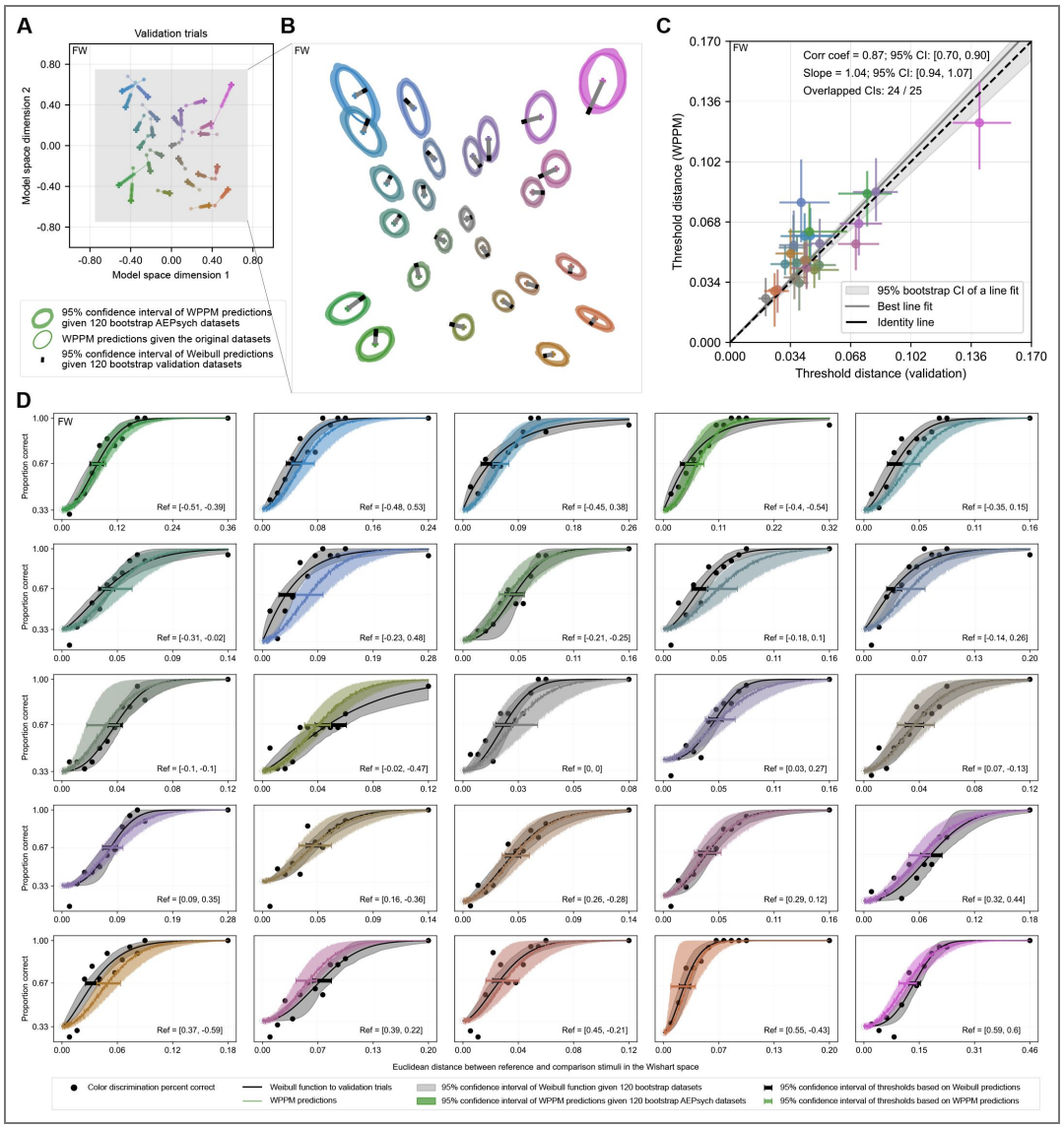

Figure S12. Validation for participant FW.

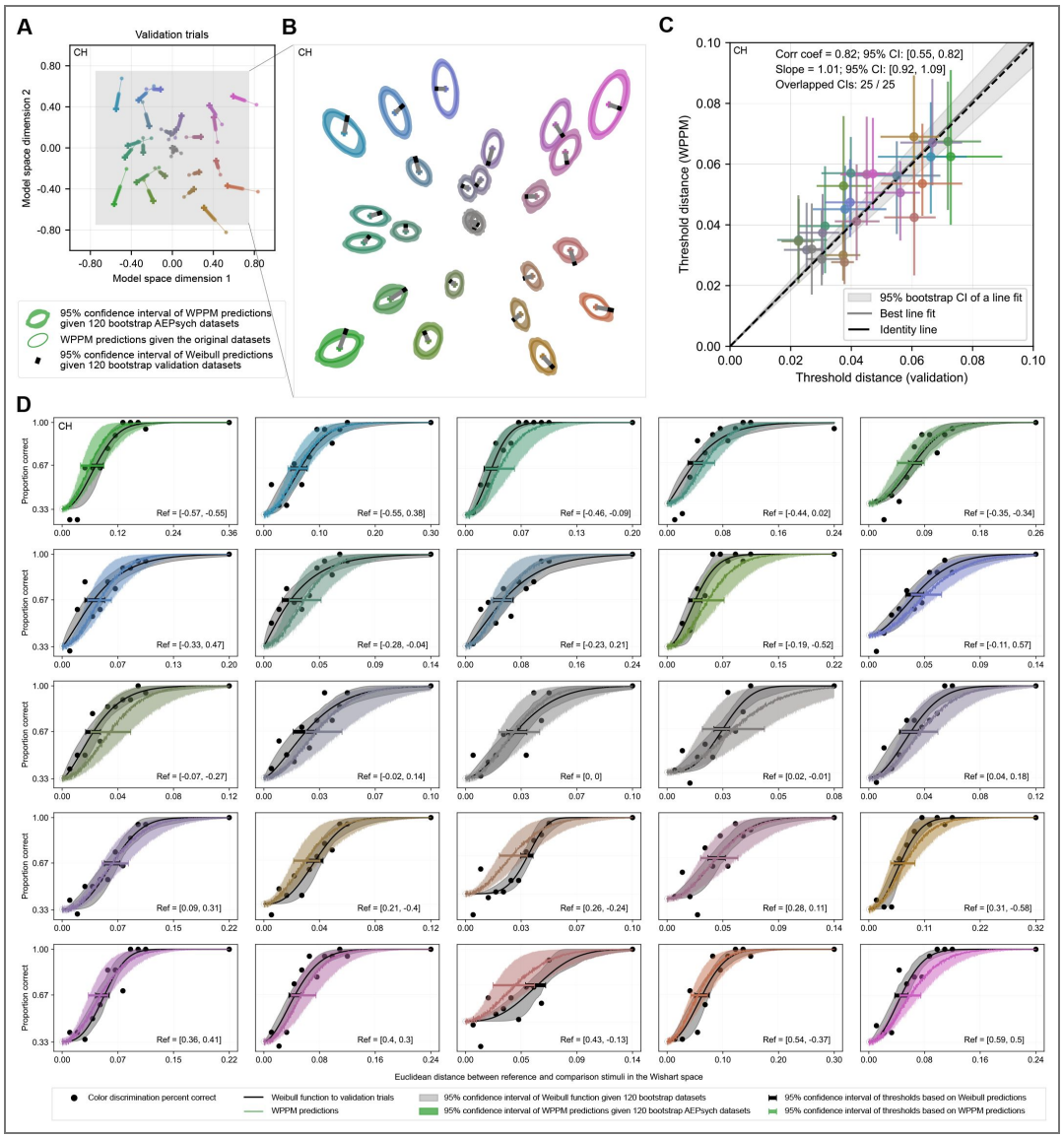

Figure S13. Validation for participant CH.

Appendix 4.2: Analysis of differences between WPPM and validation thresholds

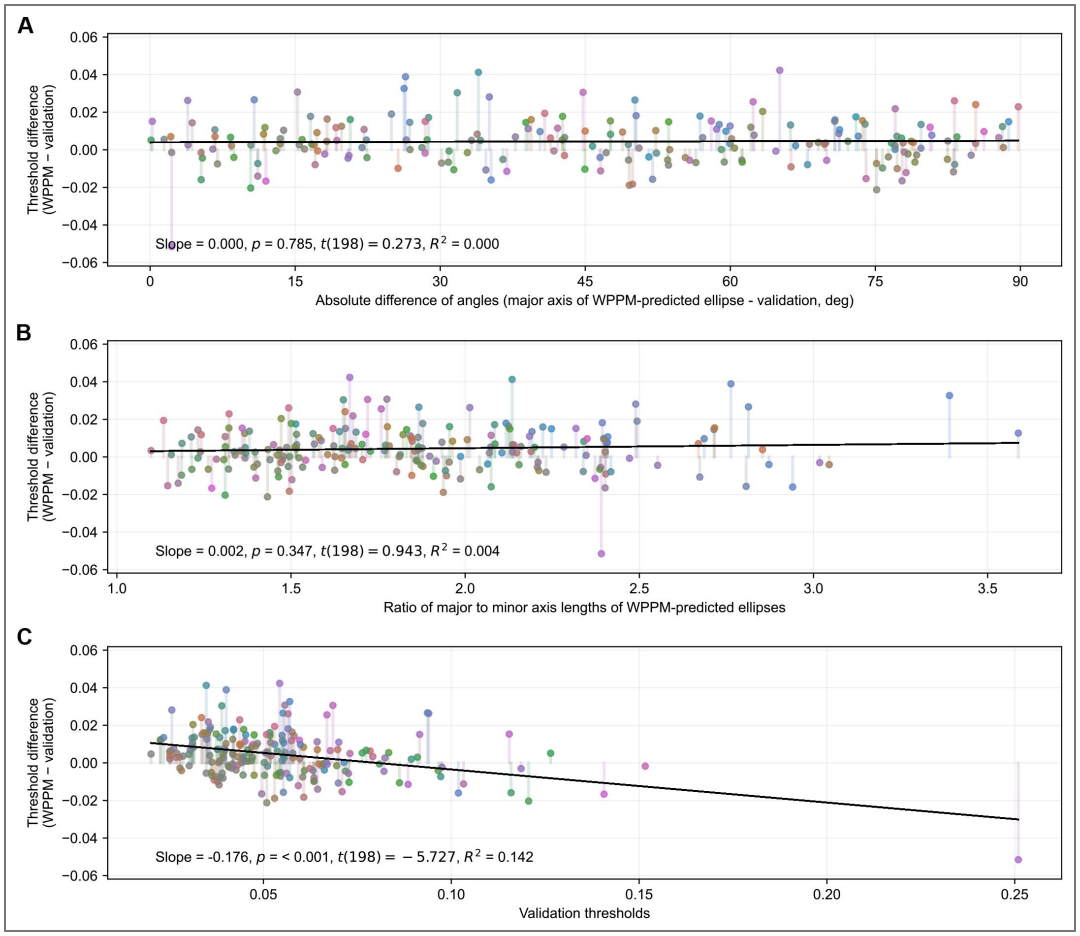

Figure S14. Threshold residuals. Data are pooled across all validation conditions and all participants (N = 8). For all panels, color codes for the surface color of the reference stimulus, and the y-axis limits are set to ± the mean of the validation thresholds. (A) Residuals as a function of the absolute angular difference between the major axis of the elliptical threshold contours read out from the WPPM fits and the chromatic direction of the validation condition. (B) Residuals as a function of the aspect ratio (major/minor axis) of the WPPM threshold contours. (C) Residuals as a function of thresholds estimated from validation trials.

We assessed whether the residuals (the differences between the WPPM and validation thresholds) exhibited systematic patterns. We found no significant correlation between the residuals and the absolute angular difference between the chromatic direction of the validation condition and the major axis of the elliptical threshold contours read out from the WPPM fits (Figure S14A), nor with the aspect ratio of the contours (Figure S14B). Thus, there is no evidence that the residuals vary systematically with the orientation or shape of the contours read out from the WPPM fits (see statistical summary in Table S3).

In contrast, we found a significant negative correlation between the residuals and the magnitude of the validation thresholds (Figure S14C; slope = −0.176, $t(198) = -5.727$, $p < 0.001$, $R^2 = 0.142$), indicating that the WPPM tends to slightly overestimate thresholds when they are small and underestimate them when they are large. However, the magnitude of this bias is small relative to the range of observed validation thresholds.

Predictor	Term	Coef	Std Err	t	p	[0.025, 0.975] CI	R^2
Absolute difference of angles	intercept	0.004	0.002	2.254	0.025	[0.000, 0.007]	0.000
	slope	0.000	0.000	0.273	0.785	[-0.000, 0.000]	
Aspect ratio	intercept	0.001	0.004	0.319	0.750	[-0.006, 0.008]	0.004
	slope	0.002	0.002	0.943	0.347	[-0.002, 0.005]	
Validation thresholds	intercept	0.014	0.002	7.511	0.000	[0.010, 0.018]	0.142
	slope	-0.176	0.031	-5.727	0.000	[-0.237, -0.116]	

Table S3. Linear regression results assessing the relationship between WPPM–validation threshold residuals and three predictors:(1) the absolute angular difference between the chromatic direction of the validation condition and the major axis of the contours read out from the WPPM fits, (2) the aspect ratio of the contours, and (3) the magnitude of the validation threshold. This analysis was done on human data.

Appendix 4.3: Analysis of percent-correct performance for catch trials

For each validation condition, we used the method of constant stimuli to sample 12 comparison levels: 11 were evenly spaced, and one was selected to serve as an easily discriminable catch trial. Participants completed 500 catch trials (1/12 of 6,000 validation trials). These catch trials were included to assess participants' attentiveness and establish a criterion for potential data exclusion. As shown in Table S4, participants except for DK performed near ceiling on these trials, indicating high task engagement throughout the experiment. Although DK's performance was somewhat lower, this likely reflects lower overall sensitivity rather than frequent lapses (Figure S8), as the "easy" trials may not have been as easily discriminable for this participant.

Participant	ME	SG	DK	BH	FM	HG	FW	CH
Proportion correct	0.996	0.996	0.948	0.992	0.998	0.988	0.988	0.998
Lower bound	0.977	0.974	0.868	0.975	0.975	0.954	0.957	0.980
Upper bound	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Table S4. Catch trial performance summary across all sessions. The proportion correct reflects the total number of correct responses divided by the total number of catch trials. Lower and upper bounds indicate the participant's lowest and highest session-level performance, respectively.

Appendix 5 Simulated observer

To evaluate how well thresholds read out from the WPPM fits aligned with those estimated via Weibull fits, we simulated a dataset with a known ground truth. The following subsections outline the key steps in this process.

Appendix 5.1: Derivation of the comparison stimuli at threshold on the isoluminant plane

We used CIELab $\Delta E94$ as the ground truth metric for deriving color discrimination performance. For any given reference color and any given chromatic direction, both were affine-transformed from the model space to the RGB space. The RGB values were then converted to CIE 1931 XYZ and then to CIELab space, where ΔE computations were performed. In the XYZ-to-Lab transformation, we used the monitor gray point ($R = G = B = 0.5$) as the reference white. We then searched along each chromatic direction in the RGB space to find a comparison stimulus $\mathbf{x}_1 = (x_{1,\text{dim1}}, x_{1,\text{dim2}})$ such that ΔE in CIELab was equal to 2.5 (Figure S15A). This procedure was repeated across multiple directions. The resulting comparison stimuli were then mapped back into the model space, where we fit an ellipse to define the iso-distance contour (Figure S15B).

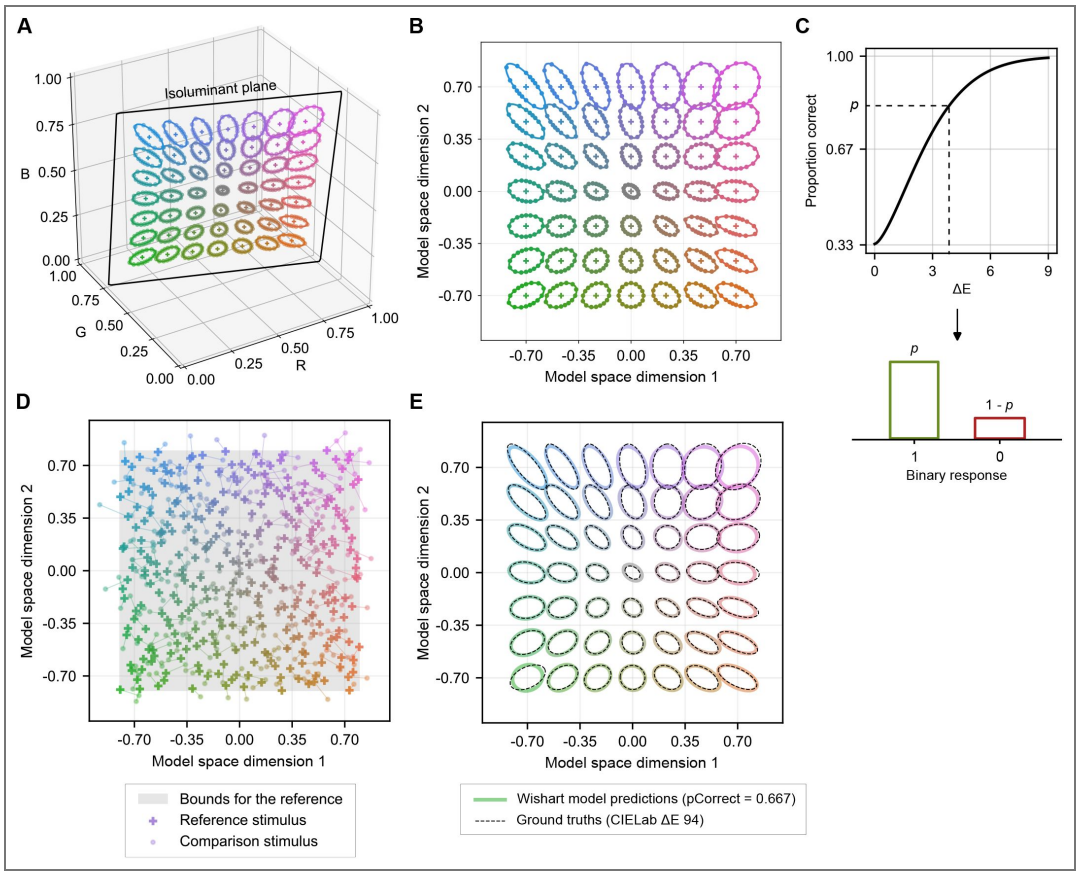

Figure S15. Derivation of the ground-truth Wishart fits based on CIELab Δ*E*94. (A–B) Comparison stimuli at the iso-distance contours in the isoluminant plane, shown in both RGB and model spaces. Note that the reference grid and fixed set of directions shown here are for illustration only; the actual sampling did not use a fixed grid or evenly spaced chromatic directions. (C) The Weibull psychometric function used to simulate binary (correct or incorrect) responses given Δ*E* values. (D) Sampled reference-comparison stimulus pairs. Reference colors and chromatic directions were sampled using Sobol' sequences, and comparison stimuli were jittered around the iso-distance contour. A total of 18,000 trials were simulated; only the first 200 are shown here for clarity. (E) Comparison between readouts from the WPPM fit and from CIELab Δ*E*94. The WPPM fit was subsequently treated as the ground truth for simulating AEPsych and validation trials.

Appendix 5.2: Simulation of trials near threshold contours

To introduce some variability, we added bivariate Gaussian noise to each comparison stimulus at the iso-distance contour in the model space. The noise standard deviation was proportional to the Euclidean distance between the reference stimulus $\mathbf{x}_0$ and the comparison stimulus $\mathbf{x}_1$. The jittered comparison stimulus $\mathbf{x}_1'$ was computed as:

$$\mathbf{x}_1' = \mathbf{x}_1 + \epsilon, \quad \epsilon \sim \mathcal{N}\left(\mathbf{0}, 0.3 \cdot |\mathbf{x}_1 - \mathbf{x}_0|^2 \cdot \mathbf{I}\right). \tag{S10}$$

We modeled performance using a Weibull psychometric function, which took the Δ*E* between the reference and jittered comparison stimuli as input and returned the predicted percent correct:

$$\Psi(\Delta E) = \gamma + (1-\gamma)\left(1 - e^{-(\Delta E/\alpha)^{\beta}}\right) = \frac{1}{3} + \frac{2}{3}\left(1 - e^{-(\Delta E/3.189)^{1.505}}\right). \tag{S11}$$

The values of α and β were selected such that the psychometric curve returns 66.7% correct when Δ*E* = 2.5 (Figure S15C). A binary (correct or incorrect) response was sampled from a Bernoulli distribution using this predicted probability:

$$r \sim \text{Bernoulli}(\Psi(\Delta E)) \tag{S12}$$

The comparison stimuli were selected to be near threshold, while the reference stimuli and chromatic directions were Sobol' sampled to ensure uniform coverage of the model space without repeated trials (Figure S15D). In total, we simulated 18,000 trials.

