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†]

***For correspondence:** fh862@sas.upenn.edu (FH)

[†]These authors contributed equally to this work

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

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 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 suprathreshold 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 individual stimulus dimensions. For example, pedestal functions characterize contrast discrimination threshold 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 significances 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 form is not known in advance.

Here, we show that it is possible to obtain a comprehensive characterization of the color discrimination psychometric field in the isoluminant plane. We achieved this by efficiently sampling reference-comparison stimulus pairs obtained using a non-parametric adaptive trial-placement procedure (***Owen et al., 2021***; ***Letham et al., 2022***), and then fitting the data *post hoc* with a semi-parametric 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 11.

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 1***A). The comparison stimulus was pseudo-randomly assigned to one of the three positions within the triangular arrangement. Participants were asked to identify the odd one out. Stimuli were rendered using the Unity graphics engine, and color was controlled by varying the specified surface reflectance using RGB (red, green, blue) coordinates, with other scene aspects held constant. We used naturalistic stimuli to increase the relevance of our results for understanding color vision in the real world. Hedjar and colleagues (***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 1***B; Appendix 1). We refer the space in which the transformed plane is as the *model space* because it is directly related to the way we formulated our semi-parametric model, and also served as a convenient representation for the non-parametric adaptive trial-placement procedure we used (see details in Covariance matrix field).

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 1***C-D). These covariance matrices, in turn, determine the entire psychometric field.

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

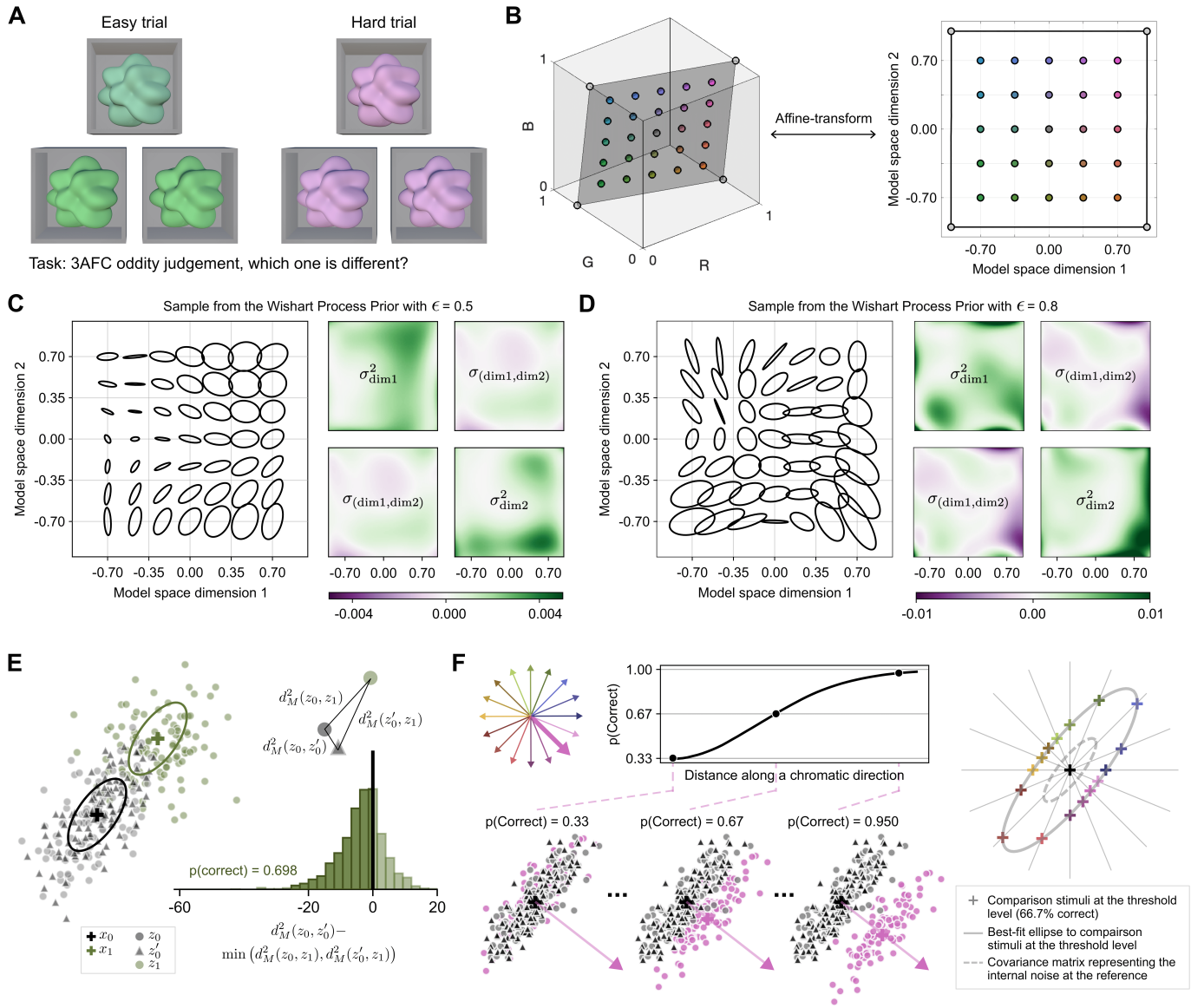

Figure 1. Task, stimuli and the WPPM. (A) 3AFC oddity task. On each trial, participants viewed a triplet of stimuli—two identical references and one different comparison—and identified the odd one out. (B) Stimuli were constrained to lie in the isoluminant plane the display's gray point. Data were represented and fit in a transformation of this plane which we refer to as *model space*. The grid of dots illustrates the transformation between the plane in the RGB and model space. (C) Example of a smoothly varying covariance matrix field produced by the WPPM. The field was generated by sampling from a finite-basis Wishart random process with a smooth prior ($\epsilon = 0.5$; see Prior over the weight matrix). Although the field is illustrated on a 7×7 grid, it specifies a covariance matrix $[\sigma^2_{\text{dim1}}, \sigma_{(\text{dim1,dim2})}; \sigma_{(\text{dim1,dim2})}, \sigma^2_{\text{dim2}}]$ for every stimulus in the plane, as shown in the heat maps. (D) Example of a less smoothly varying covariance matrix field. This field was obtained by drawing from a finite-basis Wishart random process with a less smooth prior ($\epsilon = 0.8$). (E) Observer model. For each stimulus triplet $[x_0, x'_0, x_1]$, internal representations $[z_0, z'_0, z_1]$ are drawn from multivariate Gaussian distributions centered at the reference stimulus with noise characterized by the corresponding covariance matrices. The model determines whether the observer correctly identifies the odd stimulus by comparing the squared Mahalanobis distances d^2_M between the representation pairs. (F) Derivation of elliptical threshold contour. One-dimensional psychometric functions are approximated using Monte Carlo simulations (10,000 samples per stimulus pair shown for illustration; 2,000 used during model fitting). For each selected chromatic direction, we derive the threshold distance corresponding to 66.7% correct. An ellipse is then fit to the threshold distances to describe the discrimination threshold contour. Notably, while the threshold contour and the noise ellipse shown at one standard deviation have well-matched shapes, the threshold contour is larger because it corresponds to 66.7% correct.

overly complex or erratic patterns of the covariance matrix field by giving preference to simple, smoothly varying covariance matrix fields. A single hyperparameter, ϵ, controls the strength of

this regularization (***Figure 1***C-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 (i.e., 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 1***F, 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 1***F, right panel). Given our assumption that internal noise follows a multivariate Gaussian distribution, these threshold distances form approximately elliptical contours, and we fit ellipses to them for data representation. This approach is consistent with prior work showing that ellipses provide a good approximation of color discrimination thresholds (***MacAdam, 1942***; ***Brown and MacAdam, 1949***; ***Noorlander et al., 1981***, ***1983***; ***Poirson and Wandell, 1990***; ***Krauskopf and Karl, 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 (Appendix 2); 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***). See ***Figure 2***A for an illustration of the procedure, and Design for more details.