Appendix 5.3: Fit the WPPM and treating the model fits as the ground truth

We fitted the WPPM to the full set of 18,000 trials in the model space, and treated the resulting fit as the ground truth for simulating performance for both AEPsych and validation trials (Figure S15E, color lines). We chose to use the WPPM fit as the ground truth—rather than percent-correct performance derived from CIELab Δ*E*94 with a Weibull psychometric function—because our goal here was to evaluate how well the WPPM can recover ground truth that is itself described by the WPPM.

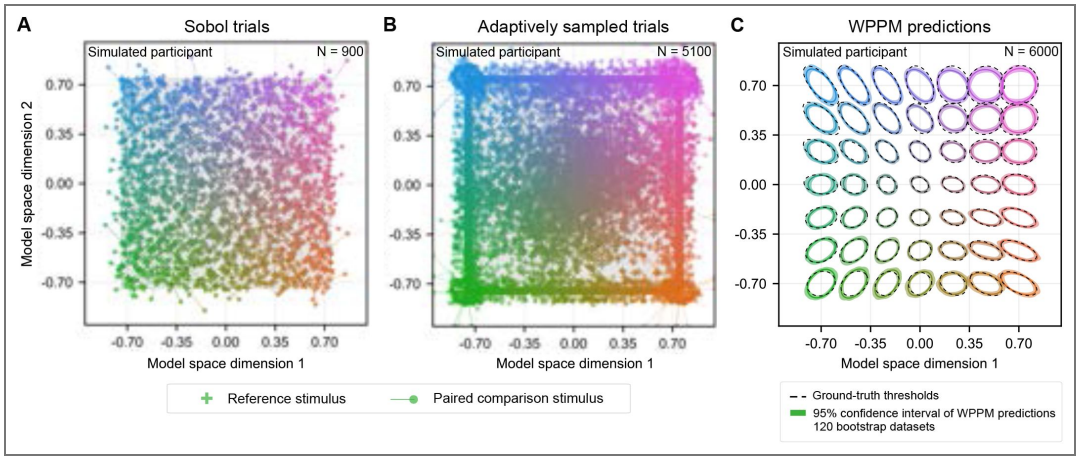

Figure S16. AEPsych-driven trials and WPPM readouts for a simulated observer. Note that the ground-truth thresholds shown in (C) is the same WPPM readouts from Figure S15E.

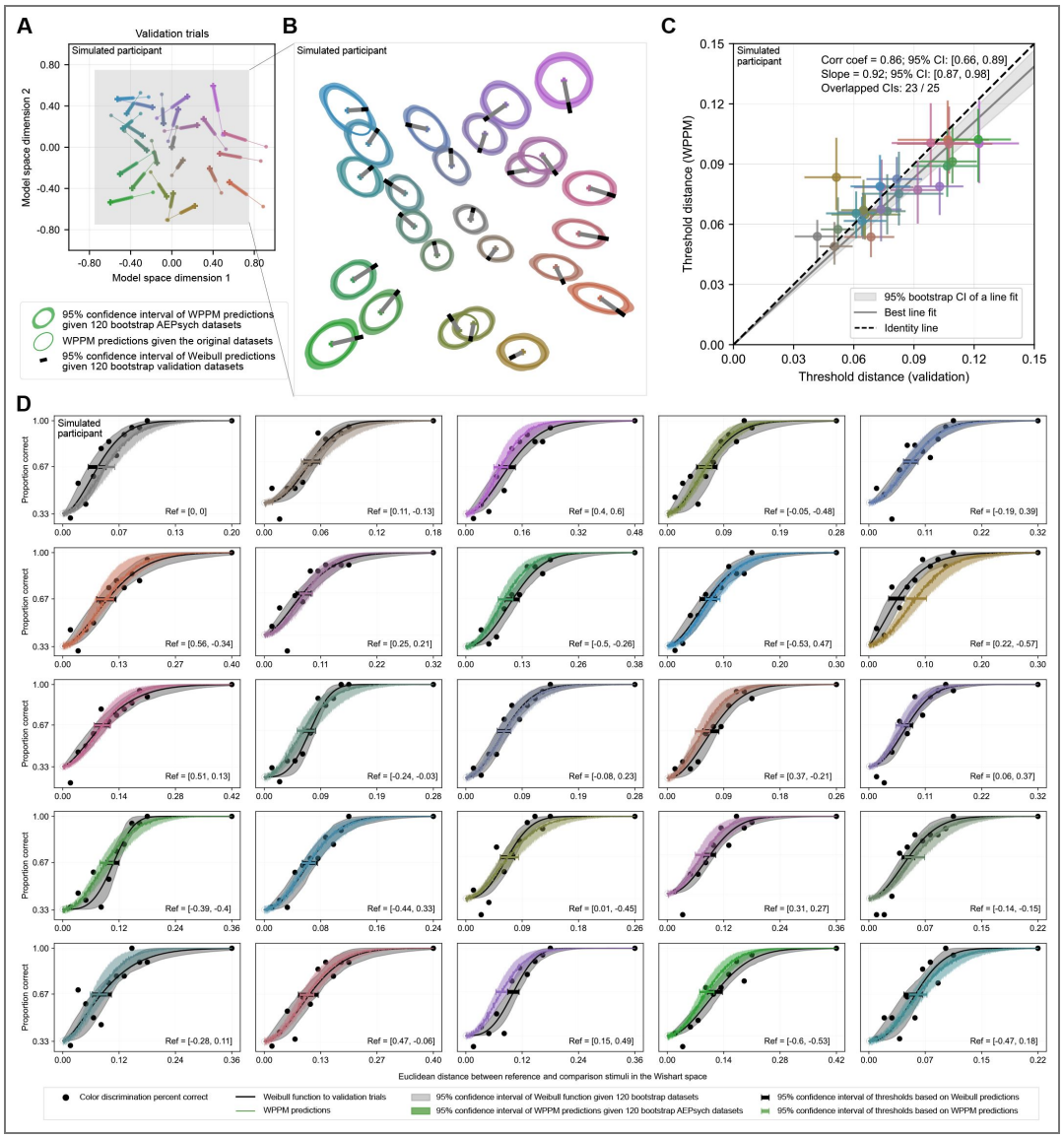

Figure S17. **Validation trials and WPPM readouts for a simulated observer.**

Predictor	Term	Coef	Std Err	t	p	[0.025, 0.975] CI	R^2
Absolute difference of angles	intercept	-0.007	0.005	-1.532	0.139	[-0.017, 0.003]	0.028
	slope	0.000	0.000	0.812	0.425	[-0.000, 0.000]	
Aspect ratio	intercept	-0.007	0.012	-0.579	0.568	[-0.033, 0.019]	0.003
	slope	0.002	0.008	0.263	0.795	[-0.014, 0.018]	
Validation thresholds	intercept	0.028	0.006	4.429	0.000	[0.015, 0.041]	0.547
	slope	-0.393	0.075	-5.273	0.000	[-0.547, -0.239]	

Table S5. **Linear regression results for the simulated dataset.**

Appendix 5.4: Fit the WPPM to simulated AEPsych trials

Based on the ground-truth WPPM fit, we simulated 900 Sobol' trials (Figure S16A), followed by 5,100 adaptively sampled trials using AEPsych (Figure S16B), just like the design for the actual experiment. For each pair of reference and comparison stimuli, we approximated percent-correct performance using Monte Carlo simulation (N = 2,000), and generated binary responses by drawing from a Bernoulli distribution. We then fit the WPPM to this simulated dataset. To approximate the variability of the WPPM readouts, we bootstrapped the data 120 times, maintaining the same Sobol'-to-adaptive trial ratio within each bootstrapped dataset. As shown in

Figure S16C, the WPPM was able to reliably recover the ground-truth model, with only minor deviations. This good agreement provides provides context for the analyses in the following subsections, which are then compared with the corresponding analyses of the human data.

Appendix 5.5: Validation trials and Weibull predictions

In addition to the 6,000 AEPsych trials, we also simulated 6,000 validation trials, mirroring the design of the actual experiment. Unlike the experimental design, these validation trials were simulated separately rather than interleaved, since sequential effects or perceptual learning are not factors in simulation. The WPPM thresholds confidence intervals agreed with 23 of 25 validation threshold confidence intervals (Figure S17). A linear regression fit to the validation thresholds (x-axis) and WPPM thresholds (y-axis) yielded a slope of 0.94 and a correlation coefficient of 0.83. These values fall within the range observed for human data (Appendix 4.1).

Appendix 5.6: Statistical analysis of residuals between WPPM readouts and validation thresholds

We applied the same statistical analysis to the simulated data as we did for the human data (***subsection***). Consistent with the human results (Figure S14), we found no strong evidence that residuals systematically varied with the orientation or shape of the elliptical threshold contours read out from the WPPM fits. However, we did observe a significant negative correlation between the residuals and the magnitude of the validation thresholds (slope = −0.393, $t(23) = -5.273$, $p < 0.001$, $R^2 = 0.547$; Figure S18; Table S5). As noted earlier, the size of this bias is small compared to the overall range of validation thresholds.

Appendix 5.7: Comparison between WPPM estimates and simulation ground truth

To evaluate whether the WPPM readouts systematically deviated from the ground truth, we sampled thresholds over a fine grid of reference locations (15 × 15 points evenly spaced between –0.7 and 0.7 in model space) and compared them with the corresponding ground-truth thresholds. As a comparison metric, we used the Bures-Wasserstein (BW) distance (Bhatia et al., 2019), which quantifies the dissimilarity between two positive semi-definite covariance matrices, Σ_1 and Σ_2. Intuitively, it captures the “effort” required to morph one ellipse into another. Mathematically, the BW distance is defined as

$$d\left(\Sigma_1, \Sigma_2\right) = \left[\operatorname{tr}\left(\Sigma_1\right) + \operatorname{tr}\left(\Sigma_2\right) - 2 \cdot \operatorname{tr}\left(\Sigma_1^{1/2} \cdot \Sigma_2 \cdot \Sigma_1^{1/2}\right)^{1/2}\right]^{1/2}. \tag{S13}$$

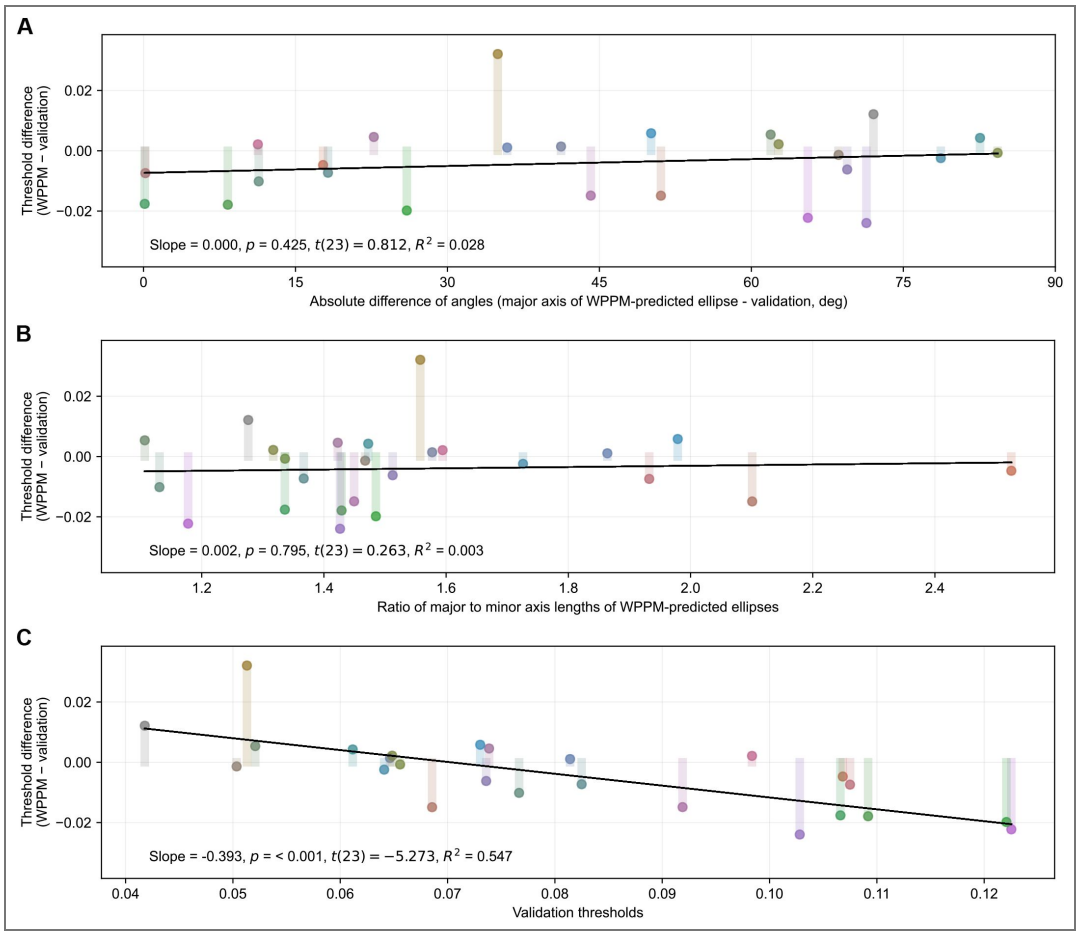

Figure S18. Threshold residuals for a simulated dataset. For all panels, color codes for the surface color of the reference stimulus, and the y-axis limits are set to ± the mean of the validation thresholds. (A) Residuals as a function of the absolute angular difference between the major axis of the elliptical threshold contours read out from the WPPM fits and the chromatic direction of the validation condition. (B) Residuals as a function of the aspect ratio (major/minor axis) of the WPPM threshold contours. (C) Residuals as a function of thresholds estimated from validation trials.

The BW distance is non-negative and equals zero only when the two matrices being compared are identical. Smaller distances indicate greater similarity between the threshold ellipses.

The results showed that BW distance generally increased as the reference color moved farther from the achromatic point (Figure S19A), suggesting that the WPPM has more difficulty accurately capturing large threshold contours in regions with higher internal noise. To provide a benchmark for what constitutes a substantial mismatch, we computed the BW distance between each ground-truth ellipse and a circle with radius being the largest major axis length among all ground-truth ellipses. The maximum of these values served as a reference point (shown as the upper limit of the color bar in Figure S19A). Overall, the mismatches observed in our simulations were modest— well below the level expected if the model were fundamentally mischaracterizing the threshold shapes.

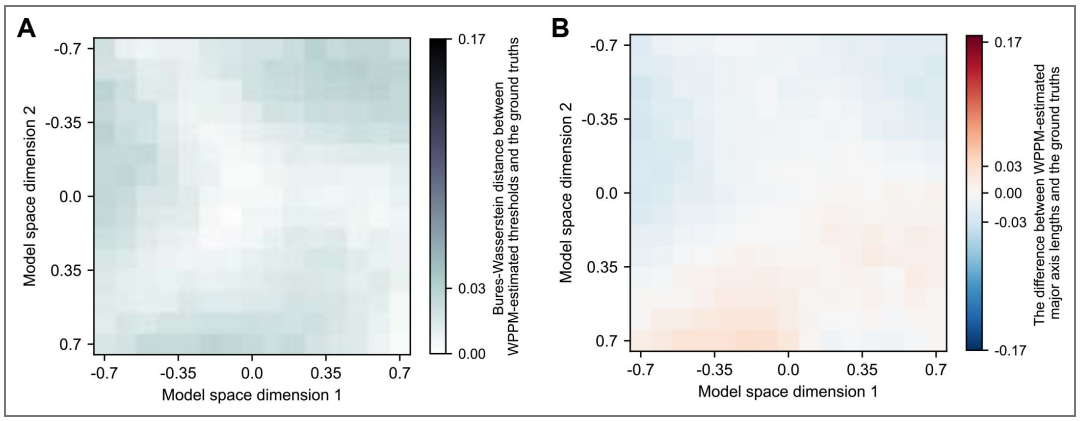

Figure S19. Deviation of WPPM estimates from the ground truth. (A) BW distance between WPPM-estimated thresholds and the ground-truth ellipses. The upper limit of the color map (0.17) corresponds to the maximum BW distance between each ground-truth ellipse and a reference circle whose radius equals the largest major axis length among all ground-truth ellipses. The maximum BW distance between WPPM estimates and the ground truth (0.03) is substantially lower than this reference value. (B) Difference in major axis length between WPPM-readouts and ground-truth ellipses. The colormap limits (±0.17) reflect the ± maximum ground-truth major axis length. Again, the maximum deviation observed (0.03) is small relative to this range.

We also examined differences in the estimated major axis lengths. The WPPM showed slight underestimation in the upper region and overestimation in the lower region of the space (Figure S19B; also apparent but small in Figure S16C). Similar to the BW analysis, these deviations were relatively small compared to the overall range of ground-truth values. Together, these results indicate that the WPPM provides a close and robust approximation of the true threshold contours, with only minor local deviations.

Appendix 6 Comparison with MacAdam ellipses (1942)

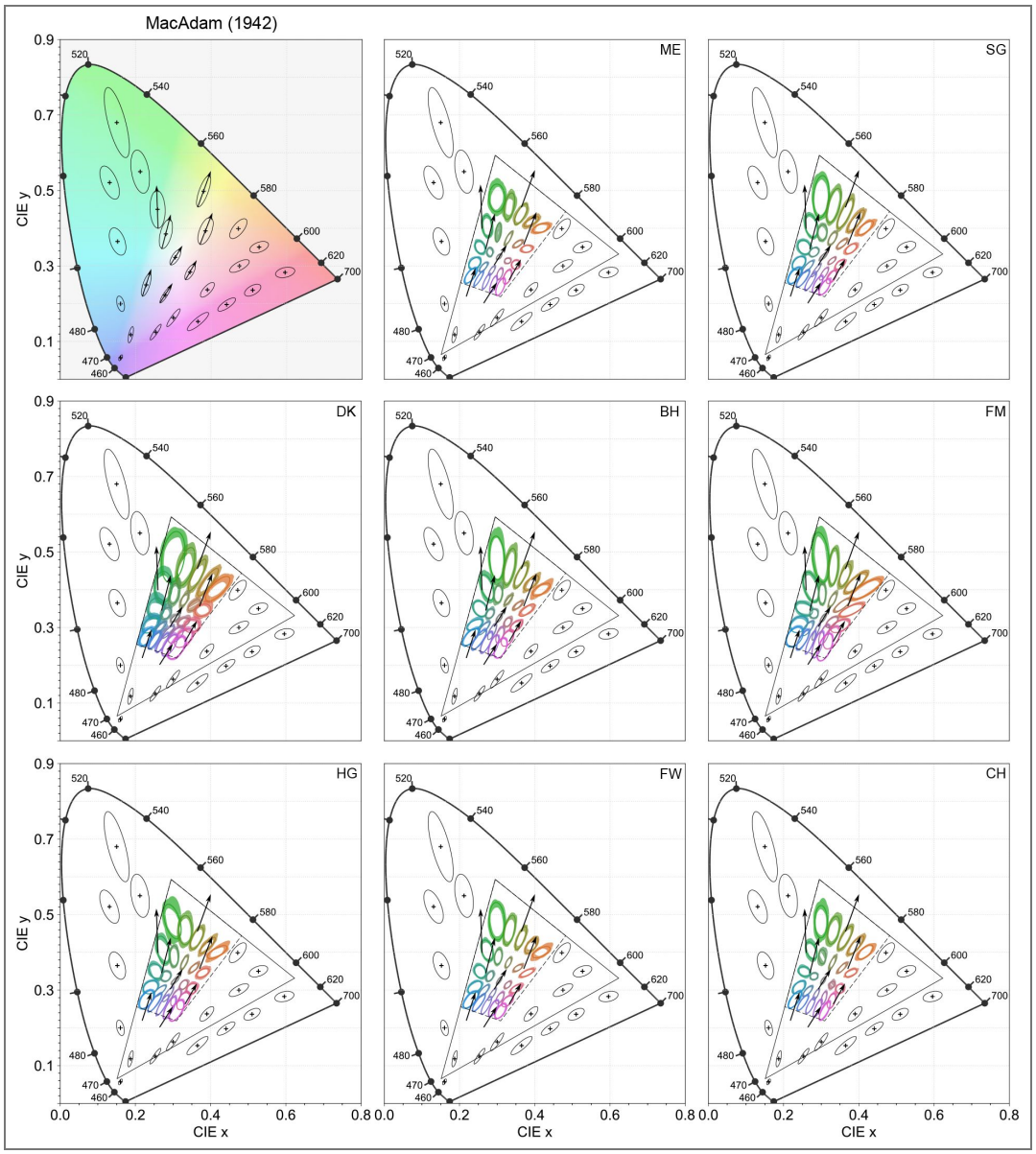

Figure S20. Comparison with MacAdam 1942. First panel: MacAdam's original threshold ellipses, magnified 10× for visualization. Rest panels: 66.7% threshold contours (colored lines) measured from all participants in our study and transformed from the model space into the CIE 1931 chromaticity diagram. Colored shaded regions indicate 95% confidence intervals computed from 120 bootstrapped datasets. Reference stimuli were sampled from a 5 × 5 grid spanning [−0.7, 0.7] along each dimension of the model space. To reduce visual clutter, MacAdam ellipses falling within the gamut of our isoluminant plane (parallelogram) are shown only by arrows indicating their major axes. For visual comparability, our ellipses are magnified 2× to approximately match the scale of MacAdam's data. The triangle indicates the monitor gamut.