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. These trials were included along with the 6000 AEPsych-driven trials when fitting the WPPM.

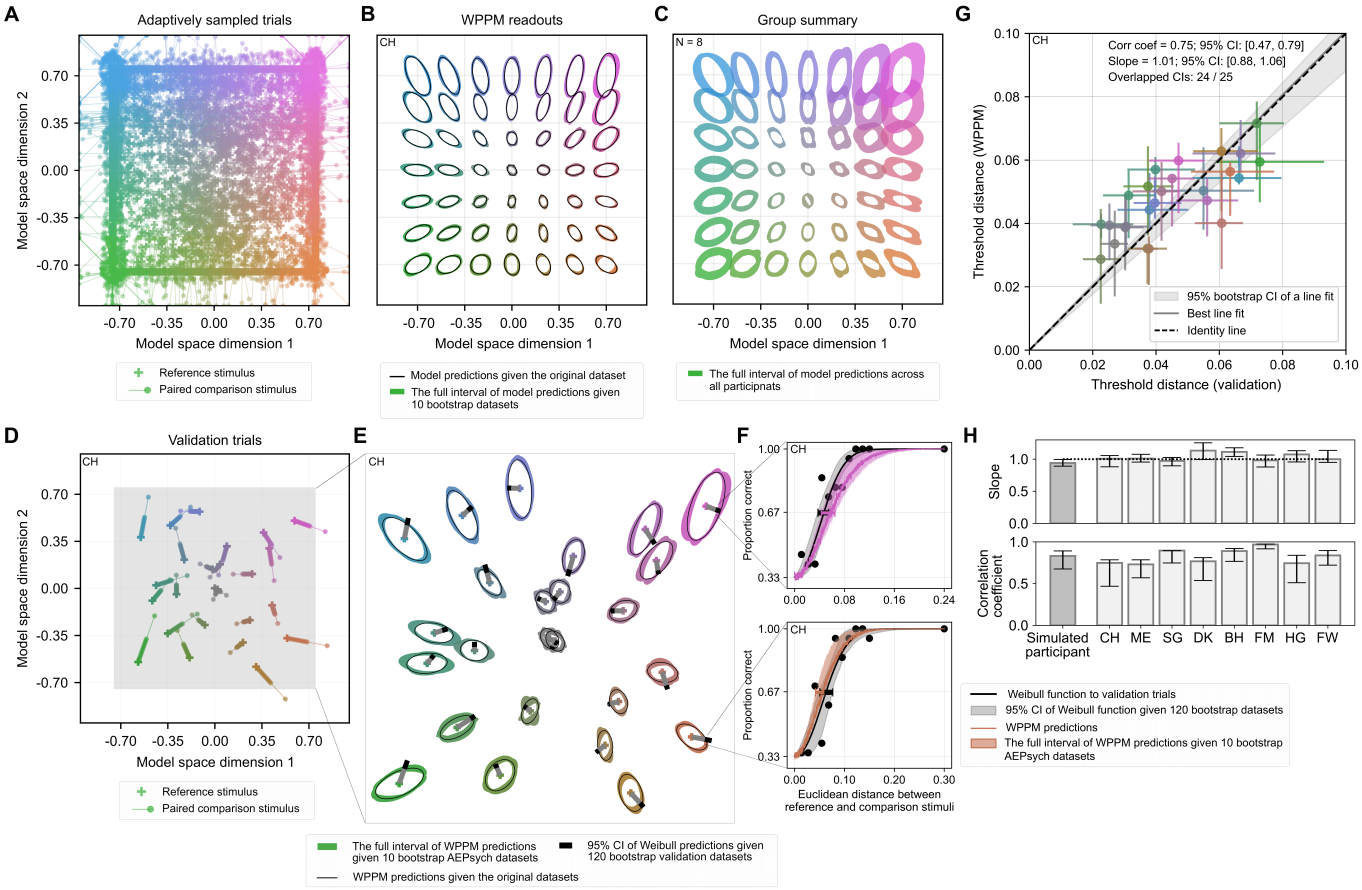

Figure 2. Threshold results and validation. (A) Adaptively sampled trials. AEPsych-driven stimulus pairs that were most informative for estimating thresholds across the entire psychometric field. Of the 6,000 trials, the first 900 were Sobol'-sampled; the remaining 5,100 (shown) were adaptively selected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

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 2***B for a representative participant). The threshold contours revealed three key regularities: (1) thresholds were lowest for references near the achromatic point defined by the background behind the blobby stimuli, (2) thresholds increased with the distance of the reference from the achromatic point, and (3) the major axes of the elliptical threshold contours were consistently radially oriented with respect to the achromatic point. These regularities were consistent with previous results in the color discrimination literature, as explained further in Comparison with previous measurements.

The data were notably consistent across participants, with all three regularities observed in every individual (***Figure 2***C; Appendix 2 for individual plots for the remaining participants). When the variability across participants was assessed as Euclidean distance in the model space representation, it was lowest near the achromatic point where sensitivity was highest and it increased with threshold magnitude. This trend has been observed in other perceptual discrimination tasks (***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 (***Figure 2***D): 11 were evenly spaced, and one was selected to provide easily discriminable catch trials (Appendix 4.3). The comparison levels were selected based on a pilot dataset to account for variability in thresholds across different reference stimuli and chromatic directions (see 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 2***F). We then used the WPPM fit (constrained using the non-overlapping set of adaptively-sampled trials) to extract elliptical contours corresponding to the 66.7% threshold level for each validation reference (***Figure 2***E). Finally, to directly compare these two methods, we read out the WPPM threshold along the MOCS chromatic direction for each reference. The confidence intervals of the WPPM estimates overlapped with those from the Weibull fits in 24 out of 25 conditions for participant CH (***Figure 2***E) and in 21 to 25 conditions across other participants (Appendix 4.1). These results demonstrate strong agreement between thresholds derived from the WPPM psychometric field and the 25 discrete MOCS validation thresholds. This strong agreement indicates that the weak Wishart process prior we imposed to favor smoothly varying threshold contours did not lead to substantial over-smoothing, as the validation thresholds were estimated independently without any such constraint.

To quantify the agreement, we performed a linear (slope only) regression between the thresholds read out from the WPPM fit and those obtained using the validation trials (***Figure 2***G). The results further support agreement between the two sets of estimates (mean correlation coefficient = 0.82, range = 0.73 - 0.97). For 7 out of 8 participants, the regression slope was not significantly different from 1 (mean slope = 1.04, range = 0.97 - 1.13), indicating no systematic bias in the WPPM fits with respect to scale (***Figure 2***H; see Appendix 4.1 for individual participant fits). To assess whether there were more subtle sources of bias not captured by the regression slope, we analyzed the residuals—the discrepancies between the WPPM and validation thresholds. While we found no evidence that residuals depended on the orientation or shape of threshold contours read out from the WPPM fit, we did observe one small but statistically significant relation: the model slightly overestimated thresholds when validation thresholds were low and underestimated them when validation thresholds were high (slope = -0.151, $t(198) = -4.632$, $p < 0.001$, $R^2 = 0.098$). However, the magnitude of this bias in residuals was small (Appendix 4.2).

As an additional benchmark, we simulated trials and responses based on a ground-truth WPPM instance chosen to approximate the predictions from the CIELab ΔE 94 metric (Appendix 5.1 - Appendix 5.5). Using this simulated dataset, we repeated the same validation analyses described above. The thresholds read out from the WPPM fit agreed with 23 out of 25 validation thresholds, based on overlapping confidence intervals. A linear regression revealed a slope of 0.94 and a correlation coefficient of 0.83—well within the range observed in human data (***Figure 2***H, left bar). Residual analysis showed a negative correlation between the residuals and the magnitude of the

ground-truth validation thresholds (Appendix 5.6), consistent with trends observed in human participants. Finally, we assessed the accuracy of the WPPM against simulation ground truth, across a dense grid of reference stimuli in the model space. Although there were some deviations (Appendix 5.7), they were small relative to the inter-participant variation in thresholds (compare ***Figure 2***C and ***Figure S15***).

Taken together, these results validate the accuracy of the WPPM and highlight the remarkable efficiency of our approach. With 6000 trials, conventional psychophysical methods only allowed us to estimate percent-correct performance along one chromatic direction across 25 references. In contrast, our new approach—combining non-parametric, adaptive trial placement with *post hoc* fitting of the semi-parametric WPPM—allowed us to map the entire psychometric field, providing the percent-correct performance for any reference-comparison stimulus pair in the isoluminant plane, using the same number of trials.

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, with the first three shown in the colorimetric representations used in the original studies.

We first compared the overall pattern of threshold variation in the isoluminant plane with measurements made by MacAdam using the method of adjustment (***MacAdam, 1942***) (see Appendix 6 for details). In his seminal work, the ellipses do not represent discrimination thresholds *per se*, but rather the standard deviation of color matches for each reference stimulus. Nevertheless, we consider his measurements to be comparable to ours, based on the linking assumption that discrimination thresholds are proportional to the internal noise that governs the variability of the appearance-based matches (***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 3***A; MacAdam ellipses magnified by $10\times$ while ours magnified by $2\times$). 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 Karl, 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 trials enabled us to characterize discrimination performance across all in-gamut reference–comparison pairs.

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 minor methodological differences, our results replicated the overall pattern of variation in ellipse orientation and size across the color space. In particular, thresholds were smallest near the adapting point, increased with distance from it, and the ellipses generally pointed towards the adapting point. Given the sim-

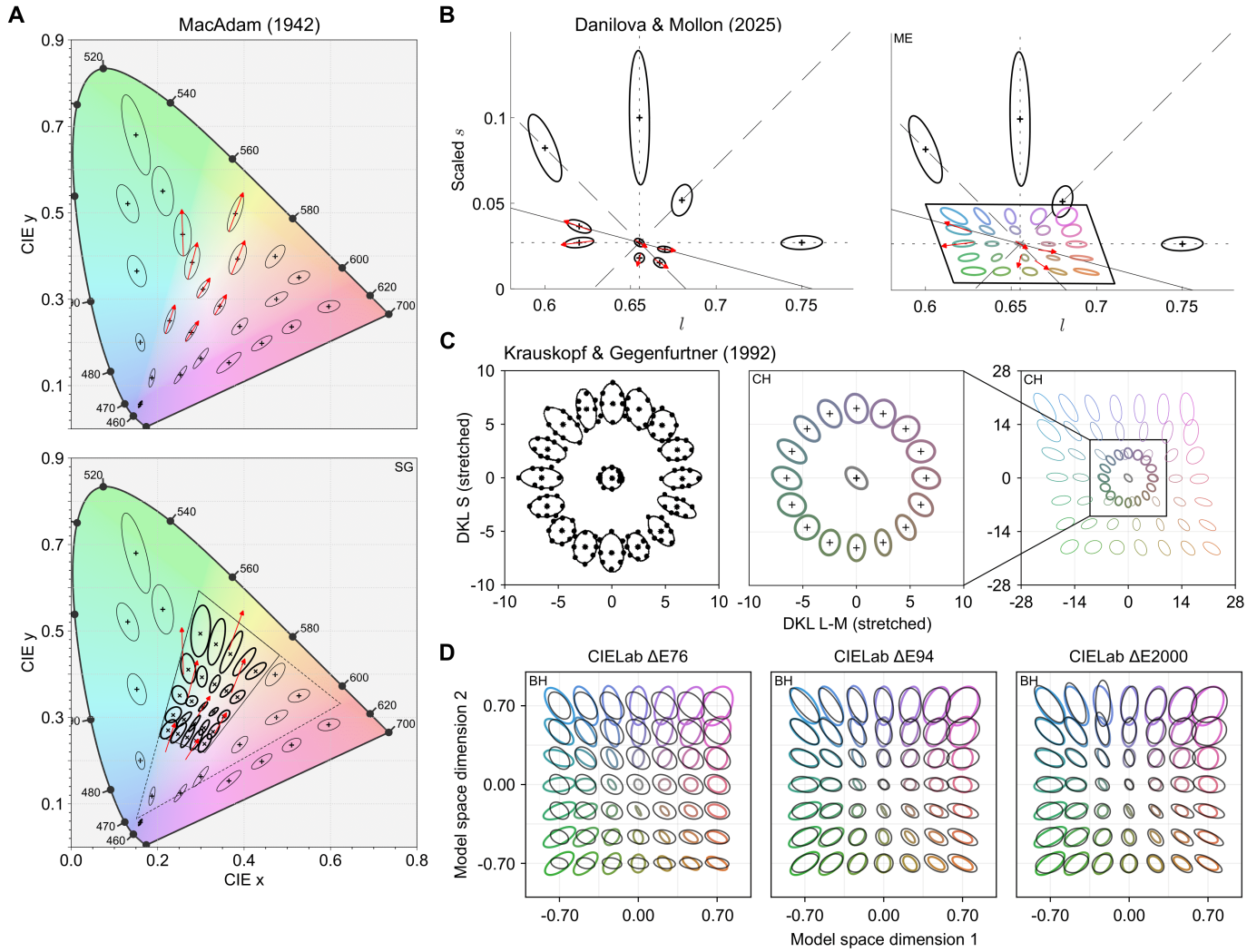

Figure 3. Comparison of color discrimination thresholds with previous measurements. (A) ***MacAdam 1942***. Top: MacAdam's original threshold contours, magnified by $10\times$ for visualization. Bottom: Threshold contours from one participant in our study, transformed from the model space into CIE 1931 chromaticity space. reference stimuli were sampled from a 5×5 grid evenly spaced from –0.75 to 0.75 along each dimension of the model space. To reduce visual clutter, MacAdam ellipses that fall within the gamut of our isoluminant plane (parallelogram) are shown as red arrows indicating only their major axes. For visual comparability, our ellipses are magnified $2\times$ to approximate the size of those in MacAdam's data. Triangle: gamut of our monitor. (B) ***Danilova and Mollon 2025***. Left: Original threshold contours (79.4% correct) from their study, magnified by $4\times$. Right: Threshold contours from one participant in our study (colored ellipses), transformed from the model space into a scaled MacLeod–Boynton space. Reference points were sampled on a 5×5 grid ranging from –0.7 to 0.7. As in (A), to reduce visual clutter, their ellipses that fall within the gamut of our isoluminant plane (parallelogram) are shown as red arrows indicating only their major axes. For visual comparability, our ellipses are magnified by $1.5\times$. (C) ***Krauskopf and Karl 1992***. Left: Original threshold contours (79.4% correct converged by a three-down-one-up staircase) from their study (Fig. 14 from this study, reproduced under Creative Commons CC BY-NC-ND 4.0). Right: Threshold contours from one participant in our study, transformed into a stretched DKL space. All contours are shown at their original sizes. (D) CIELab ΔE 76, ΔE 94, and ΔE 2000. ΔE values were converted into percent correct using a Weibull psychometric function, and threshold is defined as the $\Delta E = 2.5$. Colored lines represent the measured thresholds from one participant, shown at their original sizes. For visual comparability, the predicted threshold contours from each CIELab metric (black lines) were magnified by factors of 5, 2.5, and 2.5, for ΔE 76, ΔE 94, ΔE 2000 metrics, respectively, to approximately match the scale of the measured thresholds in our study. See Appendix 6 - Appendix 9 for additional details.

ilarity in experimental design, it is not surprising that we observe closer agreement in the absolute sizes of the threshold contours than in comparison to MacAdam's data (***Figure 3***B; their ellipses were magnified by $4\times$, ours by $1.5\times$).