Appendix 7 Comparison with Danilova & Mollon (2025)

In this section, we compare our measurements with those from Danilova and Mollon 2025 by transforming our results into the chromaticity space used in their study—a scaled version of the MacLeod–Boynton space (MacLeod and Boynton, 1979). While a direct transformation path

exists from our model space to theirs (model space → RGB → LMS → MacLeod–Boynton → scaled MacLeod–Boynton), it assumes that the adaptation point and isoluminant plane are identical between the two studies, which is not the case. To account for these differences, we instead took a detour through the DKL space (Derrington et al., 1984), where cone-opponent mechanisms are explicitly defined and adaptation is more easily controlled. Specifically, we followed the transformation chain: model space → RGB_{us} → LMS_{us} → ΔLMS_{us} → DKL → ΔLMS_{dm} → LMS_{dm} → MacLeod–Boynton → scaled MacLeod–Boynton. Here, the subscript "*us*" refers to values computed using our study's cone fundamentals, luminosity function and adaptation point, while "*dm*" denotes those based on Danilova and Mollon 2025. This approach allowed us to approximate how our stimuli would be represented in their perceptual framework, enabling visual comparison of the threshold contours. The comparison reveals a general qualitative agreement between their measurements and ours (Figure S21).

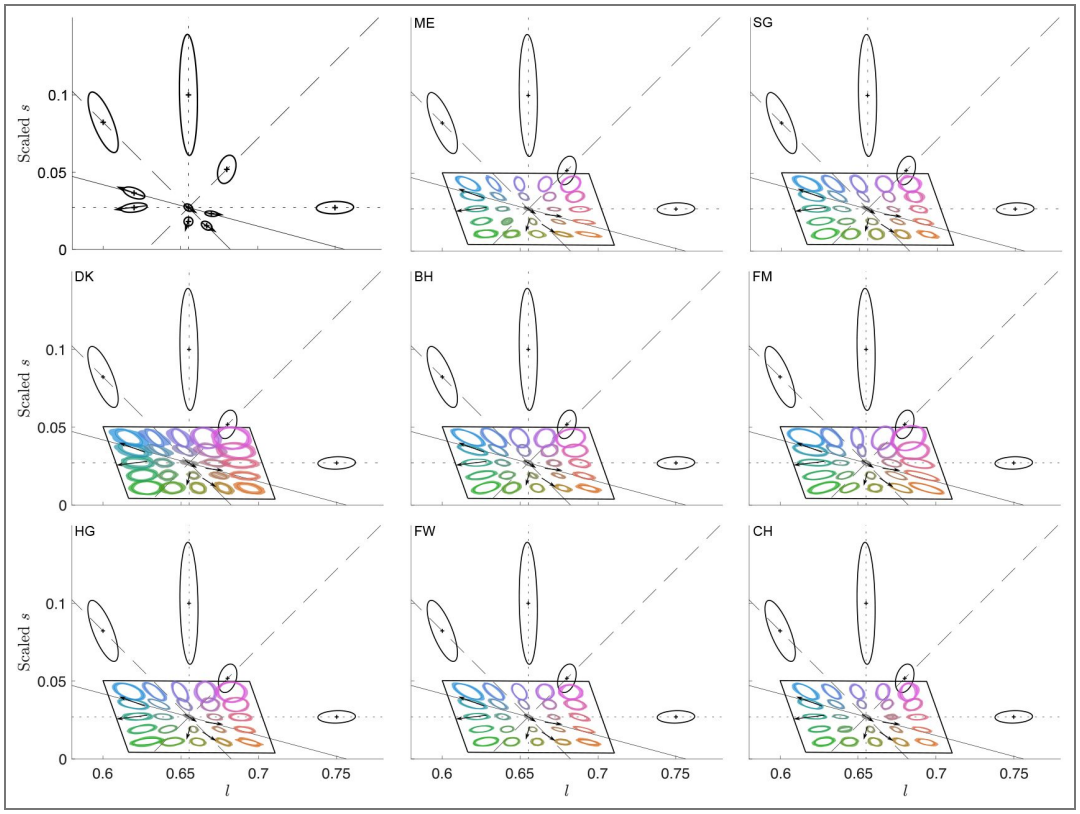

Figure S21. Comparison with Danilova and Mollon 2025 in the scaled MacLeod–Boynton space. Top left: threshold contours from their study (black ellipses), enlarged by 4×. Remaining panels: threshold contours from all participants (colored ellipses). We sampled a grid of reference points evenly spaced from −0.7 to 0.7 (5 steps) in our model space, read out the corresponding threshold contours, and transformed them into the same scaled MacLeod–Boynton space. The parallelogram indicates the gamut of the isoluminant plane. To reduce visual clutter, ellipses from Danilova & Mollon that fall within our gamut are represented by red arrows indicating only their major axes. For visual comparability, our ellipses are enlarged by 1.5× to roughly match the size of those in their study.

Appendix 8 Comparison with Krauskopf & Gegenfurtner (1992)

We compared our threshold estimates with those reported by Krauskopf and Gegenfurtner 1992. To do so, we transformed our estimates into the color space used in their measurements through a series of steps. We first read out, for each participant, the threshold contour at the achromatic reference color in the model space, which was then transformed to the DKL space (Derrington et

al., 1984). We then normalized the DKL cardinal axes so that the threshold contour at the achromatic reference had unit length along both axes. This normalized space—referred to here as the stretched DKL space—is the coordinate system in which Krauskopf and Gegenfurtner 1992 conducted their measurements. Finally, we converted the 16 reference stimuli used in their study into our model space, read out the corresponding threshold ellipses, and transformed them into this stretched DKL space to enable direct comparison (Figure S22). We observed generally good agreement between their measurements and those of some participants (e.g., CH and FM). Notably, however, individual differences are evident, particularly in the upper-right quadrant of the stretched DKL space (Figure S23). In addition, at the adapting chromaticity our ellipses are consistently rotated relative to the DKL axes, as noted in the main text.

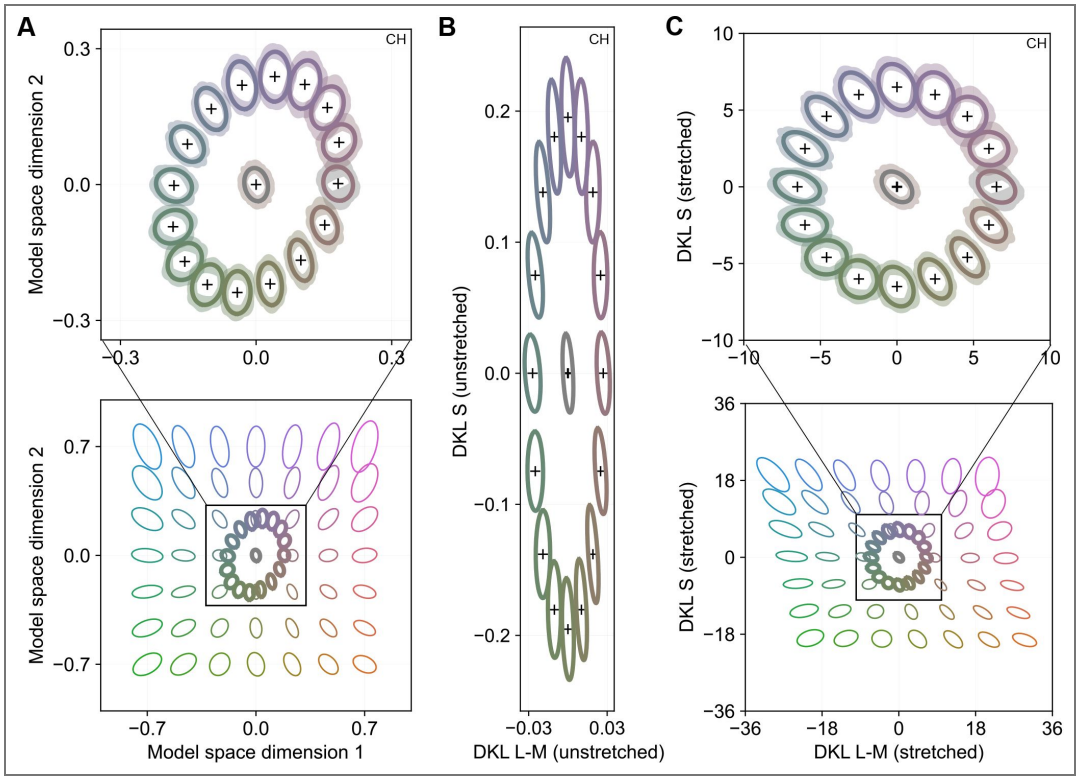

Figure S22. Transformation from the model space to a stretched DKL space used in Krauskopf and Gegenfurtner 1992 for participant CH. (A) Model space. Threshold contours were read out in this space based on each participant's WPPM fit. Notably, our data were collected on a much larger region of the isoluminant plane than they characterized. (B) The intermediate, unstretched DKL space. Transformations between this space and both the model space and the stretched DKL space are affine. (C) Stretched DKL space, in which the cardinal axes of the original DKL space are rescaled such that the threshold at the achromatic reference point is normalized.

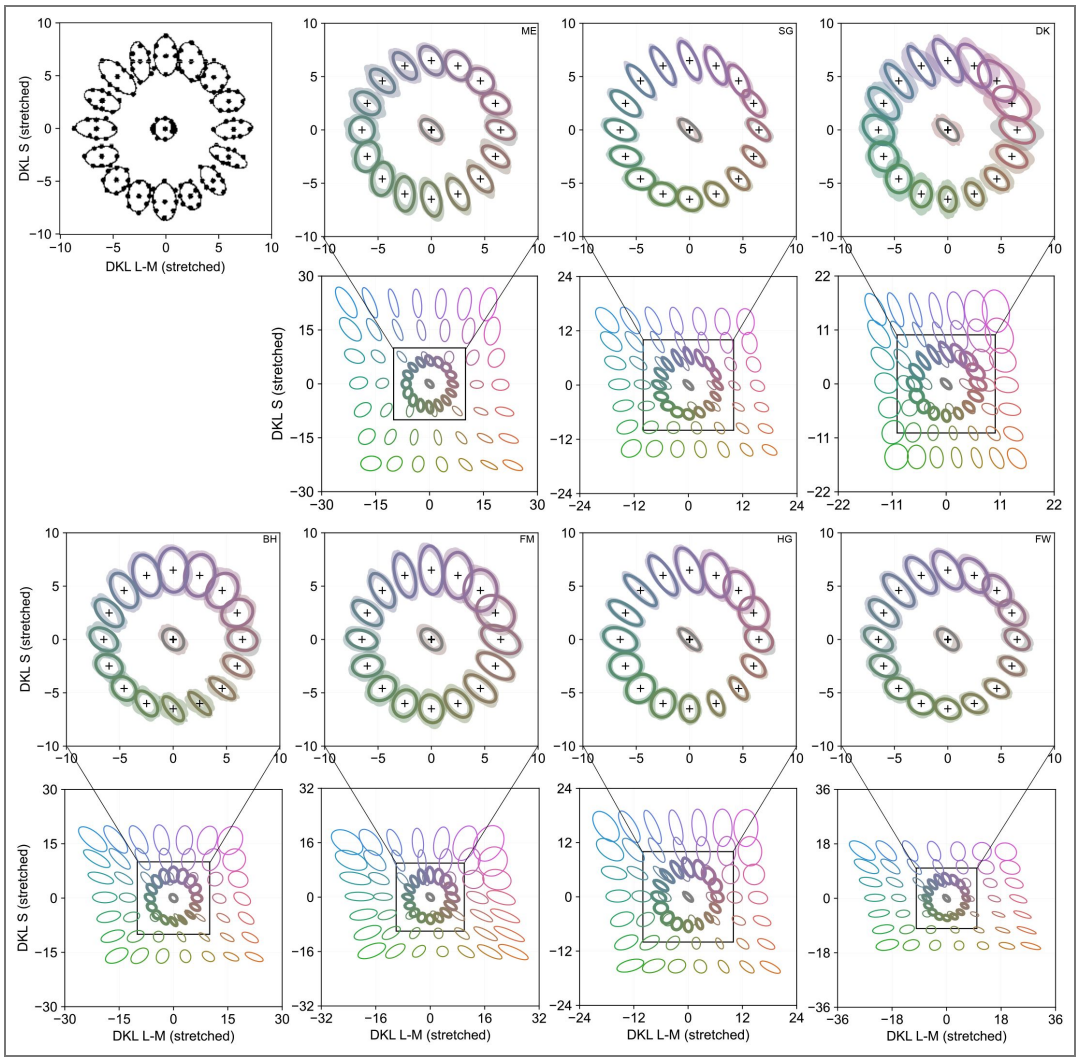

Figure S23. Comparison with Krauskopf and Gegenfurtner 1992 for the remaining seven participants. Top left: original threshold contours reported by Krauskopf and Gegenfurtner 1992, reproduced under Creative Commons CC BY-NC-ND 4.0). Remaining panels: 66.7% threshold contours (colored lines) for the remaining participants, transformed into the stretched DKL space using participant-specific scaling of the cardinal axes. Colored shaded regions indicate 95% confidence intervals computed from 120 bootstrapped datasets. All contours are plotted at their original sizes.

Appendix 9 Comparison with CIELab Δ*E*76, Δ*E*94, Δ*E*00

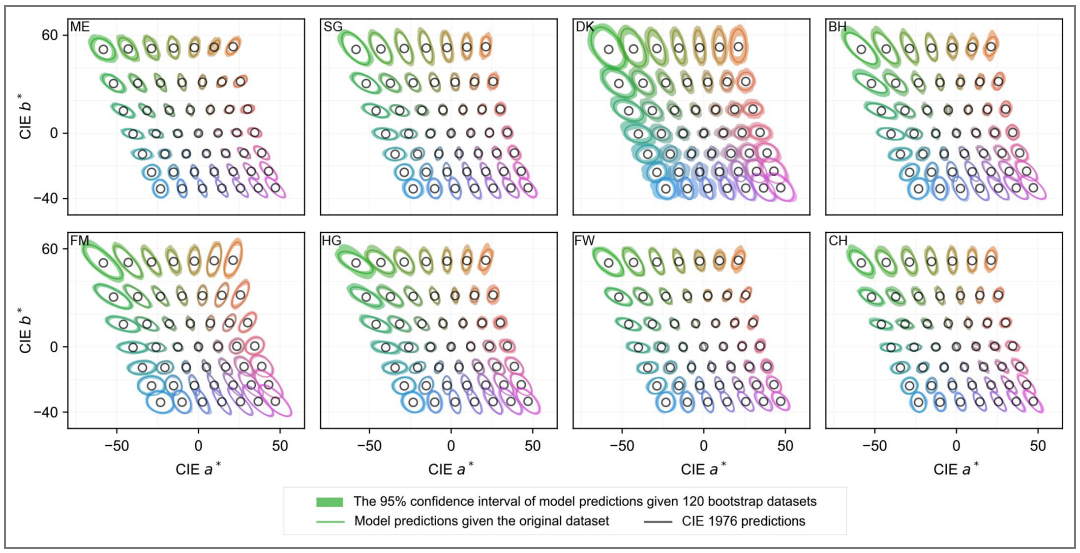

Figure S24. Comparison with CIELab Δ*E*76 color-difference (Robertson et al., 1977). Δ*E* values were converted to percent correct using a Weibull psychometric function, and threshold was defined as Δ*E* = 2.5, chosen to approximately match the scale of the measured thresholds in our data. Black contours represent the CIE predictions, whereas colored contours represent the measured thresholds transformed from the model space into CIELab space. Colored shaded regions indicate 95% confidence intervals computed from 120 bootstrapped datasets. Measured thresholds are shown at their original scales.

We obtained the elliptical thresholds on a grid of reference stimuli in the model space from the WPPM fits to participants' data (Figure S1). We then transformed these ellipses from the model space to CIELab space. Specifically, values in the model space were first converted to RGB values via the affine transformation (Appendix 1.3). The resulting RGB values were converted to LMS via a transformation matrix (Table S6), computed using the 2° cone fundamentals from CIE 2006, then to XYZ using the color matching functions reported in CIE 2015. Finally, the XYZ values were converted to CIELab using a Python implementation (Taylor, 2017). The adapting background ($R = G = B = 0.5$) was used as the reference white in the XYZ-to-Lab transformation. In addition, we computed threshold contours directly in CIELab space, defining them as iso-distance contours at a fixed perceptual distance of $\Delta E = 2.5$.

$$\mathbf{M}_{\mathrm{LMS}\rightarrow\mathrm{XYZ}} = \begin{bmatrix} 1.947 & -1.415 & 0.365 \\ 0.690 & 0.248 & 0.000 \\ 0.000 & 0.000 & 1.935 \end{bmatrix}$$

Table S6. Transformation matrix from LMS to XYZ.

Comparisons revealed that the iso-distance contours from Δ*E*94 and Δ*E*00 provided reasonable approximations to our model-predicted thresholds (Figure S25 – Figure S26), with only modest deviations. In contrast, the Δ*E*76 contours—despite their continued widespread use—diverged substantially from our measured thresholds (Figure S24).

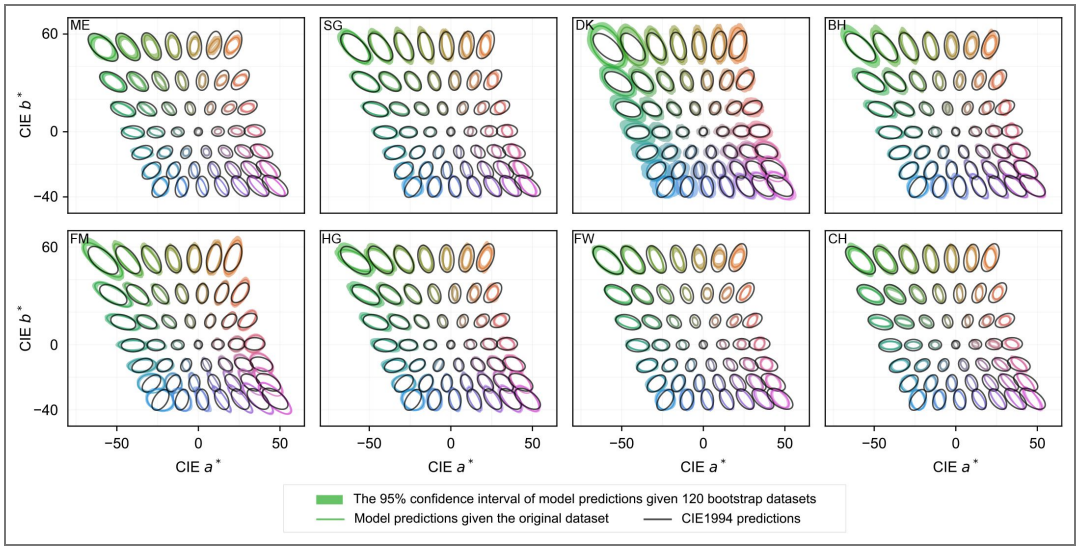

Figure S25. Comparison with predictions based on the CIELab Δ*E*94 color-difference metric (McDonald and Smith, 1995).

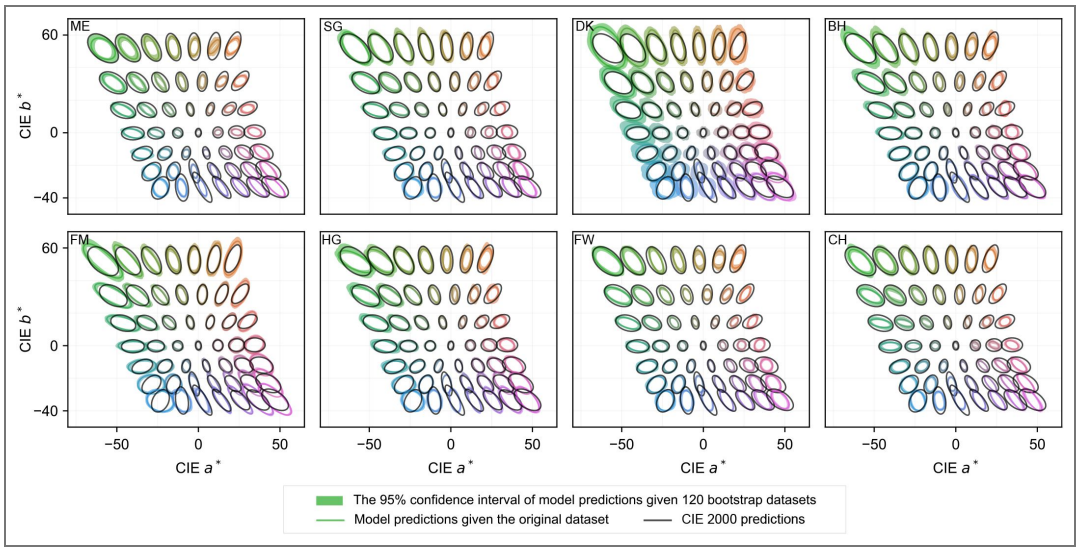

Figure S26. **Comparison with CIELab ΔE00 color-difference metric (Sharma et al., 2005).**

Appendix 10 Hyperparameters that control the smoothness prior

Appendix 10.1: The effects of ϵ and γ on WPPM predicted psychometric field

We assume that the internal noise limiting color discrimination varies smoothly across model space. Smoothness is implemented as the variance of the basis weights (Equation 16), where ϵ determines how rapidly the variance decreases with increasing polynomial order, and γ sets the overall variance scale (Figure S27).

To evaluate the influence of each hyperparameter, we first varied ϵ across a broad range while fixing $\gamma = 0.0003$. Each partipant's dataset was divided into five folds for cross-validation. For each value of ϵ, we fit the WPPM to a training set consisting of four folds of the data, while leaving the held-out fold as the test set. Model parameters were obtained by minimizing the negative log posterior probability on the training data. To evaluate model performance on both the training and test sets, we computed the negative log likelihood (nLL) without the prior to enable a fair comparison.

After each fold had been treated once as the test set, we computed the mean and the full range of nLL across the five repetitions. As expected, the mean nLL evaluated on the training data decreased monotonically with increasing ϵ, reflecting improved fit with greater model flexibility. In contrast, the mean nLL evaluated on the held-out test data initially decreased and then gradually increased (Figure S28–Figure S29). This pattern reflects a tradeoff governed by the smoothness prior. When smoothness is enforced too strongly, the model produces an overly uniform threshold field that fails to capture the data. As the smoothness constraint is relaxed, predictive performance improves and remains relatively stable over a range of ϵ values. Beyond this range, further reductions in smoothness lead to increased variability in the estimated psychometric field across the five cross-validation repetitions.

We performed the same analysis to examine the influence of the hyperparameter γ on the estimated covariance field and observed the same qualitative pattern (Figure S30–Figure S31). Importantly, the hyperparameter values used in the main analyses ($\epsilon = 0.4$ and $\gamma = 0.0003$) lie within the regime that balances oversmoothing against increasing variability.

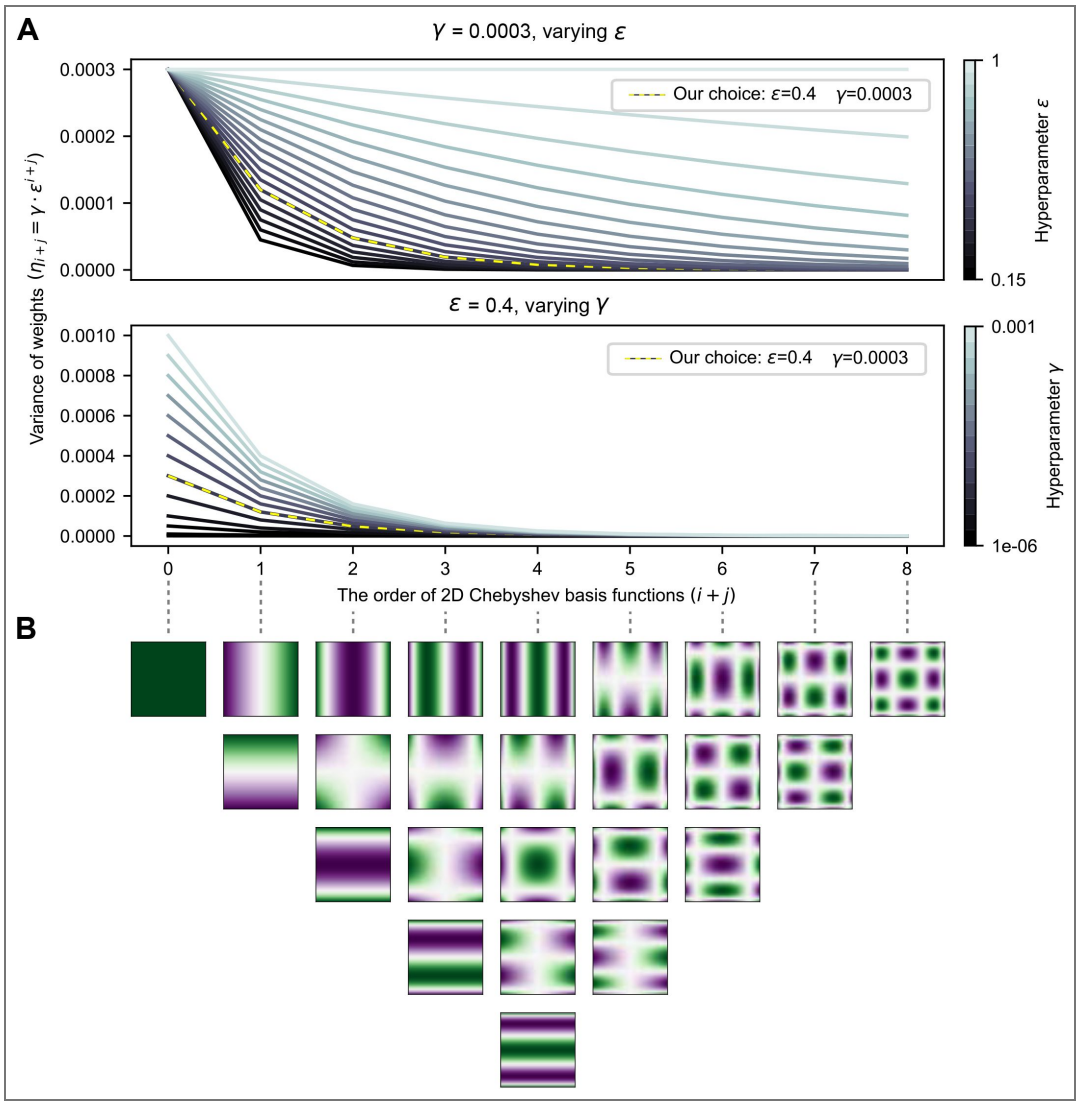

Figure S27. The effects of ϵ and γ on the variance of model weights. (A) Variance of the Chebyshev basis weights as a function of polynomial order ($i + j$). The top panel illustrates the effect of varying ϵ while holding γ fixed, whereas the bottom panel shows the effect of varying γ while holding ϵ fixed. The yellow dashed curve indicates the hyperparameter values used in the main analyses ($\epsilon = 0.4$, $\gamma = 0.0003$). (B) Two-dimensional Chebyshev basis functions arranged in order of increasing total polynomial degree ($i + j$).

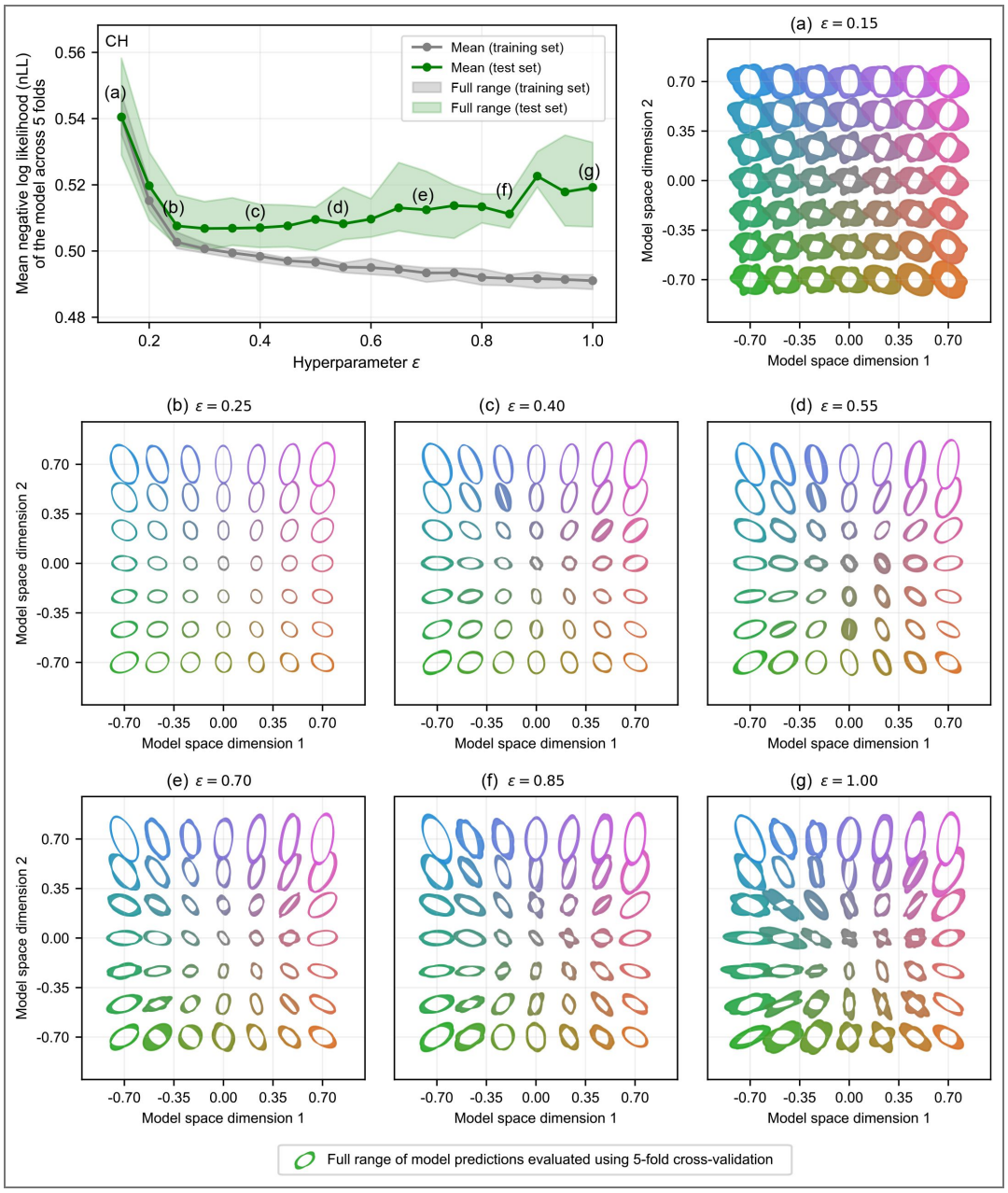

Figure S28. Effects of ϵ on WPPM-predicted psychometric field for a representative participant. Gray line and shaded region indicate the mean and full range of negative log likelihood (nLL) on the training set across five repetitions of five-fold cross-validation. Green line and shaded region indicate the mean and full range of nLL on the test set. Panels (a)–(g) show the model-predicted thresholds on a 7 × 7 reference grid for selected values of ϵ.

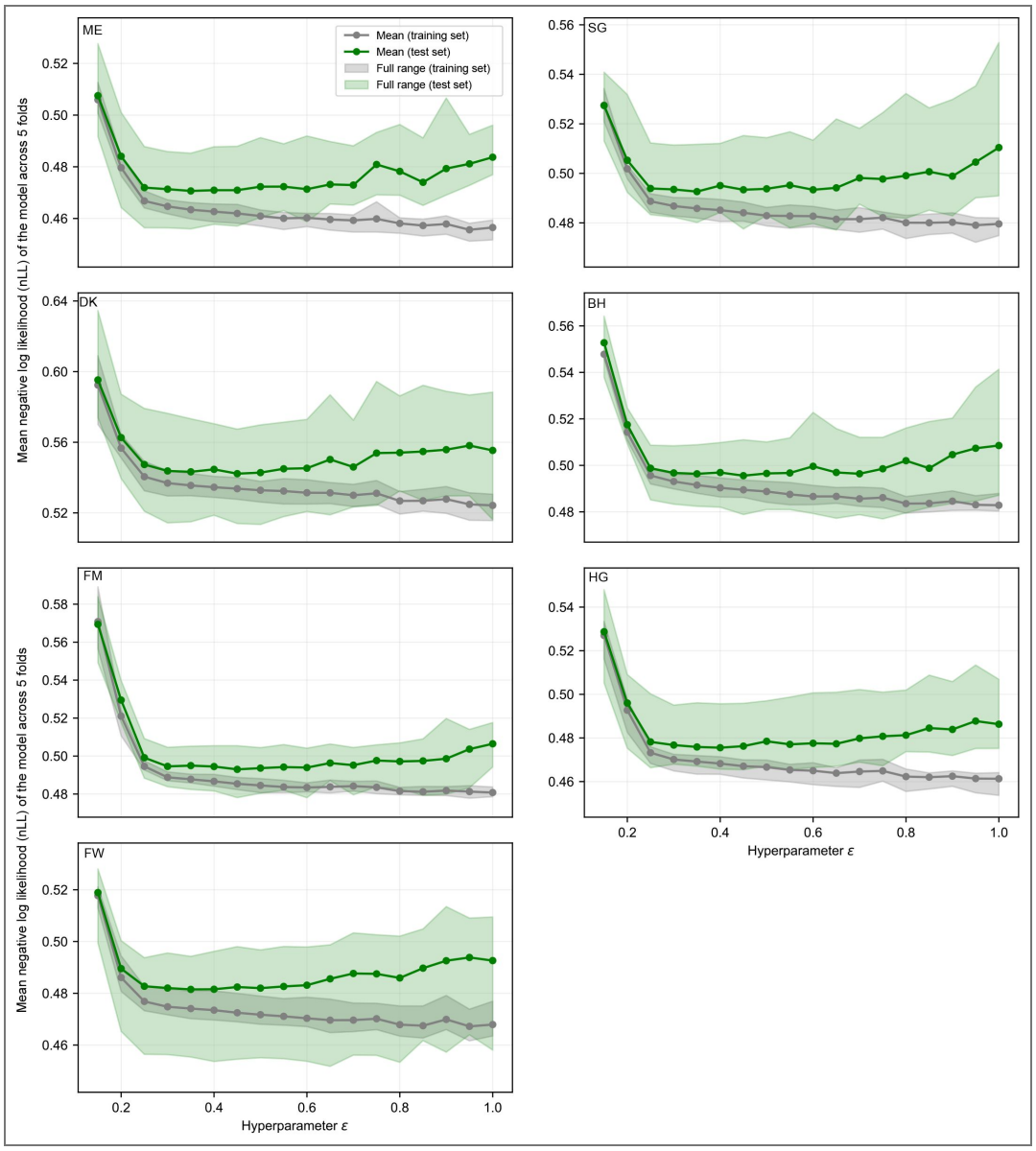

Figure S29. **Effects of ϵ on model performance for the remaining seven participants.**

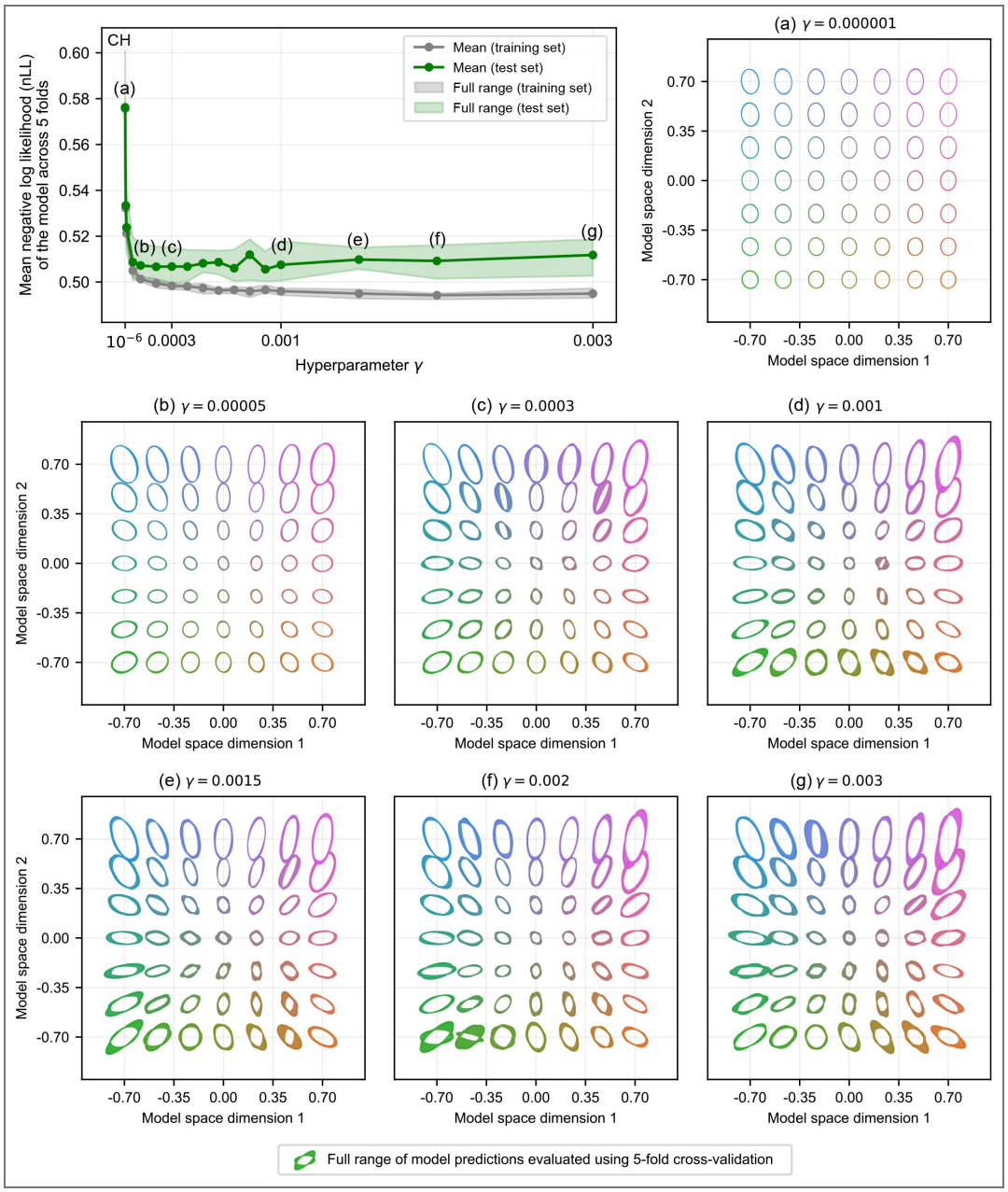

Figure S30. Effects of *γ* on WPPM-predicted psychometric field for a representative participant. Gray line and shaded region indicate the mean and full range of negative log likelihood (nLL) on the training set across five repetitions of five-fold cross-validation. Green line and shaded region indicate the mean and full range of nLL on the test set. Panels (a)–(g) show the model-predicted thresholds on a 7 × 7 reference grid for selected values of *γ*.

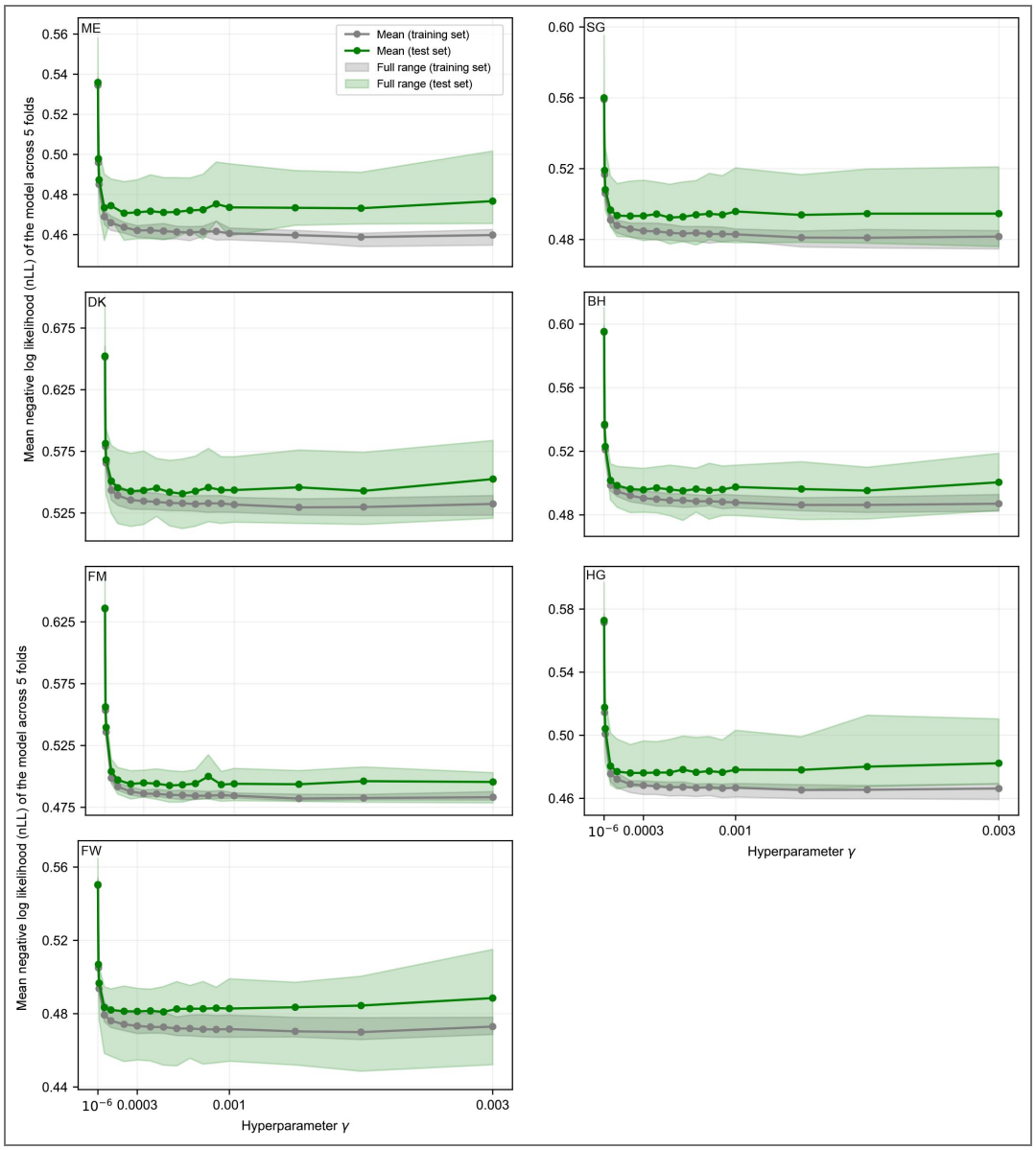

Figure S31. Effects of γ on model performance for the remaining seven participants.