In the next comparison, we turned to the study by ***Krauskopf and Karl 1992***, whose measure-

ments were concentrated within a small region near the achromatic point. Their experiment used a fixed adapting point and a 4AFC oddity task, with individual thresholds estimated using a three-down-one-up staircase procedure. To enable direct comparison, we read out threshold contours from our model at the same set of reference stimuli they used. Their results revealed two key features: (1) the threshold contour was smallest at the adapting point, and (2) as the reference moved away from it, the contours became increasingly elongated along the axis pointing toward the adapting point. Both features were observed in our data albeit with some inter-participant variability (***Figure 3***C; Appendix 8). However, because their measurements were limited to reference stimuli near the adapting point, their data did not reveal how thresholds vary with increasing distance from it. In contrast, our results demonstrate that these patterns extend across a much broader region of the isoluminant plane.

Lastly, we compared our results with iso-distance contours obtained with different versions of the CIELab ΔE color difference metrics (***Colorimetry, 2004***). For these comparisons, we plotted the CIELab contours in our model space. Although ΔE metrics were developed to describe supra-threshold perceptual color differences and are based on data integrated from multiple sources, comparisons with threshold-level measurements remain of interest—particularly 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 ΔE 76, which remains widely used, bear little resemblance to our threshold contours (***Figure 3***D, left panel; Appendix 9). The large deviations we observed between ΔE 76 and our data provide further caution against the practice of using ΔE 76 to predict perceptual color difference. In contrast, the more recent ΔE 94 and ΔE 2000 metrics provided a much closer match (***Figure 3***D, center and right), with only modest deviations from our measurements. These deviations may arise from differences between threshold and supra-threshold perceptual judgments, as well as from discrepancies in experimental conditions between our study and those used to constrain the parameters of the CIELab metrics. An important feature of our data is that it enables such comparison with any perceptual metric across the isoluminant plane.

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 Karl, 1992***; ***Danilova and Mollon, 2025***). Moreover, our measurements provide a more comprehensive characterization,

in that the WPPM allows direct readout of threshold contour at any reference stimulus without the need for additional measurement. Additionally, for studies examining how thresholds vary with factors such as stimulus size, presentation duration, or adaptation state, our approach offers a scalable and data-efficient approach for measuring how these factors affect the psychometric field. Finally, we have performed simulations that indicate it will be feasible to fully characterize the color psychometric discrimination field across the three-dimensional gamut of our display, a goal that was has been previously described as "hopelessly difficult" (***Schrödinger, 1920***).

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 Karl, 1992***). In addition, threshold contours were oriented toward the achromatic center, in agreement with previous findings (***Krauskopf and Karl, 1992***; ***Gegenfurtner, 2025***; ***Danilova and Mollon, 2025***). The observation that the size and orientation of the elliptical threshold contours vary with the reference stimulus rules out mechanistic models that posit a linear transformation of cone excitations into three post-receptoral channels followed by fixed additive noise. Such models predict identical ellipses across the space. Moreover, the observation that the orientation of the elliptical threshold contours changes across reference stimuli also rules out mechanistic models in which a linear transformation of cone excitations is followed by limiting noise applied independently to each of the three channels. These models allow variation in size and shape but still predict identical principal axes for all threshold contours.

Cone-opponent models that posit noise and nonlinearities at multiple stages of processing, possibly with an overcomplete cone-opponent representation, may be able to account for the observed data, as may models that invoke higher-order mechanisms (e.g. mechanisms narrowly tuned for hue). For more on relevant ideas see ***Wyszecki 1982***; ***Wandell 1995***; ***Chen et al. 2000***; ***Eskew Jr 2009***; ***Stockman et al. 2010***; ***Hansen and Gegenfurtner 2013***; ***Shevell and Martin 2017***. Although mechanistic models are often tested using additional manipulations such as adaptation and noise-masking, our dataset provides a new and comprehensive benchmark for evaluating and distinguishing competing accounts. In future work, our approach could be extended by incorporating such manipulations to further differentiating those models.

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, however, 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 *Riemannian hypothesis* requires knowledge of how internal noise (or threshold) varies across color space, as this determines the geodesics. Our measurements provide the necessary knowledge for the isoluminant plane, enabling direct empirical tests of the Riemannian hypothesis within this slice of color space, as well as elaborations of this hypothesis (***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 Riemannian hypothesis in this plane based on our current data would be considered provisional. Nonetheless, such tests would provide valuable exploration of the perceptual geometry revealed by our measurements. As noted above, our approach makes it feasible to extend the measurements to the full three-dimensional color space, which, when completed, will allow subsequent investigations to overcome this limitation.

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

Individual differences and implications

We studied a relatively young cohort of eight participants and found good consistency across individuals, with only modest individual differences. Such differences have, in general, provided useful insights into the neural mechanisms of color vision and are also of interest for understanding how much any given person may differ from an average characterization. To enable studies involving larger and more diverse populations, further improvements in the efficiency of our approach are likely feasible. In the present study, we deliberately used a non-parametric adaptive trial-placement procedure to avoid biasing data collection by assuming the accuracy of the WPPM. Now that the WPPM has been validated, future studies could incorporate adaptive trial-placement strategies tailored to this model or develop priors informed by the current dataset to further improve efficiency. Thus, our results have the potential to support investigations into how variation in the psychometric field of color discrimination relates to underlying biological factors such as pre-retinal absorption, photopigment spectral sensitivity, the ratio of L to M cones in the mosaic, and other color vision measures such as unique hues. Individual differences in these factors have been extensively studied (***Neitz and Jacobs, 1986***; ***Webster and MacLeod, 1988***; ***Brainard et al., 2000***; ***Kremers et al., 2000***; ***Carroll et al., 2002***; ***Hofer et al., 2005***; ***Bosten, 2022***; ***Emery et al., 2023***; ***Rezeanu et al., 2023***).

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 prodcedure 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 finalizing the analysis plans for other participants.

Participants

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

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 (Fig. **S25**A) was constructed in Unity (v2022.3.24f1) and rendered using its standard shader. The scene consisted of three identical blobby 3D objects, each created in Blender (v4.0) with a matte, non-reflective surface. On each trial, the surface color of the blobby objects was varied by adjusting their RGB values in Unity. The three blobby objects (2.5×2.5 dva; 203,900 pixels each) were arranged in a triangular configuration (***Figure 1***A). 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 10.1). These measurements directly characterized the relationship between the specified RGB values for the blobby stimuli and the light emitted from the display. The same calibration was repeated for all three blobby stimuli, confirming consistent color behavior across screen locations. Based on these results, a single gamma correction—derived from the bottom-right stimulus—was applied to all three objects during the experiment. This correction was validated by remeasuring the output with gamma correction applied, showing good alignment with the predicted identity line. To confirm stability over time, we repeated the calibration one month into data collection and observed negligible changes.

Additionally, we used a Klein K-10A colorimeter to verify that the system achieved sufficient color depth. For this check, a single blobby stimulus was presented at the center of the screen, rather than in the full triangular arrangement. Measurements confirmed that Unity and our video chain were able to produce at least 10-bit color precision via its native 8-bit output and implicit spatial dithering that occurred across the surface of the blobby object through the rendering process (Appendix 10.2).