Appendix 10.2: The effects of ϵ on the WPPM-validation threshold residuals

We examined how ϵ influences the agreement between WPPM-predicted thresholds and validation thresholds while fixing the variance scale γ at 0.0003. We held γ constant because cross-validation results showed that beyond a certain value the nLL for both training and test sets plateaued, indicating that further changes in γ had only a minimal impact on the model predictions (Figure S31). Using the model fits reported in Appendix 10.1 for each value of ϵ, we computed predicted thresh-olds at the 25 validation conditions and quantified the residuals between WPPM predictions and validation thresholds, following the procedure described in Appendix 4.2.

To assess systematic patterns in these residuals, we fit a linear regression model with three predictors: (1) the absolute angular difference between the chromatic direction of the validation condition and the major axis of the contours derived from the WPPM fit, (2) the aspect ratio of the contours, and (3) the magnitude of the validation threshold. The regression slopes are summarized in Table S7. Only the regression slopes are reported, as the intercepts were not as relevant.

Of the three predictors, the regression slopes relating threshold residuals to angular difference and to aspect ratio did not vary systematically with ϵ, showing no consistent or monotonic trend across different ϵ. In contrast, the regression slope relating threshold residuals to validation threshold exhibited a clear and systematic dependence on ϵ. As the smoothness prior became weaker (larger ϵ), the negative correlation between threshold residuals and validation threshold weakened. This pattern is consistent with the expected role of ϵ in regulating the strength of regularization in the psychometric field.

Predictor	ϵ	Coef	Std Err	t	p	[0.025, 0.975] CI	R^2
Absolute difference of angles	0.1	-0.001	0.000	-6.926	<0.001	[-0.001, -0.000]	0.195
	0.2	-0.000	0.000	-3.212	0.002	[-0.000, -0.000]	0.050
	0.3	0.000	0.000	1.764	0.079	[0.000, 0.000]	0.015
	0.4	0.000	0.000	0.273	0.785	[-0.000, 0.000]	0.000
	0.5	-0.000	0.000	-0.654	0.514	[-0.000, 0.000]	0.002
	0.6	-0.000	0.000	-1.393	0.165	[-0.000, 0.000]	0.010
	0.7	-0.000	0.000	-1.592	0.113	[-0.000, 0.000]	0.013
	0.8	-0.000	0.000	-1.901	0.059	[-0.000, 0.000]	0.018
	0.9	-0.000	0.000	-3.732	<0.001	[-0.000, -0.000]	0.066
	1.0	-0.000	0.000	-1.961	0.051	[-0.000, 0.000]	0.019
Aspect ratio	0.1	-0.003	0.001	-2.542	0.012	[-0.006, -0.001]	0.032
	0.2	0.001	0.005	0.282	0.778	[-0.008, 0.011]	0.000
	0.3	0.003	0.002	1.068	0.287	[-0.002, 0.007]	0.006
	0.4	0.002	0.002	0.943	0.347	[-0.002, 0.005]	0.004
	0.5	0.001	0.002	0.289	0.773	[-0.003, 0.004]	0.000
	0.6	-0.001	0.002	-0.724	0.470	[-0.004, 0.002]	0.003
	0.7	-0.004	0.001	-2.625	0.009	[-0.006, -0.001]	0.034
	0.8	-0.005	0.001	-4.596	<0.001	[-0.008, -0.003]	0.096
	0.9	-0.002	0.001	-3.266	0.001	[-0.003, -0.001]	0.051
	1.0	-0.002	0.001	-3.141	0.002	[-0.003, -0.001]	0.047
Validation thresholds	0.1	-0.964	0.061	-15.767	<0.001	[-1.085, -0.844]	0.557
	0.2	-0.417	0.049	-8.580	<0.001	[-0.513, -0.321]	0.271
	0.3	-0.266	0.032	-8.245	<0.001	[-0.329, -0.202]	0.256
	0.4	-0.176	0.031	-5.727	<0.001	[-0.237, -0.116]	0.142
	0.5	-0.144	0.034	-4.207	<0.001	[-0.211, -0.076]	0.082
	0.6	-0.134	0.035	-3.841	<0.001	[-0.203, -0.065]	0.069
	0.7	-0.157	0.039	-4.020	<0.001	[-0.233, -0.080]	0.075
	0.8	-0.104	0.043	-2.436	0.016	[-0.188, -0.020]	0.029
	0.9	-0.106	0.050	-2.124	0.035	[-0.205, -0.008]	0.022
	1.0	-0.103	0.044	-2.345	0.020	[-0.190, -0.016]	0.027

Table S7. The slope of linear regression assessing the relationship between WPPM–validation threshold residuals and three predictors reported in Table S3. The hyperparameter γ was fixed at 0.0003, while ϵ was varied.

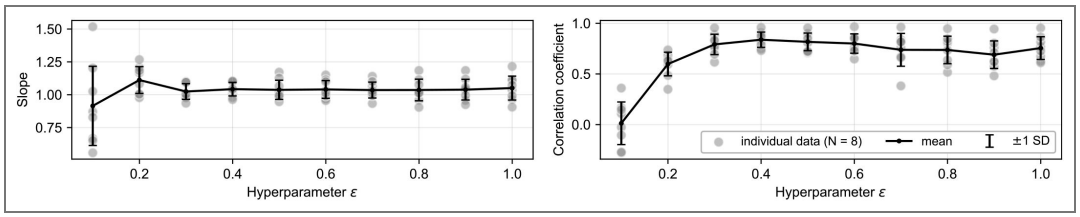

Figure S32. Effects of ϵ on the slope and correlation coefficient of the linear regression between WPPM-predicted thresholds and validation thresholds.

Appendix 10.3: The effects of ϵ on linear regression between WPPM and validation thresholds

We also examined how ϵ affects the linear regression between WPPM and validation thresholds computed at the 25 validation conditions, as illustrated in Figure S6C–Figure S13C. Across participants (N = 8), the mean regression slope did not deviate systematically from zero. However, the standard deviation of the slope across participants was large when ϵ was small (strong smoothness prior), indicating substantial bias at the individual level. The standard deviation of the slope across participants reached a minimum at $\epsilon = 0.4$ and increased again for larger values of ϵ. Additionally, the correlation coefficient between WPPM and validation thresholds was near zero under strong smoothness, peaked at $\epsilon = 0.4$, and declined again as ϵ increased further (Figure S32).

Together, these results reflect a bias–variance tradeoff. Excessive smoothness introduces bias by failing to capture structure in the data, whereas insufficient smoothness increases variance in model predictions through overfitting. These results further support our choice of $\epsilon = 0.4$ as lying near the optimal balance between bias and variance.

Appendix 11 Display characterization

Appendix 11.1: Calibration of monitor output

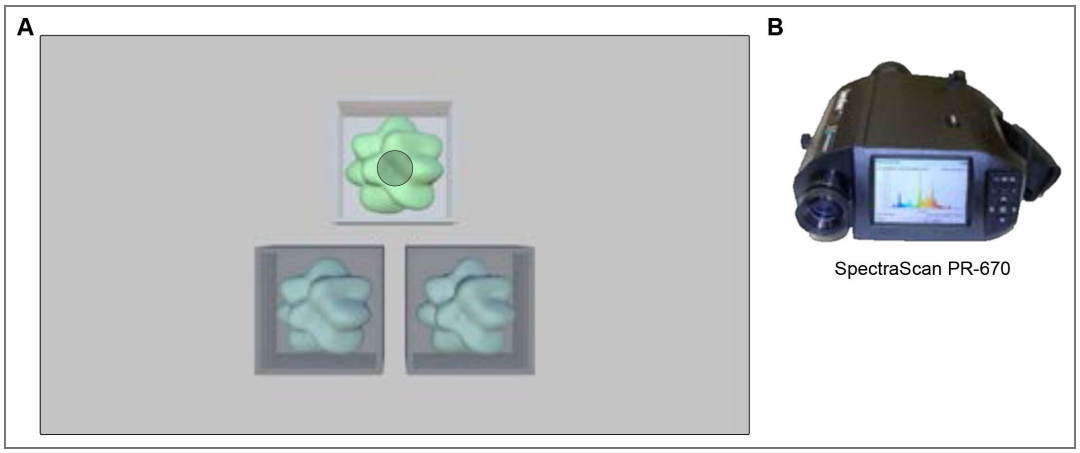

Figure S33. Stimuli and equipment used for calibration. (A) The stimulus setup during calibration was identical to that used in the main experiment. The surface color of both the cubic room and the blobby stimulus (shown here as the top-position stimulus) was varied during the calibration procedure. The shaded gray circular region on the stimulus indicates the area measured by the spectroradiometer's lens. (B) A SpectraScan PR-670 used for all calibration measurements.

Calibration was carried out with three blobby objects arranged in a triangular configuration inside the cubic room (Figure S33A). A SpectraScan PR-670 radiometer (Figure S33B), positioned at the same viewing distance as the chin-rest, recorded all measurements (Brainard et al., 2002).

We first obtained each primary's gamma function by measuring the screen output at 61 evenly spaced input levels (from 0 to 1) rendered through Unity (v2022.3.24f1) (Figure S34A). The resulting curves lie above the identity line because Unity internally applies its own assumed gamma exponent when texture values are altered. We also measured the spectral power distributions (SPDs) of the red, green, and blue primaries given different intensity levels (Figure S34B), and examined the stability of the primaries' chromaticity in the CIE diagram (Figure S34C). There was almost no drift in the chromaticities, indicating the monitor's color output remained stable across intensity levels (Figure S34D). To evaluate linearity and repeatability, we compared nominal (predicted) versus measured luminance and chromaticity for two independent

measurement runs (Figure S34E). Deviations from linearity were minimal and nearly identical across repeats, confirming reliable reproduction (Figure S34F). Finally, we tested whether the cubic room's background color affected the stimulus SPD; no measurable influence was detected (Figure S34G).

To assess consistency across different locations on the screen, we conducted the same set of measurements on each of the three blobby stimuli, and compared their primaries and chromaticities. The results showed consistent color behavior of the monitor (Figure S35), and thus we applied a single gamma correction curve to all three stimuli. This correction was derived from measurements of the bottom-right blobby stimulus. Specifically, we interpolated a gamma table for 4,096 RGB input values using a combination of linear and polynomial fits, from which we derived an inverse gamma function (Figure S36A). To validate this correction, we repeated the measurements with the gamma correction applied in Unity. The measured output closely aligned with the identity line across all three primaries, indicating accurate correction (Figure S36B).

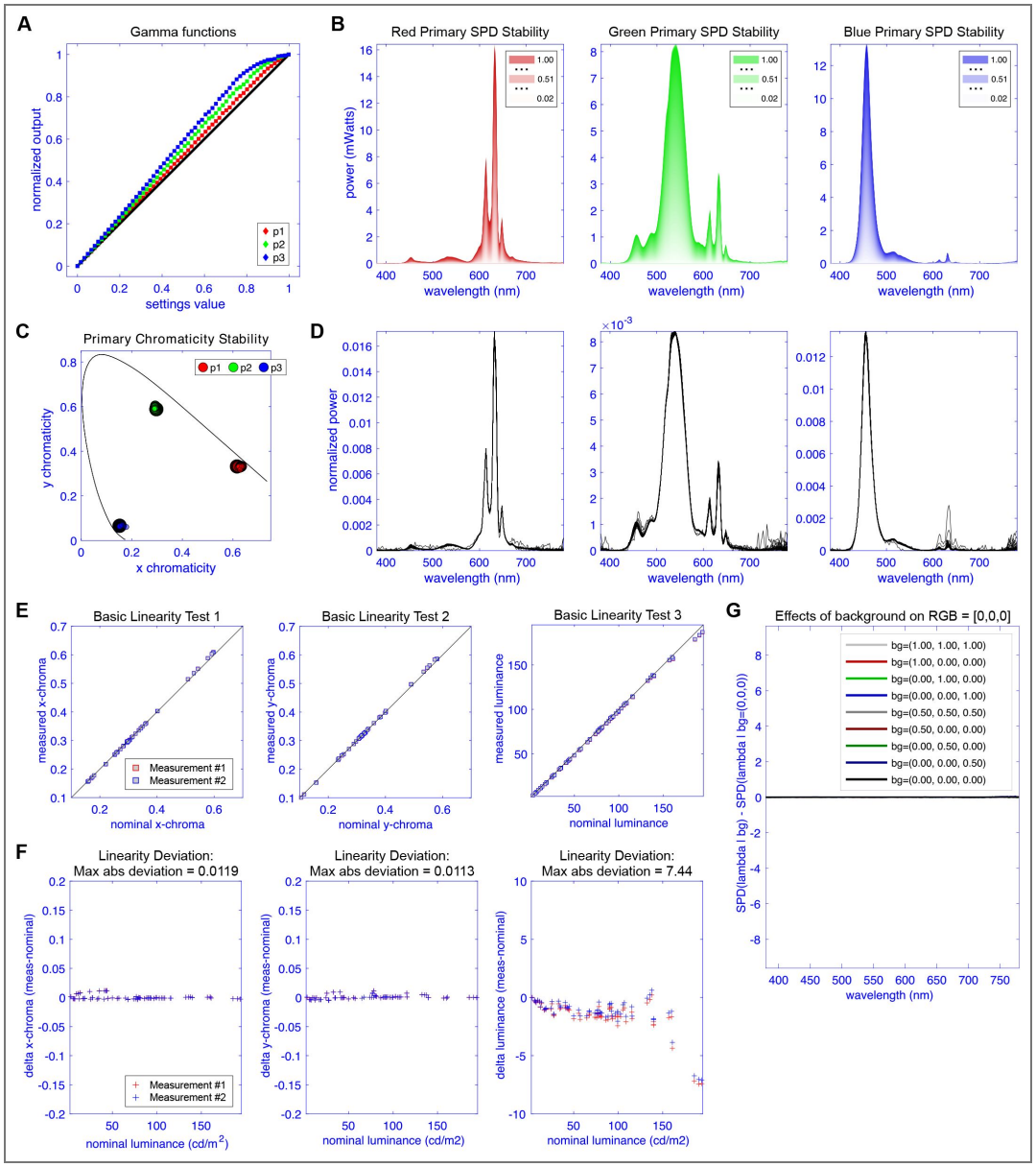

Figure S34. Characterization of display output via Unity's rendering pipeline. (A) Gamma functions for red, green and blue primaries. Note that Unity's internal correction places them above the identity line. (B) Spectral power distributions (SPDs) of the three primaries across a range of intensity levels. (C) The chromaticity of each primaries in the CIE chromaticity diagram at different intensity levels. (D) Normalized SPDs for each primary,

showing that SPD shape is stable across intensity levels. (E) Linearity tests comparing predicted and measured chromaticity and luminance across two independent measurement runs. (F) Deviations from linearity. (G) Effect of the cubic room's background color on the SPD of the blobby stimulus, showing no detectable influence.

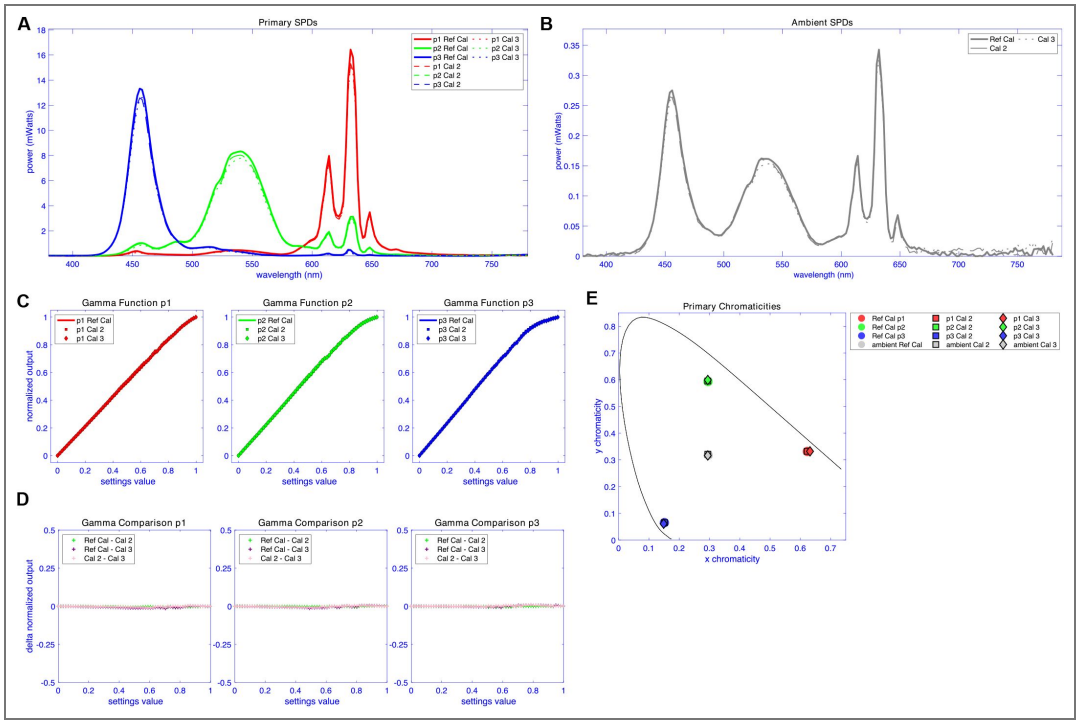

Figure S35. Comparison of display output across stimulus locations. (A) Spectral power distributions (SPDs) for each stimulus location: Ref Cal (bottom right), Cal 2 (bottom left), and Cal 3 (top). (B) Ambient light SPDs measured during calibration. (C) Gamma functions for each primary (red, green, blue) across all three stimulus locations. (D) Differences in normalized output for each pairwise comparison of stimulus locations, plotted separately for each primary. (E) Chromaticity coordinates of each primary in the CIE diagram, shown for all three stimulus locations.

Finally, to ensure the gamma correction remained stable over time, we repeated the measurements of display output via Unity's rendering pipeline on the bottom-right stimulus with gamma correction applied, approximately one month after data collection began. The results confirmed that the correction remained accurate and consistent (Figure S37).

Appendix 11.2: Assessment of color depth

Color depth measurements were conducted using a single blobby stimulus positioned at the center of the screen (Figure S38A). This stimulus was originally the top stimulus in the triangular configuration, and the camera view was adjusted to center it on the screen. Additionally, compared to the scene used in the main experiment, the other two blobby stimuli and all cubic room elements were excluded from rendering and thus were not visible. A Klein K-10A colorimeter (Figure S38A), placed directly in front of the monitor without any distance, was used to make the measurements.

Specifically, we tested RGB values ranging from 1928/4095 to 2128/4095, in increments of 1/4095. Each stimulus was displayed for 5 seconds, and the RGB values from the first frame of the frame buffer were saved in EXR format. We then compared the average RGB values across the surface of the blobby object (extracted from the EXR files) to the luminance measured through-out the full stimulus presentation. Although individual pixels exhibited quantization below 12-bit precision, the mean luminance increased with each 1/4095 increment, rather than in a staircase pattern. A similarly smooth progression was observed in the average R, G, and B channel values, with the R channel shown as an example in Figure S38B.

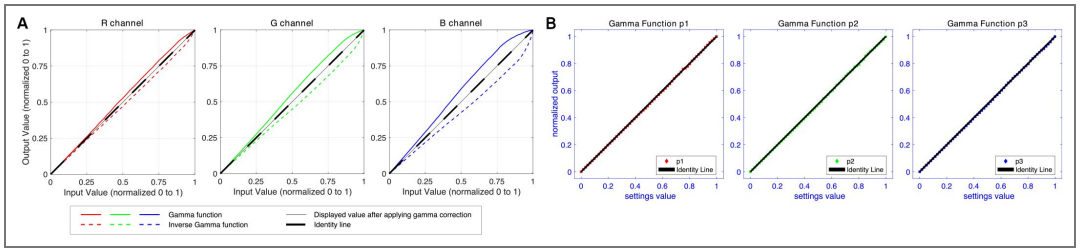

Figure S36. Gamma correction. (A) Measured gamma functions and their corresponding inverse functions for the red, green, and blue primaries, used to construct the gamma correction lookup table. (B) Gamma functions re-measured after applying the correction in Unity, showing close alignment with the identity line for all three primaries.