Design

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

We used AEPsych (v0.7) to sample a total of 6,000 reference–comparison stimulus pairs. The first 900 trials were generated using Sobol' sampling (***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 and 0.75 along each model dimension. Each element of Δ was constrained between –0.25 and 0.25 to ensure that all comparison stimuli remained within the $[-1, 1]^2$ bounds of the model space. During the initial 900 Sobol'-sampled trials, the difference vector Δ was scaled by one of three factors (1/4, 2/4, or 3/4) before being added to the reference stimulus. This controlled the distance between the reference and comparison stimuli, effectively modulating task difficulty. These scaling factors were evenly distributed and pseudo-randomized across trials. For the remaining 5,100 trials, all four variables were adaptively selected using AEPsych's optimization procedure. Specifically, the underlying GP model was updated every 20 trials, and new trials were selected using the Expected Absolute Volume Change (EAVC) acquisition function (***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 between –0.6 and 0.6 along each model dimension. For each reference, a chromatic direction was Sobol'-sampled between 0° and 360°. Each validation condition consisted of 12 stimulus levels: 11 equally spaced along the sampled direction and one easily discriminable level, with each level repeated 20 times. These levels were determined based on a pilot dataset described in the preregistration documents.

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

Procedure

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

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

The Wishart Process Psychometric Model

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

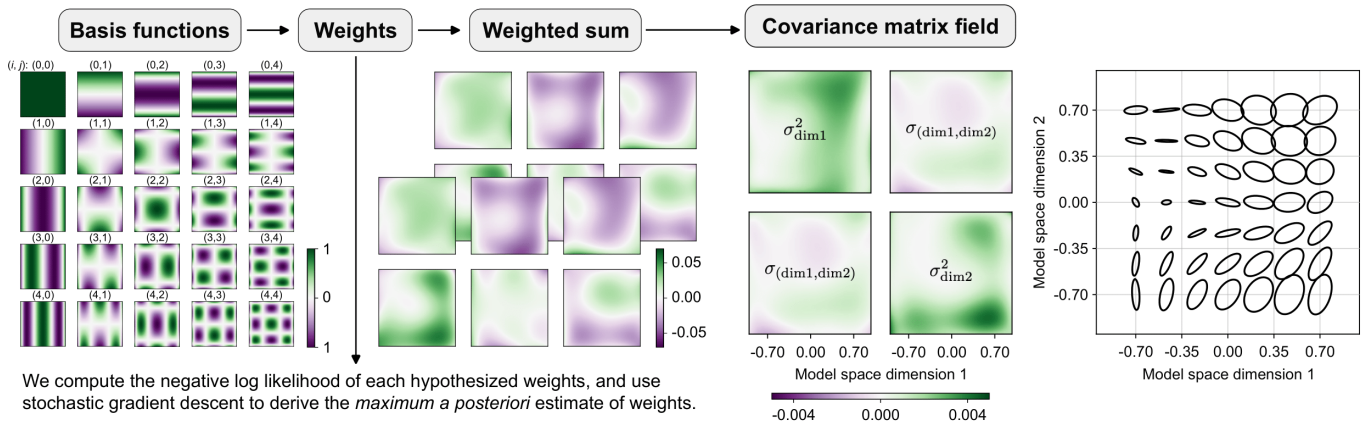

Figure 4. The finite-basis Wishart Process Psychophysical Model (WPPM). In our implementation, we used a set of 5×5 two-dimensional Chebyshev polynomial basis functions, denoted $\phi_{i,j}(x)$, where $i, j \in \{0, 1, \ldots, 4\}$. These basis functions were linearly combined using a learnable weight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

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,

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

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

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

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

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

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

$$d^2_M(z_0, z_0') = (z_0 - z_0')^\top \mathbf{S}^{-1}(z_0 - z_0') \tag{5}$$

$$d^2_M(z_0, z_1) = (z_0 - z_1)^\top \mathbf{S}^{-1}(z_0 - z_1) \tag{6}$$

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

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

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

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 1***E) 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 (σ^2_{dim1}, σ^2_{dim2}) and their covariance ($\sigma_{\text{dim1,dim2}}$) (***Figure 1***C-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 $x = [x_{\text{dim1}}, x_{\text{dim2}}]$ denote a location in the 2D model space. The basis functions are defined recursively for each model space dimension as given here for x_{dim1}:

$$T_0(x_{\text{dim1}}) = 1, \tag{9}$$

$$T_1(x_{\text{dim1}}) = x_{\text{dim1}}, \tag{10}$$

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

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

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

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., $i, j \in \{0, 1, \ldots, 4\}$, 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, $\mathbf{W} \in \mathbb{R}^{5 \times 5 \times 2 \times 3}$, where the first two dimensions index the Chebyshev basis functions along each model space dimension ($i, j \in \{0, 1, \ldots, 4\}$), and the last two dimensions index the output components ($k \in \{1, 2\}$ and $l = \{1, 2, 3\}$). The weighted basis functions are expanded into an overcomplete representation $\mathbf{U}_{k,l} \in \mathbb{R}^{n \times m}$ (***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, $\Sigma(x) \in \mathbb{R}^{2 \times 2}$ for x at any discretized point in the model space, that is,

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

This process ensures that the resulting covariance matrices are symmetric and positive semi-definite (***Figure 4***, third panel). The weight matrix serves as the free parameters of the model, controlling the smoothness of the covariance matrix field. The model is highly flexible, capable of generating a wide range of covariance matrix fields, from smooth to rapidly varying fields (***Figure 1***C-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 scalar γ controls the overall amplitude of the variance and was fixed at 3×10^{-4}. The scalar ϵ controls the rate at which the prior variance decays with increasing polynomial order, and was fixed at 0.5. A higher value of ϵ results in a prior that favors more sharply varying covariance matrix fields, while a lower value favors smoother fields. By setting $\epsilon = 0.5$, we adopted a prior that favors relatively smooth variation across the space (***Figure 4***, fourth panel). This value was chosen by hand based on examination of a pilot dataset and analysis of the data reported here for participant CH, available at a public repository. As preregistered, this decision was made before analyzing the validation trials from the remaining seven participants.

Model fitting

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

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

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:

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

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

Since we imposed a prior on the covariance matrix field to reflect the expectation of smooth variation, we combined the likelihood (***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 11).

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 1***F). 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 and derived transformation matrices to convert values from the model space to RGB space (Appendix 10). Stimulus presentation, including gamma correction, was implemented in Unity using C#.

The experiment and model fitting were conducted in Python 3.11 and JAX (***Bradbury et al., 2018***). Behavioral data were separated into AEPsych-driven plus fallback trials on the one hand and validation trials on the other. The WPPM was fit exclusively to the AEPsych and fallback trials. To assess variability in model estimates, we performed 10 bootstrap resamplings of the AEPsych-driven trials, preserving the original ratio between Sobol', adaptively sampled, and fallback trials in each resampled dataset. The WPPM was then re-fit to each of the 10 bootstrapped datasets.

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

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

Data and code availability

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

Acknowledments

We thank Nicolas P. Cottaris for his assistance with the calibration, our colleagues at the UPenn Vision Labs, and Larry Maloney from NYU for their valuable feedback.

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.

References

Aguilar G, Wichmann FA, Maertens M. Comparing sensitivity estimates from MLDS and forced-choice methods in a slant-from-texture experiment. Journal of Vision. 2017; 17(1):37–37.

Ashby FG, Soto FA. Multidimensional signal detection theory. Oxford handbook of computational and mathematical psychology. 2015; p. 13–34.

Aspinall PA, Kinnear PR, Duncan LJ, Clarke BF. Prediction of diabetic retinopathy from clinical variables and color vision data. Diabetes Care. 1983; 6(2):144–8. https://www.ncbi.nlm.nih.gov/pubmed/6851807, doi: 10.2337/diacare.6.2.144.