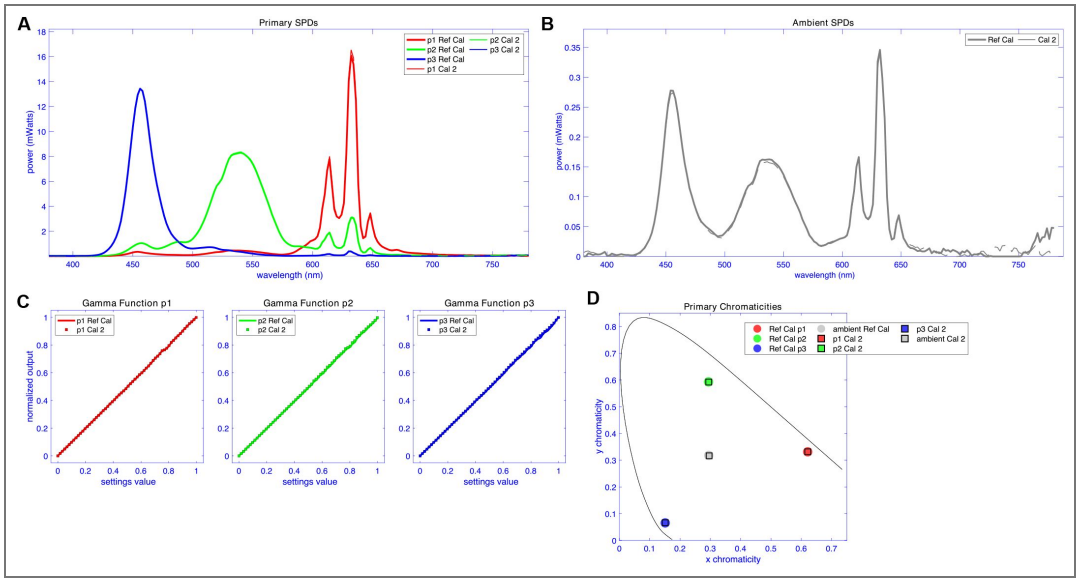

Figure S37. Comparison of display output with gamma correction over time. (A) Spectral power distributions (SPDs) when measured at the bottom-right blobby stimulus location. Ref Cal (initial measurements prior to the experiment) and Cal 2 (follow-up measurements roughly one month after data collection began). (B) Ambient light SPDs measured during each calibration. (C) Gamma functions for the red, green, and blue primaries across both sessions, with gamma correction applied. (D) Chromaticity coordinates of each primary plotted in the CIE diagram for both calibration runs.

To better understand how Unity and our video chain achieved this behavior, we analyzed horizontal slices of pixel values from the EXR files. When extracting a very thin slice—just one pixel in height—the individual pixel values exhibited staircase-like changes, consistent with 8-bit quantization. However, as we increased the height of the horizontal slice, the averaged channel values became progressively smoother. These results suggest that Unity achieves effective 12-bit color depth through internal spatial dithering (Figure S38C).

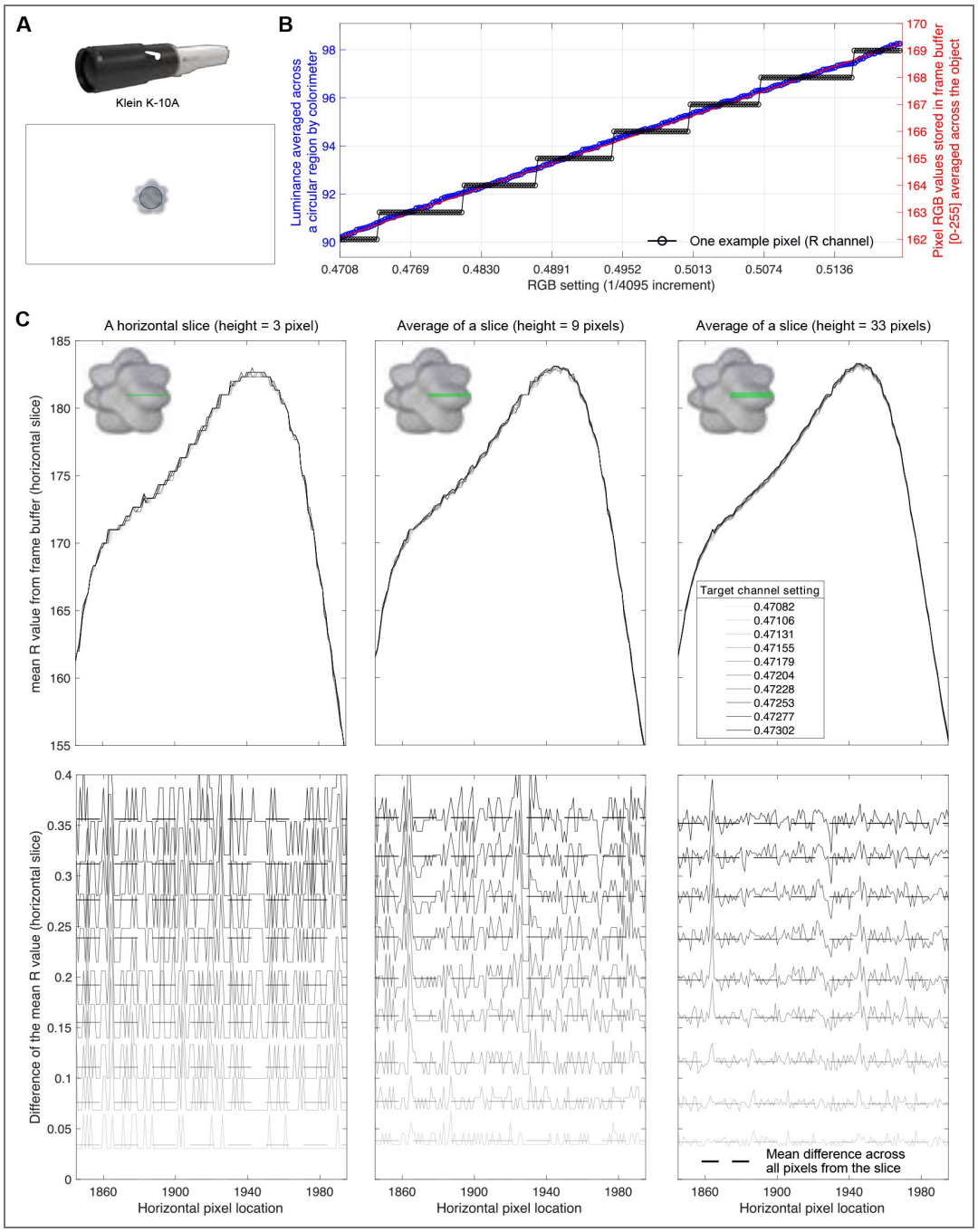

Figure S38. Evidence of spatial dithering by Unity's rendering pipeline. (A) The stimulus setup during measurements was identical to that used in the main experiment. The surface color of both the cubic room and the blobby stimulus (shown here as the top-position stimulus) was varied across trials during the measurements. The shaded gray circular region on the stimulus indicates the area measured by the colorimeter's aperture. (B) Spatial dithering by Unity's standard shader is suggested by comparing the luminance measurements from the Klein K-10A (averaged across a circular region on the blobby object) with the RGB values stored in the frame buffer. The measured luminance shows small incremental changes as the RGB settings increase in steps of 1/4095. These measurements are consistent with what we obtain by averaging over pixels in a saved image of the frame buffer (saved from Unity in .exr format). The averaged pixel values exhibit 12-bit quantization even though individual pixel values exhibit 8-bit quantization. (C) Top row: mean R channel values averaged vertically within a horizontal slice of the blobby object. Bottom row: differences in the R channel values between the minimum target R channel setting and each of the rest settings. Different shades of gray represent different target R settings. For illustration, only a portion of the horizontal slice is shown, and solid lines in the bottom row are scaled by a factor of 0.1. Dashed lines: the mean difference averaged across all pixels within each slice.

Appendix 12 Differentiable Monte Carlo approach

Recall from the main text that the log likelihood function implied by the WPPM observer model can be written in terms of

$$\Pr\left[d_M\left(z_0, z_0^{'}\right) - \min\left(d_M\left(z_0, z_1\right), d_M\left(z_0^{'}, z_1\right)\right) \leq 0 \mid \mathbf{W}\right], \tag{S14}$$

which has no simple closed form solution and must be estimated by Monte Carlo simulation. This quantity of interest takes the form of a cumulative distribution function $g(u) = \Pr[v \leq u]$ for some random variable v and scalar constant u. In this section, we describe how to approximate the log likelihood in a manner that is compatible with automatic differentiation libraries, which enables gradient-based optimization of the log posterior density.

Let P_θ denote some probability distribution parameterized by θ. Given n independent and identically distributed random variables, $v_1, \ldots, v_n \sim P_\theta$, we would like to form an estimate of the cumulative distribution function, $g(u) = \Pr[v \leq u]$ where $v \sim P_\theta$. A simple and well-known estimate is empirical cumulative distribution function:

$$\hat{g}_{\text{emp}}(u) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\left[v_i \leq u\right] \tag{S15}$$

where $\mathbf{1}[\cdot]$ is the indicator function—i.e. $\mathbf{1}[A]$ evaluates to one if the event A occurs and evaluates to zero otherwise. In many respects, this is a perfectly fine estimator. For example, the celebrated Dvoretzky–Kiefer–Wolfowitz inequality (Dvoretzky et al., 1956) states that this estimate converges exponentially fast to the true cumulative distribution function as $n \to \infty$.

In our setting, we would like to not only evaluate $g(u)$ for any given u, but to also evaluate $\partial g(u)/\partial\theta_j$ for all parameters $\theta_1, \ldots, \theta_J$ that define the underlying distribution P_θ. ***Equation S15*** provides an estimate of $g(u)$ but it is unfortunately not differentiable with respect to $v_1, \ldots, v_n$ because $\mathbf{1}[v_i \leq u]$ is a discontinuous step as a function of u. A straightforward and intuitive solution is to replace this step function with a smooth sigmoid function. We formalize this approach below, showing that it can be motivated by forming a smoothed estimate of the underlying density function.

Specifically, let $K(v)$ denote a smooth, nonnegative function that integrates to one and satisfies $K(v) = K(-v)$. Suppose that P_θ has a density function $f(v)$. Then, given $v_1, \ldots, v_n \sim P_\theta$ we can estimate the density function as:

$$\hat{f}(v) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{v - v_i}{h}\right) \tag{S16}$$

where $h > 0$ is a user-specified hyperparameter called the bandwidth. ***Equation S16*** is known as a *kernel density estimate* (Wasserman, 2006). Asymptotically, $\hat{f}$ approaches the true density function

f as $n \to \infty$ and $h \to 0$. Intuitively, larger values of h lead to smoother density estimates, which is preferable in sample-limited (i.e. small n) regimes.

Now define:

$$I_{v_i}(u) = \int_{-\infty}^{u} \frac{1}{h} K\left(\frac{v - v_i}{h}\right) \mathrm{d}v \tag{S17}$$

which is a smooth sigmoid function centered at v_i, and consider the following estimate of the cumulative distribution function:

$$\hat{g}(u) = \frac{1}{n}\sum_{i=1}^{n} I_{v_i}(u) \tag{S18}$$

Notice that in the limit of $h \to 0$, we recover the empirical cumulative distribution estimator $\hat{g} = \hat{g}_{\text{emp}}$ because, in this limit, we have that $I_{v_i}(u) = \mathbf{1}[v_i \leq u]$. We can further justify ***Equation S18*** as a reasonable estimator of g by recognizing it as the integral of the density estimate in ***Equation S16***. That is,

$$g(u) = \int_{-\infty}^{u} f(v)\mathrm{d}v \approx \int_{-\infty}^{u} \hat{f}(v)\mathrm{d}v = \int_{-\infty}^{u} \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{v - v_i}{h}\right)\mathrm{d}v \tag{S19}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{u}\frac{1}{h}K\left(\frac{v - v_i}{h}\right)\mathrm{d}v = \frac{1}{n}\sum_{i=1}^{n} I_{v_i}(u) = \hat{g}(u) \tag{S20}$$

Our refined estimator $\hat{g}$ is clearly differentiable whenever we choose $K(v)$ to be a smoothly differentiable function. In our model fitting routine, we chose $K(v)$ to be the density of a standard logistic distribution:

$$K(v) = \frac{\exp[-v]}{(1 + \exp[-v])^2} \tag{S21}$$

This smoothing kernel has heavy tails, which we reasoned would enable numerically stable autodifferentiation routines even when h is chosen to be small. Another feature is that the integrated density is the well-known *logistic function*:

$$I_{v_i}(u) = \frac{1}{1 + \exp\left[-\left(u - v_i\right)/h\right]} \tag{S22}$$

which is familiar and easy to compute.

References

Agosti G, Hadnett-Hunter J, Gegenfurtner KR (2026) Color discrimination in action: High-throughput measurement in immersive VR. *bioRxiv* https://doi.org/10.64898/2026.02.13.705758

Aguilar G, Wichmann FA, Maertens M. (2017) Comparing sensitivity estimates from MLDS and forced-choice methods in a slant-from-texture experiment. *Journal of Vision* **17**:37-37 https://doi.org/10.1167/17.1.37 | PubMed

Ashby FG, Soto FA (2015) Multidimensional signal detection theory. In: Busemeyer JR, et al (Eds). *Oxford handbook of computational and mathematical psychology* pp. 13-34 https://doi.org/10.1093/oxfordhb/9780199957996.013.2

Aspinall PA, Kinnear PR, Duncan LJ, Clarke BF (1983) Prediction of diabetic retinopathy from clinical variables and color vision data. *Diabetes Care* **6**:144-8 https://doi.org/10.2337/diacare.6.2.144 | PubMed

Barnett MA, Chin BM, Aguirre GK, Burge J, Brainard DH (2025) Temporal dynamics of human color processing measured using a continuous tracking task. *Journal of Vision* **25**:12-12 https://doi.org/10.1167/jov.25.2.12 | PubMed

Bertonati G, Amadeo MB, Campus C, Gori M. (2021) Auditory speed processing in sighted and blind individuals. *Plos one* **16**:e0257676 https://doi.org/10.1371/journal.pone.0257676 | PubMed

Bhatia R, Jain T, Lim Y. (2019) On the Bures–Wasserstein distance between positive definite matrices. *Expositiones Mathematicae* **37**:165-191 https://doi.org/10.1016/j.exmath.2018.01.002

Bosten JM (2022) Do you see what I see? Diversity in human color perception. *Annual review of vision science* **8**:101-133 https://doi.org/10.1146/annurev-vision-093020-112820 | PubMed

Bradbury J, Frostig R, Hawkins P, Johnson MJ, Leary C, Maclaurin D, Necula G, Paszke A, VanderPlas J, Wanderman-Milne S, *et al.* (2018) JAX: composable transformations of Python+NumPy programs.

Brainard DH (2003) Color Appearance and Color Difference Specification. In: Shevell SK (Ed). *The Science of Color* (2nd) Elsevier. pp. 191-216 https://doi.org/10.1016/b978-044451251-2/50006-4

Brainard DH, Pelli DG, Robson T. (2002) Display characterization. *Signal Process* **80**:2-067

Brainard DH, Roorda A, Yamauchi Y, Calderone JB, Metha A, Neitz M, Neitz J, Williams DR, Jacobs GH (2000) Functional consequences of the relative numbers of L and M cones. *Journal of the Optical Society of America A* **17**:607-614 https://doi.org/10.1364/josaa.17.000607 | PubMed

Brainard DH, Stockman A. (2010) Colorimetry. In: *Handbook of Optics Volume III: Vision and Vision Optics* McGraw Hill.

Brainard D. (1996) Cone contrast and opponent modulation color spaces. In: *Human color vision*

Brown WRJ, MacAdam DL (1949) Visual sensitivities to combined chromaticity and luminance differences. *Journal of the Optical Society of America* **39**:808-834 https://doi.org/10.1364/josa.39.000808 | PubMed

Brown W. (1952) The effect of field size and chromatic surroundings on color discrimination. *Journal of the Optical Society of America* **42**:837-844 https://doi.org/10.1364/josa.42.000837 | PubMed

Bujack R, Teti E, Miller J, Caffrey E, Turton TL (2022) The non-Riemannian nature of perceptual color space. *Proceedings of the National Academy of Sciences* **119**:e2119753119 https://doi.org/10.1073/pnas.2119753119 | PubMed

Burge J, Cormack LK (2024) Continuous psychophysics shows millisecond-scale visual processing delays are faithfully preserved in movement dynamics. *Journal of Vision* **24**:4-4 https://doi.org/10.1167/jov.24.5.4 | PubMed

Campbell FW, Robson JG (1968) Application of Fourier analysis to the visibility of gratings. *The Journal of physiology* **197**:551 https://doi.org/10.1113/jphysiol.1968.sp008574 | PubMed

Carlile S, Leung J. (2016) The perception of auditory motion. *Trends in hearing* **20**:2331216516644254 https://doi.org/10.1177/2331216516644254 | PubMed

Carroll J, Neitz J, Neitz M. (2002) Estimates of L: M cone ratio from ERG flicker photometry and genetics. *Journal of vision* **2**:1-1 https://doi.org/10.1167/2.8.1 | PubMed

Champion RA, Freeman TC (2010) Discrimination contours for the perception of head-centered velocity. *Journal of Vision* **10**:14-14 https://doi.org/10.1167/10.6.14 | PubMed

Chebyshev PL (1853) *Théorie des mécanismes connus sous le nom de parallélogrammes* Imprimerie de l'Académie impériale des sciences.

Chen CC, Foley JM, Brainard DH (2000) Detection of chromoluminance patterns on chromoluminance pedestals II: model. *Vision Research* **40**:789-803 https://doi.org/10.1016/s0042-6989(99)00228-x | PubMed

Churchland PM (1986) Some reductive strategies in cognitive neurobiology. *Mind* **95**:279-309

Cicchini GM, Anobile G, Burr DC (2016) Spontaneous perception of numerosity in humans. *Nature communications* **7**:12536 https://doi.org/10.1038/ncomms12536 | PubMed

Cicchini GM, Anobile G, Burr DC (2019) Spontaneous representation of numerosity in typical and dyscalculic development. *Cortex* **114**:151-163 https://doi.org/10.1016/j.cortex.2018.11.019 | PubMed

Cicchini GM, Anobile G, Burr DC, Marchesini P, Arrighi R. (2023) The role of non-numerical information in the perception of temporal numerosity. *Frontiers in Psychology* **14**:1197064 https://doi.org/10.3389/fpsyg.2023.1197064 | PubMed

CIE (2004) *Colorimetry*

CIE (2006) *Fundamental Chromaticity Diagram with Physiological Axes – Part 1* Vienna, Austria: CIE Central Bureau.

CIE (2015) *Fundamental Chromaticity Diagram with Physiological Axes – Part 2: Spectral Luminous Efficiency Functions and Chromaticity Diagrams* Vienna, Austria: CIE Central Bureau.

Craik K. (1938) The effect of adaptation on differential brightness discrimination. *The Journal of Physiology* **92**:406 https://doi.org/10.1113/jphysiol.1938.sp003612 | PubMed

Crozier WJ, Holway AH (1937) On the law for minimal discrimination of intensities: I. *Proceedings of the National Academy of Sciences* **23**:23-28 https://doi.org/10.1073/pnas.23.1.23 | PubMed

Danilova M, Mollon J. (2025) Effect of stimulus size on chromatic discrimination. *Journal of the Optical Society of America A* **42**:B167-B177 https://doi.org/10.1364/josaa.545292 | PubMed

Derrington AM, Krauskopf J, Lennie P. (1984) Chromatic mechanisms in lateral geniculate nucleus of macaque. *The Journal of physiology* **357**:241-265 https://doi.org/10.1113/jphysiol.1984.sp015499 | PubMed

Dvoretzky A, Kiefer J, Wolfowitz J. (1956) Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics* 642-669 https://doi.org/10.1214/aoms/1177728174

Efron B. (2012) *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction* **1** Cambridge University Press.

Ennis DM, Mullen K. (2014) 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

Ennis RJ, Zaidi Q. (2019) Geometrical structure of perceptual color space: Mental representations and adaptation invariance. *Journal of vision* **19**:1-1 https://doi.org/10.1167/19.12.1 | PubMed

Eskew RT (2009) Higher order color mechanisms: A critical review. *Vision research* **49**:2686-2704 https://doi.org/10.1016/j.visres.2009.07.005 | PubMed

Fechner GT (1860) *Elemente der psychophysik* **2** Breitkopf u. Härtel.