Bertonati G, Amadeo MB, Campus C, Gori M. Auditory speed processing in sighted and blind individuals. Plos one. 2021; 16(9):e0257676.

Bhatia R, Jain T, Lim Y. On the Bures–Wasserstein distance between positive definite matrices. Expositiones Mathematicae. 2019; 37(2):165–191.

Bosten JM. Do you see what I see? Diversity in human color perception. Annual review of vision science. 2022; 8(1):101–133.

Bradbury J, Frostig R, Hawkins P, Johnson MJ, Leary C, Maclaurin D, Necula G, Paszke A, VanderPlas J, Wanderman-Milne S, Zhang Q, JAX: composable transformations of Python+NumPy programs; 2018. http://github.com/jax-ml/jax.

Brainard DH. Color Appearance and Color Difference Specification. In: Shevell SK, editor. *The Science of Color*, 2nd ed. Elsevier; 2003.p. 191–216. doi: 10.1016/B978-044451251-2/50006-4.

Brainard DH, Pelli DG, Robson T. Display characterization. Signal Process. 2002; 80:2–067.

Brainard DH, Roorda A, Yamauchi Y, Calderone JB, Metha A, Neitz M, Neitz J, Williams DR, Jacobs GH. Functional consequences of the relative numbers of L and M cones. Journal of the Optical Society of America A. 2000; 17(3):607–614.

Brainard DH, Stockman A. Colorimetry. In: McGraw Hill; 2010.

Brainard D. Cone contrast and opponent modulation color spaces. Human color vision. 1996; .

Brown WRJ, MacAdam DL. Visual sensitivities to combined chromaticity and luminance differences. Journal of the Optical Society of America. 1949; 39(10):808–834.

Brown W. The effect of field size and chromatic surroundings on color discrimination. Journal of the Optical Society of America. 1952; 42(11):837–844.

Bujack R, Teti E, Miller J, Caffrey E, Turton TL. The non-Riemannian nature of perceptual color space. Proceedings of the National Academy of Sciences. 2022; 119(18):e2119753119.

Campbell FW, Robson JG. Application of Fourier analysis to the visibility of gratings. The Journal of physiology. 1968; 197(3):551.

Carlile S, Leung J. The perception of auditory motion. Trends in hearing. 2016; 20:2331216516644254.

Carroll J, Neitz J, Neitz M. Estimates of L: M cone ratio from ERG flicker photometry and genetics. Journal of vision. 2002; 2(8):1–1.

Champion RA, Freeman TC. Discrimination contours for the perception of head-centered velocity. Journal of Vision. 2010; 10(6):14–14.

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

Chen CC, Foley JM, Brainard DH. Detection of chromoluminance patterns on chromoluminance pedestals II: model. Vision Research. 2000; 40(7):789–803.

Churchland PM. Some reductive strategies in cognitive neurobiology. Mind. 1986; 95(379):279–309.

Cicchini GM, Anobile G, Burr DC. Spontaneous perception of numerosity in humans. Nature communications. 2016; 7(1):12536.

Cicchini GM, Anobile G, Burr DC. Spontaneous representation of numerosity in typical and dyscalculic development. Cortex. 2019; 114:151–163.

Cicchini GM, Anobile G, Burr DC, Marchesini P, Arrighi R. The role of non-numerical information in the perception of temporal numerosity. Frontiers in Psychology. 2023; 14:1197064.

Colorimetry C, Commission Internationale de l'Èclairage: Vienna. Austria; 2004.

Craik K. The effect of adaptation on differential brightness discrimination. The Journal of Physiology. 1938; 92(4):406.

Crozier WJ, Holway AH. On the law for minimal discrimination of intensities: I. Proceedings of the National Academy of Sciences. 1937; 23(1):23–28.

Danilova M, Mollon J. Effect of stimulus size on chromatic discrimination. Journal of the Optical Society of America A. 2025; 42(5):B167–B177.

Derrington AM, Krauskopf J, Lennie P. Chromatic mechanisms in lateral geniculate nucleus of macaque. The Journal of physiology. 1984; 357(1):241–265.

Dvoretzky A, Kiefer J, Wolfowitz J. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. The Annals of Mathematical Statistics. 1956; p. 642–669.

Emery KJ, Volbrecht VJ, Peterzell DH, Webster MA. Fundamentally different representations of color and motion revealed by individual differences in perceptual scaling. Proceedings of the National Academy of Sciences. 2023; 120(4):e2202262120.

Ennis DM, Mullen K. A general probabilistic model for triad discrimination, preferential choice, and two-alternative identification. In: *Multidimensional models of perception and cognition* Psychology Press; 2014.p. 115–122.

Ennis RJ, Zaidi Q. Geometrical structure of perceptual color space: Mental representations and adaptation invariance. Journal of vision. 2019; 19(12):1–1.

Eskew Jr RT. Higher order color mechanisms: A critical review. Vision research. 2009; 49(22):2686–2704.

Fechner GT. Elemente der psychophysik, vol. 2. Breitkopf u. Härtel; 1860.

Foley JM, Legge GE. Contrast detection and near-threshold discrimination in human vision. Vision research. 1981; 21(7):1041–1053.

Freeman TC, Leung J, Wufong E, Orchard-Mills E, Carlile S, Alais D. Discrimination contours for moving sounds reveal duration and distance cues dominate auditory speed perception. PloS one. 2014; 9(7):e102864.

Garside DJ, Chang AL, Selwyn HM, Conway BR. The origin of color categories. Proceedings of the National Academy of Sciences. 2025; 122(1):e2400273121.

Gegenfurtner KR. The Verriest Lecture: Color vision from pixels to objects. Journal of the Optical Society of America A. 2025; 42(5):B313–B328.

Girshick AR, Landy MS, Simoncelli EP. Cardinal rules: visual orientation perception reflects knowledge of environmental statistics. Nature neuroscience. 2011; 14(7):926–932.

Green DM, Swets JA, et al. Signal detection theory and psychophysics, vol. 1. Wiley New York; 1966.

Hansen T, Gegenfurtner KR. Higher order color mechanisms: Evidence from noise-masking experiments in cone contrast space. Journal of vision. 2013; 13(1):26–26.

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

Hecht S, Shlaer S, Pirenne MH. Energy, quanta, and vision. Journal of General Physiology. 1942; 25(6):819–840.

Hedjar L, Toscani M, Gegenfurtner KR. Importance of hue: color discrimination of three-dimensional objects and two-dimensional discs. Journal of the Optical Society of America A. 2025; 42(5):B296–B304.

Hillis JM, Brainard DH. Distinct mechanisms mediate visual detection and identification. Current Biology. 2007; 17(19):1714–1719.

Hofer H, Carroll J, Neitz J, Neitz M, Williams DR. Organization of the human trichromatic cone mosaic. Journal of Neuroscience. 2005; 25(42):9669–9679.

Hong F, Badde S, Landy MS. Causal inference regulates audiovisual spatial recalibration via its influence on audiovisual perception. PLOS Computational Biology. 2021 11; 17(11):1–37. https://doi.org/10.1371/journal.pcbi.1008877, doi: 10.1371/journal.pcbi.1008877.

Horiuchi S, Nagai T. Color discrimination repetition distorts color representations. Scientific Reports. 2024; 14(1):9615.

Hurvich LM, Hurvich-Jameson D. Opponent chromatic induction and wavelength discrimination. Springer; 1961.

Johnson CA, Wall M, Thompson HS. A history of perimetry and visual field testing. Optometry and Vision Science. 2011; 88(1):E8–E15.

Knoblauch K, Maloney LT. Testing the indeterminacy of linear color mechanisms from color discrimination data. Vision research. 1996; 36(2):295–306.

Knoblauch K, Maloney LT. Modeling psychophysical data in R, vol. 32. Springer Science & Business Media; 2012.

Koenderink JJ. Color for the Sciences. MIT press; 2010.

Krauskopf J, Karl G. Color discrimination and adaptation. Vision research. 1992; 32(11):2165–2175.

Kremers J, Scholl HP, Knau H, Berendschot TT, Usui T, Sharpe LT. L/M cone ratios in human trichromats assessed by psychophysics, electroretinography, and retinal densitometry. Journal of the Optical Society of America A. 2000; 17(3):517–526.

de Lange Dzn H. Research into the dynamic nature of the human fovea$\bigcirc$ cortex systems with intermittent and modulated light. I. Attenuation characteristics with white and colored light. Journal of the Optical Society of America. 1958; 48(11):777–784.

Lesmes LA, Lu ZL, Baek J, Albright TD. Bayesian adaptive estimation of the contrast sensitivity function: the quick CSF method. Journal of vision. 2010; 10(3):17–17.

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

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

MacAdam DL. Visual sensitivities to color differences in daylight. Journal of the Optical Society of America. 1942; 32(5):247–274.

Macadam DL. Judd's contributions to color metrics and evaluation of color differences. Color Research & Application. 1979; 4(4):177–193.

MacLeod DI, Boynton RM. Chromaticity diagram showing cone excitation by stimuli of equal luminance. Journal of the Optical Society of America. 1979; 69(8):1183–1186.

McDonald R, Smith KJ. CIE94-a new colour-difference formula. Journal of the Society of Dyers and Colourists. 1995; 111(12):376–379.

Mullen K, Ennis DM. A simple multivariate probabilistic model for preferential and triadic choices. Psychometrika. 1991; 56(1):69–75.

Najemnik J, Geisler WS. Optimal eye movement strategies in visual search. Nature. 2005; 434(7031):387–391.

Neitz J, Jacobs GH. Polymorphism of the long-wavelength cone in normal human colour vision. Nature. 1986; 323(6089):623–625.

Newton JR, Eskew RT. Chromatic detection and discrimination in the periphery: a postreceptoral loss of color sensitivity. Visual neuroscience. 2003; 20(5):511–521.

Niwa Y, Muraki S, Naito F, Minamikawa T, Ohji M. Evaluation of acquired color vision deficiency in glaucoma using the Rabin cone contrast test. Invest Ophthalmol Vis Sci. 2014; 55(10):6686–90. https://www.ncbi.nlm.nih.gov/pubmed/25168899, doi: 10.1167/iovs.14-14079.

Noorlander C, Heuts MJ, Koenderink JJ. Sensitivity to spatiotemporal combined luminance and chromaticity contrast. Journal of the Optical Society of America. 1981; 71(4):453–459.

Noorlander C, Koenderink JJ, Den Olden RJ, Edens BW. Sensitivity to spatiotemporal colour contrast in the peripheral visual field. Vision Research. 1983; 23(1):1–11.

Olkkonen M, McCarthy PF, Allred SR. The central tendency bias in color perception: Effects of internal and external noise. Journal of vision. 2014; 14(11):5–5.

Owen L, Browder J, Letham B, Stocek G, Tymms C, Shvartsman M. Adaptive nonparametric psychophysics. arXiv preprint arXiv:210409549. 2021; .

Palmer J, Ames CT, Lindsey DT. Measuring the effect of attention on simple visual search. Journal of Experimental Psychology: Human Perception and Performance. 1993; 19(1):108.

Pointer MR. Color discrimination as a function of observer adaptation. Journal of the Optical Society of America. 1974; 64(6):750–759.

Poirson AB, Wandell BA. The ellipsoidal representation of spectral sensitivity. Vision research. 1990; 30(4):647–652.

Prins N, et al. Psychophysics: a practical introduction. Academic Press; 2016.

Reisbeck TE, Gegenfurtner KR. Velocity tuned mechanisms in human motion processing. Vision research. 1999; 39(19):3267–3286.

Rezeanu D, Neitz M, Neitz J. From cones to color vision: a neurobiological model that explains the unique hues. Journal of the Optical Society of America A. 2023; 40(3):A1–A8.

Roberti V. Helmholtz, Schrödinger, and the First Non-Euclidean Model of Perceptual Color Space. Annalen der Physik. 2024; 536(5):2300536.

Robertson AR, Lozano RD, Alman DH, Orchard S, Keitch J, Connely R, Graham L, Acree W, John R, Hoban R, et al. CIE recommendations on uniform color spaces, color-difference equations, and metric color terms. Color Res Appl. 1977; 2(5-6):3.

Schrödinger Ev. Outline of a theory of color measurement for daylight vision. Physics Annual. 1920; 63(4):397–520.

Sharma G, Wu W, Dalal EN. The CIEDE2000 color-difference formula: Implementation notes, supplementary test data, and mathematical observations. Color Research & Application: Endorsed by Inter-Society Color Council, The Colour Group (Great Britain), Canadian Society for Color, Color Science Association of Japan, Dutch Society for the Study of Color, The Swedish Colour Centre Foundation, Colour Society of Australia, Centre Français de la Couleur. 2005; 30(1):21–30.

Shepard TG, Lahlaf SI, Eskew RT. Labeling the lines: A test of a six-mechanism model of chromatic detection. Journal of Vision. 2017; 17(13):9–9.

Shepard TG, Swanson EA, McCarthy CL, Eskew RT. A model of selective masking in chromatic detection. Journal of vision. 2016; 16(9):3–3.

Shevell SK, Martin PR. Color opponency: tutorial. Journal of the Optical Society of America A. 2017; 34(7):1099–1108.

Sobol IM. The distribution of points in a cube and the approximate evaluation of integrals. USSR Computational mathematics and mathematical physics. 1967; 7:86–112.

Stark E, Turton TL, Bujack R. Diminishing Returns in Perceptual Color Space-Now in Color. . 2025; .

Stockman A, Brainard DH, et al. Color vision mechanisms. OSA handbook of optics. 2010; 3:11–1.

Vemala R, Sivaprasad S, Barbur JL. Detection of Early Loss of Color Vision in Age-Related Macular Degeneration - With Emphasis on Drusen and Reticular Pseudodrusen. Invest Ophthalmol Vis Sci. 2017; 58(6):BIO247–BIO254. https://www.ncbi.nlm.nih.gov/pubmed/28846119, doi: 10.1167/iovs.17-21771.

Wandell BA. Color measurement and discrimination. Journal of the Optical Society of America A. 1985; 2(1):62–71.

Wandell BA. Foundations of vision. Sinauer Associates; 1995.

Wardle SG, Alais D. Evidence for speed sensitivity to motion in depth from binocular cues. Journal of Vision. 2013; 13(1):17–17.

Wasserman L. All of nonparametric statistics. Springer Science & Business Media; 2006.

Watson AB. QUEST+: A general multidimensional Bayesian adaptive psychometric method. Journal of Vision. 2017; 17(3):10–10.

Webster MA, MacLeod DI. Factors underlying individual differences in the color matches of normal observers. Journal of the Optical Society of America A. 1988; 5(10):1722–1735.

Williams CK, Rasmussen CE. Gaussian processes for machine learning, vol. 2. MIT press Cambridge, MA; 2006.

Wilson AG, Ghahramani Z. Generalised Wishart processes. In: *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence* UAI'11, Arlington, Virginia, USA: AUAI Press; 2011. p. 736–744.

Winawer J, Witthoft N. Effects of color terms on color perception and cognition. In: *Encyclopedia of color science and technology* Springer; 2023.p. 777–785.