Foley JM, Legge GE (1981) Contrast detection and near-threshold discrimination in human vision. *Vision research* **21**:1041-1053 https://doi.org/10.1016/0042-6989(81)90009-2 | PubMed

Freeman TC, Leung J, Wufong E, Orchard-Mills E, Carlile S, Alais D. (2014) Discrimination contours for moving sounds reveal duration and distance cues dominate auditory speed perception. *PloS one* **9**:e102864 https://doi.org/10.1371/journal.pone.0102864 | PubMed

Garside DJ, Chang AL, Selwyn HM, Conway BR (2025) The origin of color categories. *Proceedings of the National Academy of Sciences* **122**:e2400273121 https://doi.org/10.1073/pnas.2400273121 | PubMed

Gegenfurtner KR (2025) The Verriest Lecture: Color vision from pixels to objects. *Journal of the Optical Society of America A* **42**:B313-B328 https://doi.org/10.1364/josaa.544136 | PubMed

Girshick AR, Landy MS, Simoncelli EP (2011) Cardinal rules: visual orientation perception reflects knowledge of envi-ronmental statistics. *Nature neuroscience* **14**:926-932 https://doi.org/10.1038/nn.2831 | PubMed

Gravesen J. (2015) The metric of colour space. *Graphical Models* **82**:77-86 https://doi.org/10.1016/j.gmod.2015.06.005

Green DM, Swets JA, et al. (1966) *Signal detection theory and psychophysics* **1** New York: Wiley.

Hansen T, Gegenfurtner KR (2013) Higher order color mechanisms: Evidence from noise-masking experiments in cone contrast space. *Journal of vision* **13**:26-26 https://doi.org/10.1167/13.1.26 | PubMed

Hautus MJ, Macmillan NA, Creelman CD (2021) *Detection theory: A user's guide* Routledge.

Hecht S, Shlaer S, Pirenne MH (1942) Energy, quanta, and vision. *Journal of General Physiology* **25**:819-840 https://doi.org/10.1085/jgp.25.6.819 | PubMed

Hedjar L, Toscani M, Gegenfurtner KR (2025) Importance of hue: color discrimination of three-dimensional objects and two-dimensional discs. *Journal of the Optical Society of America A* **42**:B296-B304 https://doi.org/10.1364/josaa.544380 | PubMed

Hillis JM, Brainard DH (2007) Distinct mechanisms mediate visual detection and identification. *Current Biology* **17**:1714-1719 https://doi.org/10.1016/j.cub.2007.09.012 | PubMed

Hofer H, Carroll J, Neitz J, Neitz M, Williams DR (2005) Organization of the human trichromatic cone mosaic. *Journal of Neuroscience* **25**:9669-9679 https://doi.org/10.1523/jneurosci.2414-05.2005 | PubMed

Hong F, Badde S, Landy MS (2021) Causal inference regulates audiovisual spatial recalibration via its influence on audiovisual perception. *PLOS Computational Biology* **17**:1-37 https://doi.org/10.1371/journal.pcbi.1008877 | PubMed

Hong F, Chow J, Guan P, Brainard DH, Williams AH (2026) The Geometry of Color Space: Suprathreshold Differences from Discrimination Thresholds. *Journal of Vision*

Horiuchi S, Nagai T. (2024) Color discrimination repetition distorts color representations. *Scientific Reports* **14**:9615 https://doi.org/10.1038/s41598-024-60283-4 | PubMed

Hurvich LM, Hurvich-Jameson D. (1961) *Opponent chromatic induction and wavelength discrimination* Springer.

Johnson CA, Wall M, Thompson HS (2011) A history of perimetry and visual field testing. *Optometry and Vision Science* **88**:E8-E15 https://doi.org/10.1097/opx.0b013e3182004c3b | PubMed

Knoblauch K, Maloney LT (1996) Testing the indeterminacy of linear color mechanisms from color discrimination data. *Vision research* **36**:295-306 https://doi.org/10.1016/0042-6989(95)00098-k | PubMed

Knoblauch K, Maloney LT (2012) *Modeling psychophysical data in R* **32** Springer Science & Business Media.

Koenderink J, van Doorn A, Braun DI, Gegenfurtner KR (2026) An empirical three-dimensional metric field for color space. *bioRxiv* https://doi.org/10.64898/2026.03.09.710376

Koenderink JJ (2010) *Color for the Sciences* MIT press.

Krauskopf J, Gegenfurtner KR (1992) Color discrimination and adaptation. *Vision research* **32**:2165-2175 https://doi.org/10.1016/0042-6989(92)90077-v | PubMed

Kremers J, Scholl HP, Knau H, Berendschot TT, Usui T, Sharpe LT (2000) L/M cone ratios in human trichromats as-sessed by psychophysics, electroretinography, and retinal densitometry. *Journal of the Optical Society of America A* **17**:517-526 https://doi.org/10.1364/josaa.17.000517 | PubMed

de Lange Dzn H. (1958) Research into the dynamic nature of the human fovea cortex systems with intermittent and modulated light. I. Attenuation characteristics with white and colored light. *Journal of the Optical Society of America* **48**:777-784 https://doi.org/10.1364/josa.48.000777 | PubMed

Lesmes LA, Lu ZL, Baek J, Albright TD (2010) Bayesian adaptive estimation of the contrast sensitivity function: the quick CSF method. *Journal of vision* **10**:17-17 https://doi.org/10.1167/10.3.17 | PubMed

Letham B, Guan P, Tymms C, Bakshy E, Shvartsman M. (2022) Look-ahead acquisition functions for Bernoulli level set estimation. In: International Conference on Artificial Intelligence and Statistics. PMLR. pp. 8493-8513

Loomis JM, Berger T. (1979) Effects of chromatic adaptation on color discrimination and color appearance. *Vision Research* **19**:891-901 https://doi.org/10.1016/0042-6989(79)90023-3 | PubMed

MacAdam DL (1942) Visual sensitivities to color differences in daylight. *Journal of the Optical Society of America* **32**:247-274 https://doi.org/10.1364/josa.32.000247

Macadam DL (1979) Judd's contributions to color metrics and evaluation of color differences. *Color Research & Application* **4**:177-193 https://doi.org/10.1002/col.5080040402

MacLeod DI, Boynton RM (1979) Chromaticity diagram showing cone excitation by stimuli of equal luminance. *Journal of the Optical Society of America* **69**:1183-1186 https://doi.org/10.1364/josa.69.001183 | PubMed

McDonald R, Smith KJ (1995) CIE94-a new colour-difference formula. *Journal of the Society of Dyers and Colourists* **111**:376-379 https://doi.org/10.1111/j.1478-4408.1995.tb01688.x

Mullen K, Ennis DM (1991) A simple multivariate probabilistic model for preferential and triadic choices. *Psychome-trika* **56**:69-75 https://doi.org/10.1007/bf02294586

Muzellec B, Cuturi M. (2018) Generalizing Point Embeddings using the Wasserstein Space of Elliptical Distributions. In: Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R (Eds). *Advances in Neural Information Processing Systems* **31** Curran Associates, Inc. https://doi.org/10.48550/arxiv.1805.07594

Najemnik J, Geisler WS (2005) Optimal eye movement strategies in visual search. *Nature* **434**:387-391 https://doi.org/10.1038/nature03390 | PubMed

Neitz J, Jacobs GH (1986) Polymorphism of the long-wavelength cone in normal human colour vision. *Nature* **323**:623-625 https://doi.org/10.1038/323623a0 | PubMed

Newton JR, Eskew RT (2003) Chromatic detection and discrimination in the periphery: a postreceptoral loss of color sensitivity. *Visual neuroscience* **20**:511-521 https://doi.org/10.1017/s0952523803205058 | PubMed

Niwa Y, Muraki S, Naito F, Minamikawa T, Ohji M. (2014) Evaluation of acquired color vision deficiency in glaucoma using the Rabin cone contrast test. *Invest Ophthalmol Vis Sci* **55**:6686-90 https://doi.org/10.1167/iovs.14-14079 | PubMed

Noorlander C, Heuts MJ, Koenderink JJ (1981) Sensitivity to spatiotemporal combined luminance and chromaticity contrast. *Journal of the Optical Society of America* **71**:453-459 https://doi.org/10.1364/josa.71.000453 | PubMed

Noorlander C, Koenderink JJ, Den Olden RJ, Edens BW (1983) Sensitivity to spatiotemporal colour contrast in the peripheral visual field. *Vision Research* **23**:1-11 https://doi.org/10.1016/0042-6989(83)90035-4 | PubMed

Olkkonen M, McCarthy PF, Allred SR (2014) The central tendency bias in color perception: Effects of internal and external noise. *Journal of vision* **14**:5-5 https://doi.org/10.1167/14.11.5 | PubMed

Owen L, Browder J, Letham B, Stocek G, Tymms C, Shvartsman M. (2021) Adaptive nonparametric psychophysics. *arXiv* https://doi.org/10.48550/arxiv.2104.09549

Palmer J, Ames CT, Lindsey DT (1993) Measuring the effect of attention on simple visual search. *Journal of Experi-mental Psychology: Human Perception and Performance* **19**:108 https://doi.org/10.1037//0096-1523.19.1.108 | PubMed

Pointer MR (1974) Color discrimination as a function of observer adaptation. *Journal of the Optical Society of America* **64**:750-759 https://doi.org/10.1364/josa.64.000750 | PubMed

Poirson AB, Wandell BA (1990) The ellipsoidal representation of spectral sensitivity. *Vision research* **30**:647-652 https://doi.org/10.1016/0042-6989(90)90075-v | PubMed

Prins N, et al. (2016) *Psychophysics: a practical introduction* Academic Press.

Rad KR, Paninski L. (2010) Efficient, adaptive estimation of two-dimensional firing rate surfaces via Gaussian process methods. *Network: Computation in Neural Systems* **21**:142-168 https://doi.org/10.3109/0954898x.2010.532288 | PubMed

Reisbeck TE, Gegenfurtner KR (1999) Velocity tuned mechanisms in human motion processing. *Vision research* **39**:3267-3286 https://doi.org/10.1016/s0042-6989(99)00017-6 | PubMed

Rezeanu D, Neitz M, Neitz J. (2023) From cones to color vision: a neurobiological model that explains the unique hues. *Journal of the Optical Society of America A* **40**:A1-A8 https://doi.org/10.1364/josaa.477227 | PubMed

Roberti V. (2024) Helmholtz, Schrödinger, and the First Non-Euclidean Model of Perceptual Color Space. *Annalen der Physik* **536**:2300536 https://doi.org/10.1002/andp.202300536

Robertson AR, Lozano RD, Alman DH, Orchard S, Keitch J, Connely R, Graham L, Acree W, John R, Hoban R, *et al.* (1977) CIE recommendations on uniform color spaces, color-difference equations, and metric color terms. *Color Res Appl* **2**:3 https://doi.org/10.1002/j.1520-6378.1977.tb00102.x

Savin C, Tkacik G. (2016) Estimating nonlinear neural response functions using gp priors and kronecker methods. In: Advances in Neural Information Processing Systems. **29**

Schrödinger Ev. (1920) Outline of a theory of color measurement for daylight vision. *Physics Annual* **63**:397-520

Sharma G, Wu W, Dalal EN (2005) The CIEDE2000 color-difference formula: Implementation notes, supplementary test data, and mathematical observations. *Color Research & Application* **30**:21-30 https://doi.org/10.1002/col.20070

Shepard TG, Lahlaf SI, Eskew RT (2017) Labeling the lines: A test of a six-mechanism model of chromatic detection. *Journal of Vision* **17**:9-9 https://doi.org/10.1167/17.13.9 | PubMed

Shepard TG, Swanson EA, McCarthy CL, Eskew RT (2016) A model of selective masking in chromatic detection. *Journal of vision* **16**:3-3 https://doi.org/10.1167/16.9.3 | PubMed

Shevell SK, Martin PR (2017) Color opponency: tutorial. *Journal of the Optical Society of America A* **34**:1099-1108 https://doi.org/10.1364/josaa.34.001099 | PubMed

Sobol IM (1967) The distribution of points in a cube and the approximate evaluation of integrals. *USSR Computational mathematics and mathematical physics* **7**:86-112 https://doi.org/10.1016/0041-5553(67)90144-9

Stark E, Turton TL, Bujack R. (2025) Diminishing Returns in Perceptual Color Space-Now in Color. In: EuroVis 2025 - 27th EG Conference on Visualization.

Stein EM, Shakarchi R. (2011) *Fourier analysis: an introduction* **1** Princeton University Press.

Stockman A, Brainard DH, et al. (2010) Color vision mechanisms. In: *OSA handbook of optics* **3** pp. 11-1

Taylor G (2017) python-colormath.

Thibos LN, Applegate RA, Schwiegerling JT, Webb R. (2000) Standards for reporting the optical aberrations of eyes. In: Vision science and its applications. pp. SuC1

Vemala R, Sivaprasad S, Barbur JL (2017) 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 https://doi.org/10.1167/iovs.17-21771 | PubMed

Wandell BA (1985) Color measurement and discrimination. *Journal of the Optical Society of America A* **2**:62-71 https://doi.org/10.1364/josaa.2.000062 | PubMed

Wandell BA (1995) *Foundations of vision* Sinauer Associates.

Wardle SG, Alais D. (2013) Evidence for speed sensitivity to motion in depth from binocular cues. *Journal of Vision* **13**:17-17 https://doi.org/10.1167/13.1.17 | PubMed

Wasserman L. (2006) *All of nonparametric statistics* Springer.

Watson AB (2017) QUEST+: A general multidimensional Bayesian adaptive psychometric method. *Journal of Vision* **17**:10-10 https://doi.org/10.1167/17.3.10 | PubMed

Waz S, Wang Y, Lu ZL (2025) Quantification of retinotopic maps with a Gaussian process modeling. *Journal of Vision* **25**:20-20 https://doi.org/10.1167/jov.25.8.20 | PubMed

Williams CK, Rasmussen CE (2006) *Gaussian processes for machine learning* **2** Cambridge, MA: MIT press.

Wilson AG, Ghahramani Z. (2011) Generalised Wishart processes. In: Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence UAI'11. Arlington, United States. AUAI Press. pp. 736-744

Winawer J, Witthoft N. (2023) Effects of color terms on color perception and cognition. In: Shamey R (Ed). *Encyclopedia of color science and technology* Springer. pp. 777-785 https://doi.org/10.1007/978-3-030-89862-5_77

Wuerger SM, Maloney LT, Krauskopf J. (1995) Proximity judgments in color space: tests of a Euclidean color geometry. *Vision research* **35**:827-835 https://doi.org/10.1016/0042-6989(94)00170-q | PubMed

Wyszecki G. (1982) *Concepts and methods, quantitative data and formulae* pp. 130-175

Zaidi Q (2001) Is there a perceptual color space?. *Color Res Appl* https://doi.org/10.1002/col.1041

Zernike F. (1934) Beugungstheorie des Schneidenverfahrens und seiner verbesserten Form, der Phasenkontrast-methode. *Physica* **1**:689-704 https://doi.org/10.1016/s0031-8914(34)80259-5

Zhang C, Wang K, Chow J, Guan P, Williams AH, Brainard DH, Hong F. (2026) Chromatic Adaptation Systematically Reshapes Human Color Discrimination Thresholds. *Journal of Vision*

Zhou J, Duong LR, Simoncelli EP (2024) A unified framework for perceived magnitude and discriminability of sensory stimuli. *Proceedings of the National Academy of Sciences* **121**:e2312293121 https://doi.org/10.1073/pnas.2312293121 | PubMed

Peer reviews

Reviewer #1 (Public review):

Summary:

This paper presents an ambitious and technically impressive attempt to map how well humans can discriminate between colours across the entire isoluminant plane. The authors introduce a novel Wishart Process Psychophysical Model (WPPM) - a Bayesian method that estimates how visual noise varies across colour space. Using an adaptive sampling procedure, they then obtain a dense set of discrimination thresholds from relatively few trials, producing a smooth, continuous map of perceptual sensitivity. They validate their procedure by comparing actual and predicted thresholds at an independent set of sample points. The work is a valuable contribution to computational psychophysics and offers a promising framework for modelling other perceptual stimulus fields more generally.

Strengths:

The approach is elegant and well-described, and the data are of high quality. The writing throughout is clear and the figures are clean (elegant in fact) and do a good job of explaining how the analysis was performed. The whole paper is tremendously thorough and the technical appendices and attention to detail are impressive (for example, a huge amount of data about calibration, variability of the stim system over time etc). This should be a touchstone for other papers that use calibrated colour stimuli.

Comments on revised version:

The authors have addressed all the issues I raised to my satisfaction.

https://doi.org/10.7554/eLife.108943.2.sa2

Reviewer #3 (Public review):

Summary:

This study presents a powerful and rigorous approach for characterizing stimulus discriminability throughout a sensory manifold, and is applied to the specific context of predicting color discrimination thresholds across the chromatic plane.

Strengths:

Color discrimination has played a fundamental role in studies of human color vision and for color applications, but as the authors note, remains poorly characterized. The study leverages the assumption that thresholds should vary smoothly and systematically within the space, and validates this with their own tests and comparisons with previous studies.

Comments on revised version:

My comments have been addressed.

https://doi.org/10.7554/eLife.108943.2.sa1

Author response:

The following is the authors' response to the original reviews.

We would like to thank the editors and the reviewers for the thorough and insightful comments and suggestions. Addressing them has strengthened our manuscript. We have carefully addressed all reviewer comments, as described in detail below, as well as additional comments we received from others. In addition, we made two substantive updates to the manuscript:

(1) We improved the estimation of uncertainty in the model predictions by computing 95% confidence intervals using 120 bootstrapped datasets (instead of the 100% of 10 bootstrapped datasets in the original submission) to match the number of bootstrap for the validation dataset.

(2) We selected a slightly different hyperparameter value based on follow-up analyses suggested by Reviewer 1, which provided very useful information.

Importantly, none of these changes alter the main results or conclusions of the paper.

Beyond these changes and those outlined below, we also worked to improve the clarity of the prose throughout as well as added various additional citations to the literature.

> ***Public Reviews:***
>
> ***Reviewer #1 (Public review):***
>
> *Summary:*
>
> *This paper presents an ambitious and technically impressive attempt to map how well humans can discriminate between colours across the entire isoluminant plane. The authors introduce a novel Wishart Process Psychophysical Model (WPPM) - a Bayesian method that estimates how visual noise varies across colour space. Using an adaptive sampling procedure, they then obtain a dense set of discrimination thresholds from relatively few trials, producing a smooth, continuous map of perceptual sensitivity. They validate their procedure by comparing actual and predicted thresholds at an independent set of sample points. The work is a valuable contribution to computational psychophysics and offers a promising framework for modelling other perceptual stimulus fields more generally.*
>
> *Strengths:*
>
> *The approach is elegant and well-described (I learned a lot!), and the data are of high quality. The writing throughout is clear, and the figures are clean (elegant in fact) and do a good job of explaining how the analysis was performed. The whole paper is tremendously thorough, and the technical appendices and attention to detail are impressive (for example, a huge amount of data about calibration, variability of the stim system over time, etc). This should be a touchstone for other papers that use calibrated colour stimuli.*
>
> *Weaknesses:*
>
> *Overall, the paper works as a general validation of the WPPM approach. Importantly, the authors validate the model for the particular stimuli that they use by testing model predictions against novel sample locations that were not part of the fitting procedure*

(Figure 2). The agreement is pretty good, and there is no overall bias (perhaps local bias?), but they do note a statistically-significant deviation in the shape of the threshold ellipses. The data also deviate significantly from historical measurements, and I think the paper would be considerably stronger with additional analyses to test the generality of its conclusions and to make clearer how they connect with classical colour vision research. In particular, three points could use some extra work:

(1) Smoothness prior.

The WPPM assumes that perceptual noise changes smoothly across colour space, but the degree of smoothness (the eta parameter) must affect the results. I did not see an analysis of its effects - it seems to be fixed at 0.5 (line 650). The authors claim that because the confidence intervals of the MOCS and the model thresholds overlap (line 223), the smoothing is not a problem, but this might just be because the thresholds are noisy. A systematic analysis varying this parameter (or at least testing a few other values), and reporting both predictive accuracy and anisotropy magnitude, would clarify whether the model's smoothness assumption is permitting or suppressing genuine structure in the data. Is the gamma parameter also similarly important? In particular, does changing the underlying smoothness constraint alter the systematic deviation between the model and the MOCS thresholds? The authors have thought about this (of course! - line 224), but also note a discrepancy (line 238). I also wonder if it would be possible to do some analysis on the posterior, which might also show if there are some regions of color space where this matters more than others? The reason for doing this is, in part, motivated by the third point below - it's not clear how well the fits here agree with historical data.

Thank you for raising this important point. We have now added analyses of the effects of the two smoothness-related hyperparameters, ε and γ (see Appendix 10).

First, we swept a range of values for each hyperparameter (ε: 0.1 – 1; γ: 0.000001 – 0.003) and evaluated model performance using 5-fold cross-validation of the dataset used to fit the WPPM, quantifying predictive accuracy on held-out test data. We used the mean negative log likelihood averaged across the held-out data in the cross validation as our measure of predictive accuracy (Figs. S27-31).

The two hyperparameters affect cross-validation accuracy in a similar manner. With γ fixed at 0.0003, predictive accuracy is highest for ε in the range of approximately 0.3–0.5 and drops quite rapidly for ε < 0.3. We attribute this drop to oversmoothing. Cross-validation accuracy also decreases, albeit more gradually, for ε > 0.5. We attribute this to increased variance due to undersmoothing relative to the power of our datasets. Similarly, with ε fixed at 0.4, predictive accuracy is highest for γ values between approximately 0.0001 and 0.001, declines rapidly for smaller γ (oversmoothing), and more slowly for larger γ (undersmoothing).

Second, we examined how the hyperparameter ε affected the agreement between the WPPM fit and the MOCS validation data. Specifically, at each ε, for each participant, we computed the linear regression between WPPM thresholds and validation thresholds at 25 reference locations. Then, we examined the slope and correlation coefficient of all participants as a function of ε. We found a classic bias–variance tradeoff. Excessive smoothness introduces bias by failing to capture structure in the data, whereas insufficient smoothness increases variance in model predictions. These results further support a choice of ε = 0.4 as lying near the optimal balance between bias and variance (Fig. S32).

Based on these analyses, we selected for the final analysis ε = 0.4, slightly smaller than the preregistered value used in the original submission (0.5), while retaining the original value of γ (0.0003).

We now discuss these reasons for changing this value in the revision, as well as provide a more general discussion of the importance and practicalities of hyperparameter choice in Bayesian approaches to analyzing data (Discussion / Prior specification).

> *(2) Comparison with simpler models. It would help to see whether the full WPPM is genuinely required. Clearly, the data (both here and from historical papers) require some sort of anisotropy in the fitting - the sensitivities decrease as the stimuli move away from the adaptation point. But it's >not< clear how much the fits benefit from the full parameterisation used here. Perhaps fits for a small hierarchy of simpler models - starting with isotropic Gaussian noise (as a sort of 'null baseline') and progressing to a few low-dimensional variants - would reveal how much predictive power is gained by adding spatially varying anisotropy. This would demonstrate that the model's complexity is justified by the data.*

In the 5-fold cross-validation analysis described above (and now presented in Appendix 10), we found that when ε or γ is small, the stronger smoothness constraint leads to threshold ellipses that are nearly identical to each other across color space. Under these conditions, model predictions show poor accuracy on held-out test data and lead to poor predictions of the validation data. This observation addresses the underlying point raised by the reviewer, albeit in a different way than suggested: it shows that a degree of spatially varying anisotropy is necessary to capture the structure of the data. We now make this point in the paper (Discussion / Prior specification).

More broadly, we employed the WPPM as a prior that imposed smoothness but not much other obvious structure, and used this to learn about the psychometric field. We are currently working to understand how we can best use our current data to improve the prior we would apply to future measurements. There are a number of approaches to this. One would be to seek a parametric mechanistic model that can describe the current data, and to the extent this is possible formulate prior distributions over the parameters of the model. The results reported here thus provide a foundation for deriving and evaluating more structured priors that would even more efficiently leverage future datasets, but with the feature that they impose more structure. We have added this perspective to the Discussion / Extensions of the WPPM framework.

> *(3) Quantitative comparison to historical data. The paper currently compares its results to MacAdam, Krauskopf & Karl, and Danilova & Mollon only by visual inspection. It is hard to extract and scale actual data from historical papers, but from the quality of the plotting here, it looks like the authors have achieved this, and so quantitative comparisons are possible. The MacAdam data comparisons are pretty interesting - in particular, the orientations of the long axes of the threshold ellipses do not really seem to line up between the two datasets - and I thought that the orientation of those ellipses was a critical feature of the MacAdam data. Quantitative comparisons (perhaps overall correlations, which should be immune to scaling issues, axis-ratio, orientation, or RMS differences) would give concrete measures of the quality of the model. I know the authors spend a lot of time comparing to the CIE data, and this is great.... But re-expressing the fitted thresholds in CIE or DKL coordinates, and comparing them directly with classical datasets, would make the paper's claims of "agreement" much more convincing.*

Although we are sympathetic to this request, we have chosen not to implement the sort of quantitative comparison requested by the reviewer. The reason is that an important feature of color thresholds is that they depend on the spatial (e.g. Kelly, 1974; Poirson & Wandell, 1996; Danilova & Mollon, 2025) and temporal (e.g. Kelly, 1974) properties of the stimuli, and on the observer's state of adaptation (e.g. Loomis & Berger, 1979; Krauskopf & Gegenfurtner, 1992). Because (as the reviewer notes below) the spatial and temporal properties of our stimuli were not matched to those of the comparison datasets, our purpose in making these

comparisons was to examine qualitative agreement, as well as to situate our results in the literature and to demonstrate that our approach allows us to read out thresholds around the references and in the color spaces used in other studies. We would not expect detailed quantitative agreement with the current dataset because of differences in stimuli.

As a consequence of this, we think we would be overreaching to quantify the differences between our data and classic datasets. This consideration is particularly important for the MacAdam measurements, where because of the matching adjustment procedure used, the observer's state of adaptation is likely to have varied (by amounts that are difficult to estimate) from one reference to the next (e.g. Danilova & Mollon, 2025). We have clarified the manuscript with respect to these points (Results / Comparison with previous measurements).

A point to make on this topic is that an important and interesting future direction that emerges from our work is to develop efficient methods to characterize the dependence of the full discrimination field on ancillary variables, such as those that describe spatial and temporal properties and/or the state of adaptation, which we now also mention in the paper (Discussion / Implications for the mechanisms of color perception). Although not the primary motivation, doing so would enable comparison of data with a wider range of studies.

We do agree that the comparisons to CIELAB predictions work better when we express them in CIELAB, and have now done so (Fig. 3D; Fig. S24-S26).

Kelly, D. H. (1974). "Spatio-temporal frequency characteristics of color-vision mechanisms." Journal of the Optical Society of America 64(7): 983–990.

Poirson, A. B. and B. A. Wandell (1996). "Pattern-color separable pathways predict sensitivity to simple colored patterns " Vision Research 36(4): 515–526.

Danilova, M. V. and J. D. Mollon (2025). "Effect of stimulus size on chromatic discrimination." Journal of the Optical Society of America A 42(5).

Loomis, J. M. and T. Berger (1979). "Effects of chromatic adaptation on color discrimination and color appearance." Vision Research 19(8): 891–901.

Krauskopf, J., Gegenfurtner, K. (1992). "Color discrimination and adaptation." Vision Research 32(11): 2165–2175.

> *Overall, this is a creative and technically sophisticated paper that will be of broad interest to vision scientists. It is probably already a definitive method paper showing how we can sample sensitivity accurately across colour space (and other visual stimulus spaces). But I think that until the comparison with historical datasets is made clear (and, for example, how the optimal smoothness parameters are estimated), it has slightly less to tell us about human colour vision. This might actually be fine - perhaps we just need the methods?*
>
> *Related to this, I'd also note that the authors chose a very non-standard stimulus to perform these measurements with (a rendered 3D 'Greebley' blob). This does have the advantage of some sort of ecological validity. But it has the significant disadvantage that it is unlike all the other (much simpler) stimuli that have been used in the past - and this is likely to be one of the reasons why the current (fitted) data do not seem to sit in very good agreement with historical measurements.*

As the reviewer notes, our stimuli head in the direction of ecological validity (see also Hedjar et al., 2025) and indeed this was a consideration when we chose them, at the cost of limiting the degree of comparison we can make with prior studies (as discussed above). Another reason we chose our stimuli is that they enable the current data to be used as a basis of comparison with stimuli where we add specularity, change object shape, and vary object pose

in the future. These manipulations are not possible with flat matte patches. Such experiments are of interest to us, as they will tell us about how effectively color may be used to differentiate stimuli in cases where other ecologically important variables co-vary. We now mention this motivation in the paper (Results / Task and Stimuli).

Hedjar, L., M. Toscani and K. R. Gegenfurtner (2025). "Importance of hue: color discrimination of three-dimensional objects and two-dimensional discs." Journal of the Optical Society of America A 42(5).

> ***Reviewer #2 (Public review):***
>
> *Summary:*
>
> *Hong et al. present a new method that uses a Wishart process to dramatically increase the efficiency of measuring visual sensitivity as a function of stimulus parameters for stimuli that vary in a multidimensional space. Importantly, they have validated their model against their own hold-out data and against 3 published datasets, as well as against colour spaces aimed at 'perceptual uniformity' by equating JNDs. Their model achieves high predictive success and could be usefully applied in colour vision science and psychophysics more generally, and to tackle analogous problems in neuroscience featuring smooth variation over coordinate spaces.*
>
> *Strengths:*
>
> *(1) This research makes a substantial contribution by providing a new method to very significantly increase the efficiency with which inferences about visual sensitivity can be drawn, so much so that it will open up new research avenues that were previously not feasible. Secondly, the methods are well thought out and unusually robust. The authors made a lot of effort to validate their model, but also to put their results in the context of existing results on colour discrimination, transforming their results to present them in the same colour spaces as used by previous authors to allow direct comparisons. Hold-out validation is a great way to test the model, and this has been done for an unusually large number of observers (by the standards of colour discrimination research). Thirdly, they make their code and materials freely available with the intention of supporting progress and innovation. These tools are likely to be widely used in vision science, and could of course be used to address analogous problems for other sensory modalities and beyond.*
>
> *Weaknesses:*
>
> *It would be nice to better understand what constraints the choice of basis functions puts on the space of possible solutions. More generally, could there be particular features of colour discrimination (e.g., rapid changes near the white point) that the model captures less well.*

This comment bears conceptual similarity to Reviewer 1's question about the hyperparameters of our prior, as it is basically asking whether we might be oversmoothing through the choice of form and number of basis functions. The hyperparameter sweeps we now present suggest that within the choice of basis functions we used, we are operating at a reasonable point on the bias-variance tradeoff curve - we can see bias emerging with a smoother prior, and variance increasing with a less smooth prior. Our expectation is that varying the smoothness of the prior in other ways, such as by varying the form and number of the basis functions, would lead to similar tradeoffs.

We did perform one additional check that shows, within our current framework, that adding more basis functions is unlikely to change things much. This was to plot the fit weights as a function of Chebyshev basis order (Figure S4 in Appendix 2). These decline to near zero at the

highest order we used, suggesting that adding more would not alter the inferred psychometric field, given our hyperparameter choices. Although we could explore this question further by explicitly fitting the data using more basis functions along with different hyperparameter choices, or different functional forms for the basis functions, we decided not to pursue this in favor of performing the other additional analyses we now present.

We resonate with the reviewer's concern that assuming smoothness, both by assuming that isoperformance contours are elliptical and by assuming that these vary smoothly with reference, might cause us to miss features of the true underlying field in cases where that field varies rapidly or the isoperformance contours are asymmetric or non-elliptical. Our approach to this was to measure the validation thresholds and demonstrate that any bias in our WPPM-inferred field is small for these measurements. Because we shared the reviewer's intuition that the adapting point is a candidate location where there might be less smooth variation, we measured a validation threshold at this reference for every subject. Nonetheless, we only measured in one direction around the adapting reference for each subject. We considered validation approaches where we measured full ellipses at a set of validation references, but we were worried about effects of uncertainty reduction and perceptual learning which might distort thresholds at highly sampled locations.

It is the case that if one wanted to study the discrimination field in more detail around a particular reference, one could concentrate trials in a smaller model space around that reference, and for the same number of trials use a prior with less smoothness relative to the underlying stimulus space. Indeed, simply halving the size of the stimulus space that maps onto the [-1,1] model space and keeping the same prior over the model space effectively halves the degree of smoothness expressed with respect to the stimulus space. Thus our methods could prove useful in studying more rapid variations in the discrimination field if one hypothesized that they might occur around particular reference choices, but this would still rest upon the elliptical assumption. To relax that assumption, one could use the threshold field estimation methods implemented in AEPsych, which incorporate a smoothness assumption but do not assume elliptical isoperformance contours. Weakening the prior in this way would, however, increase trial demand to obtain similar measurement precision.

As a general matter, we don't think it is possible to leverage smoothness for trial efficiency on the one hand and at the same time be completely sure that there isn't some aspect to the underlying ground truth that has been smoothed over. Carefully choosing the degree of prior smoothness together with the number of experimental trials in the context of a particular content problem is an important part of bringing the WPPM and related methods to bear, and one where simulation and held-out data both play an important role.

We now bring these points out more fully in the paper (Discussion / Extensions of the WPPM framework; Discussion / Prior specification).

Chen, C.-C., J. M. Foley and D. H. Brainard (2000). "Detection of chromoluminance patterns on chromoluminance pedestals I: threshold measurements." Vision Research 40(7): 773–788.

> *The substantial individual differences evident in Figure S20 (comparison with Krauskopf and Gegenfurtner, 1992) are interesting in this context. Some observers show radial biases for the discrimination ellipses away from the white point, some show biases along the negative diagonal (with major axes oriented parallel to the blue-yellow axis), and others show a mixture of the two biases. Are these genuine individual differences, or could the model be performing less accurately in this desaturated region of colour space?*

We agree that these differences are interesting. We have now added more complete bootstrapped confidence regions in these (Appendix 8) and the other comparison figures (Appendix 6, 7, 9), so that an estimate of measurement precision is directly available in these

figures. These confidence regions suggest that the individual differences in this region of color space are real. A longer-term goal is to develop more mechanistic models that can account for individual subject data through parameter choice. This might lead to insight into what differs in the visual system across individuals.

> ***Reviewer #3 (Public review):***
>
> *Summary:*
>
> *This study presents a powerful and rigorous approach for characterizing stimulus discriminability throughout a sensory manifold, and is applied to the specific context of predicting color discrimination thresholds across the chromatic plane.*
>
> *Strengths:*
>
> *Color discrimination has played a fundamental role in studies of human color vision and for color applications, but as the authors note, it remains poorly characterized. The study leverages the assumption that thresholds should vary smoothly and systematically within the space, and validates this with their own tests and comparisons with previous studies.*
>
> *Weaknesses:*
>
> *The paper assumes that threshold variations are due to changes in the level of intrinsic noise at different stimulus levels. However, it's not clear to me why they could not also be explained by nonlinearities in the responses, with fixed noise. Indeed, most accounts of contrast coding (which the study is at least in part measuring because the presentation kept the adapt point close to the gray background chromaticity, and thus measured increment thresholds), assume a nonlinear contrast response function, which can at least as easily explain why the thresholds were higher for colors farther from the gray point. It would be very helpful if a section could be added that explains why noise differences rather than signal differences are assumed and how these could be distinguished. If they cannot, then it would be better to allow for both and refer to the variation in terms of S/N rather than N alone.*

We agree with the reviewer. We are measuring SNR and attributing it to noise, but cannot identify from the data whether changes in SNR across color spaces are due to changes in noise, to a nonlinear relationship between stimulus space and the observer's response space with noise in the response space held fixed, or both. We now make this point where we introduce the Results / Wishart Process Psychophysical Model and reiterate it in the Discussion / Extensions of the

WPPM framework.

> *Related to this point, the authors note that the thresholds should depend on a number of additional factors, including the spatial and temporal properties and the state of adaptation. However, many of these again seem to be more likely to affect the signal than the noise.*

We don't disagree. Indeed, as we noted in our response to a comment by Reviewer 1 and above in the context of individual differences, we are very interested in developing a mechanistically plausible model that accounts for the data. If we or others are able to do so, that would provide a basis for parsing performance into separate signal and noise effects. And if such a model has natural ways in which additional variables affect its predictions, measuring the effects of these variables would be a way to provide evidence in favor of the model (Discussion / Implication for the mechanisms of color perception - Extensions of the WPPM framework).

> *An advantage of the approach is that it makes no assumptions about the underlying mechanisms. However, the choice to sample only within the equiluminant plane is itself a mechanistic assumption, and these could potentially be leveraged for deciding how to sample to improve the characterization and efficiency. For example, given what we know about early color coding, would it be more (or less) efficient to select samples based on a DKL space, etc?*

The more we are willing to assume about the structure of the psychometric field, the more efficiently we can measure it. As the reviewer correctly notes, this principle applies to trial placement as well. We are currently using an adaptive method (AEPsych) that starts with a fairly weak smoothness prior and attempts to place trials using heuristics that aim to minimize the expected uncertainty in the posterior. As we learn more about the discrimination field, we should be able to leverage stronger priors to increase trial efficiency. This point is closely related to one we made above about developing stronger priors that capture what we have learned in this study. Such priors could also help improve trial placement. For a prior that has a relatively small number of parameters, for example, perhaps a mechanistic prior, methods such as Quest+ (Watson, 2017) may be used for trial placement.

Watson, A. B. (2017). "QUEST+: A general multidimensional Bayesian adaptive psychometric method." J Vis 17(3): 10.

> ***Recommendations for the authors:***
>
> ***Reviewer #1 (Recommendations for the authors):***
>
> *I do not think that the authors need to perform additional experiments. However, I would like to see some additional analyses regarding the assumptions made in the fitting procedure and how they affect the final maps.*
>
> *I also think some more quantitative comparisons with historical data would be valuable - at the moment, a lot of the comparisons are simply 'by eye'.*
>
> *It would have been nice to have the code and data available during the review procedure - I'm sure these will be released with excellent documentation?*

We addressed the first two points in the public review section. The code is now available online as is the data. These links are now provided in the paper (Methods and Materials / Data and code availability).

> ***Reviewer #2 (Recommendations for the authors):***
>
> *Minor points*
>
> *I have a few suggestions for additions and small changes.*
>
> *(1) Several examples of covariance matrix fields are shown in Figure 1, 4, but these are for simulated examples. It would be nice to see the fields actually fit the data! I would be interested in seeing this for all participants in an Appendix, and maybe for participant CH in the main paper?*

We have made the changes (see Figure 4 and Figure S3).

> *(2) I have not worked through all the math in the appendices line by line, but it seems to be complete, and the model validation results speak for themselves. I think the authors have done a pretty good job of explaining the model conceptually (not easy), but I struggled with the 'weighted sum' step in Figure 4 and the main text. I would appreciate a bit more hand-holding here, e.g, why is an 'overcomplete' representation needed as an*

> *intermediate, and providing an intuition of why there are 12 matrices in the overcomplete representation and what each matrix in this representation represents.*

We have now added more explanations in the figure legend and text (Fig. 4 and Methods and Materials / The Wishart Process Psychometric Model).

> *(3) Individual differences: There is a section on this in the manuscript, and it's concluded that there are only "modest" individual differences. However, in Figure S20, the individual differences, I think, are huge and place observers almost in qualitatively different categories! Some observers show a radial bias in discrimination ellipses, others seem to show basically a bias along the negative diagonal, and others a mixture of both biases. These ellipses are at a desaturated part of colour space - is it possible that there are some rapid changes in the underlying noise in this region that the Wishart fit has not captured due to relatively sparse sampling or the fact that the basis functions are all fairly low spatial frequency? I wondered whether the results are constrained by the choice of Cartesian rather than polar basis functions, e.g, polar basis functions may have better allowed fine-grained changes near the white point but slower changes at higher saturations away from the white point.*

We agree that the individual differences are meaningful and, in some cases, quite pronounced. Our intent in describing the differences as “modest” was to emphasize that the overall structure of the psychometric fields remains broadly consistent across observers. We have revised the Results to note and more fully describe these differences.

Regarding the possibility that sharp changes in the underlying noise near the achromatic point might not be fully captured by the current model, we agree that this is an important consideration. The current implementation uses relatively low-order Chebyshev basis functions that primarily capture smooth global variations in the psychometric field. While validation analyses indicate that these basis functions capture the dominant structure in the data, they may be less sensitive to sharp local variations such as those that could occur near the white point. Future work could address this by mapping the model space to a smaller region around the achromatic reference or by exploring alternative basis sets (e.g., polar or Zernike functions) that may better capture such localized structure. This is discussed above in this response and now addressed in Discussion / Extensions of the WPPM framework.

> *On sampling, I wondered if the results might have been biased by the strongly biased ellipse that occurs at the grey point. If not, and the model is accurate in this region of colour space, I think this figure does show some large individual differences, and it would be good to comment on these in the individual differences section of the manuscript.*

Based on our analysis of trial placement (Fig. S1), the adaptive algorithm does not appear to have disproportionately concentrated trials near the gray point. In fact, more trials were allocated to the edges of the stimulus space than to the center. This suggests that the WPPM estimates are unlikely to be driven primarily by performance in the gray region. In addition, we examined the threshold ellipses around the gray reference in DKL space and found that they are broadly consistent across participants (Figs. S22–S23). Together, these analyses suggest that the anisotropy observed near the gray point reflects a genuine property of the psychometric field rather than an artifact of the sampling procedure.

As noted just above, we have added additional text about individual differences in the Results and referenced it in the Discussion.

> *(4) The manuscript seems unusually free of typographical errors, but I noticed that in many places "Krauskopf and Karl 1992" is cited! Also, I think something has gone wrong with the legend to Figure 2 - perhaps the order of panels was swapped around, but the*

> *legend was not fully updated. There is a repeated reference to the "summary of regression slopes" which seems to be in 2 positions, after C and G. It would make more sense to label panel G as D and progress from there, or switch the order of the panels so that G is on the bottom row.*

Thank you for catching those errors. They are now fixed.

> ***Reviewer #3 (Recommendations for the authors):***
>
> *A minor point (or perhaps major if your last name is Gegenfurtner) is that the reference to Krauskopf and Karl is incorrect.*

They are now fixed.

https://doi.org/10.7554/eLife.108943.2.sa0