Wuerger SM, Maloney LT, Krauskopf J. Proximity judgments in color space: tests of a Euclidean color geometry. Vision research. 1995; 35(6):827–835.

Wyszecki G. Concepts and methods, quantitative data and formulae. Color Science. 1982; p. 130–175.

Zaidi Q, Is there a perceptual color space? Wiley Online Library; 2001.

Zhou J, Duong LR, Simoncelli EP. A unified framework for perceived magnitude and discriminability of sensory stimuli. Proceedings of the National Academy of Sciences. 2024; 121(25):e2312293121.

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 color 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-deg 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 defined 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 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 $\mathbf{d}_{\text{DKL}} = [\cos(\theta), \sin(\theta), 0]^{\top}$, 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 $\mathbf{d}_{\text{DKL}}$ to LMS cone excitations, denoted $\mathbf{e}_{\text{LMS}}$, as follows:

$$\mathbf{e}_{\text{LMS}} = \boldsymbol{M}_{\text{cone contrast}\to\text{LMS}} \cdot \boldsymbol{M}_{\Delta\text{LMS}\to\text{cone contrast}} \cdot \boldsymbol{M}_{\text{DKL}\to\Delta\text{LMS}} \cdot \mathbf{d}_{\text{DKL}}. \tag{S1}$$

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

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

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

$$\mathbf{d}_{\text{RGB}} = \boldsymbol{M}_{\text{LMS}\to\text{RGB}} \cdot \mathbf{e}_{\text{LMS, stimulus}} - \boldsymbol{M}_{\text{LMS}\to\text{RGB}} \cdot \mathbf{e}_{\text{LMS, background}} \tag{S3}$$

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

Finally, we marched outward along the direction $\mathbf{d}_{\text{RGB}}$ 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	$DKL_{L\text{-}M}$	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 $M_{\text{DKL}\to\text{W}}$ such that for each vertex pair, we have:

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

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

$$M_{\text{DKL}\to\text{W}} = \begin{bmatrix} — & \mathbf{v}_{\text{W, dim1}} & — \\ — & \mathbf{v}_{\text{W, dim2}} & — \\ — & \mathbf{1} & — \end{bmatrix} \cdot \begin{bmatrix} — & \mathbf{v}_{\text{DKL, L-M}} & — \\ — & \mathbf{v}_{\text{DKL, S}} & — \\ — & \mathbf{0} & — \end{bmatrix}^{\dagger}, \tag{S6}$$

where † denotes the pseudoinverse. Note that the last row of $M_{\text{DKL}\to\text{W}}$ 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.

$M_{\text{DKL}\to\text{W}}$	$M_{\text{RGB}\to\text{W}}$
$\begin{bmatrix} 6.724 & 0.213 & 0.000 \\ 0.076 & 1.221 & 0.000 \\ 0.000 & 0.000 & 1.000 \end{bmatrix}$	$\begin{bmatrix} 1.556 & -1.364 & -0.192 \\ -0.444 & -1.364 & 1.808 \\ 0.444 & 1.364 & 0.192 \end{bmatrix}$

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 $M_{\text{RGB}\to\text{W}}$ as:

$$M_{\text{RGB}\to\text{W}} = \begin{bmatrix} — & \mathbf{v}_{\text{W, dim1}} & — \\ — & \mathbf{v}_{\text{W, dim2}} & — \\ — & \mathbf{1} & — \end{bmatrix} \cdot \begin{bmatrix} — & \mathbf{v}_{\text{R}} & — \\ — & \mathbf{v}_{\text{G}} & — \\ — & \mathbf{v}_{\text{B}} & — \end{bmatrix}^{\dagger}, \tag{S7}$$

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:

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

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:

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

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
AEPsych-driven trials and WPPM readouts for all participants

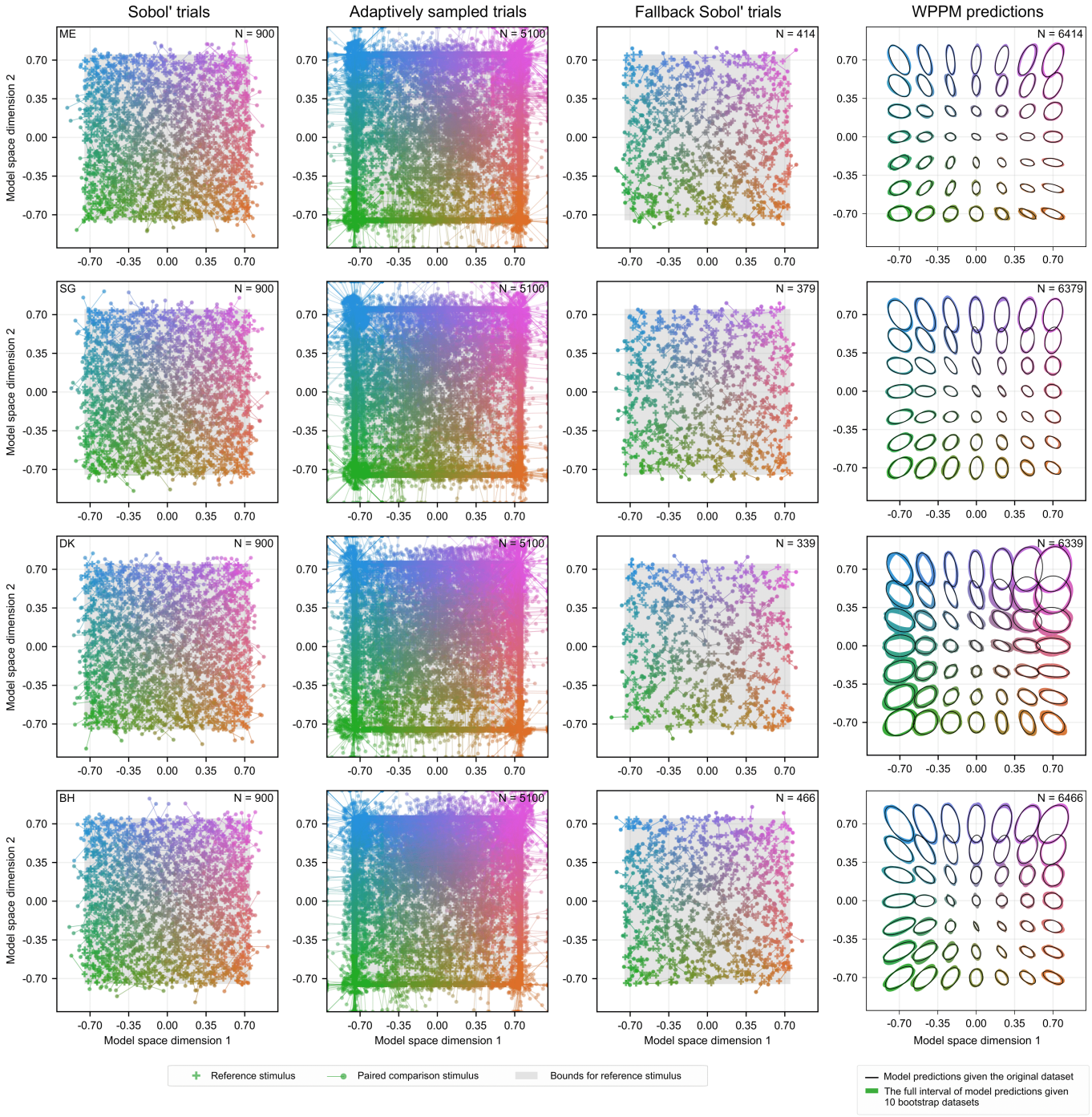

Figure S1. AEPsych-driven trials (900 Sobol'-sampled and 5,100 adaptively sampled), fallback trials, and WPPM predictions for all participants. Each row represents data from one participant.

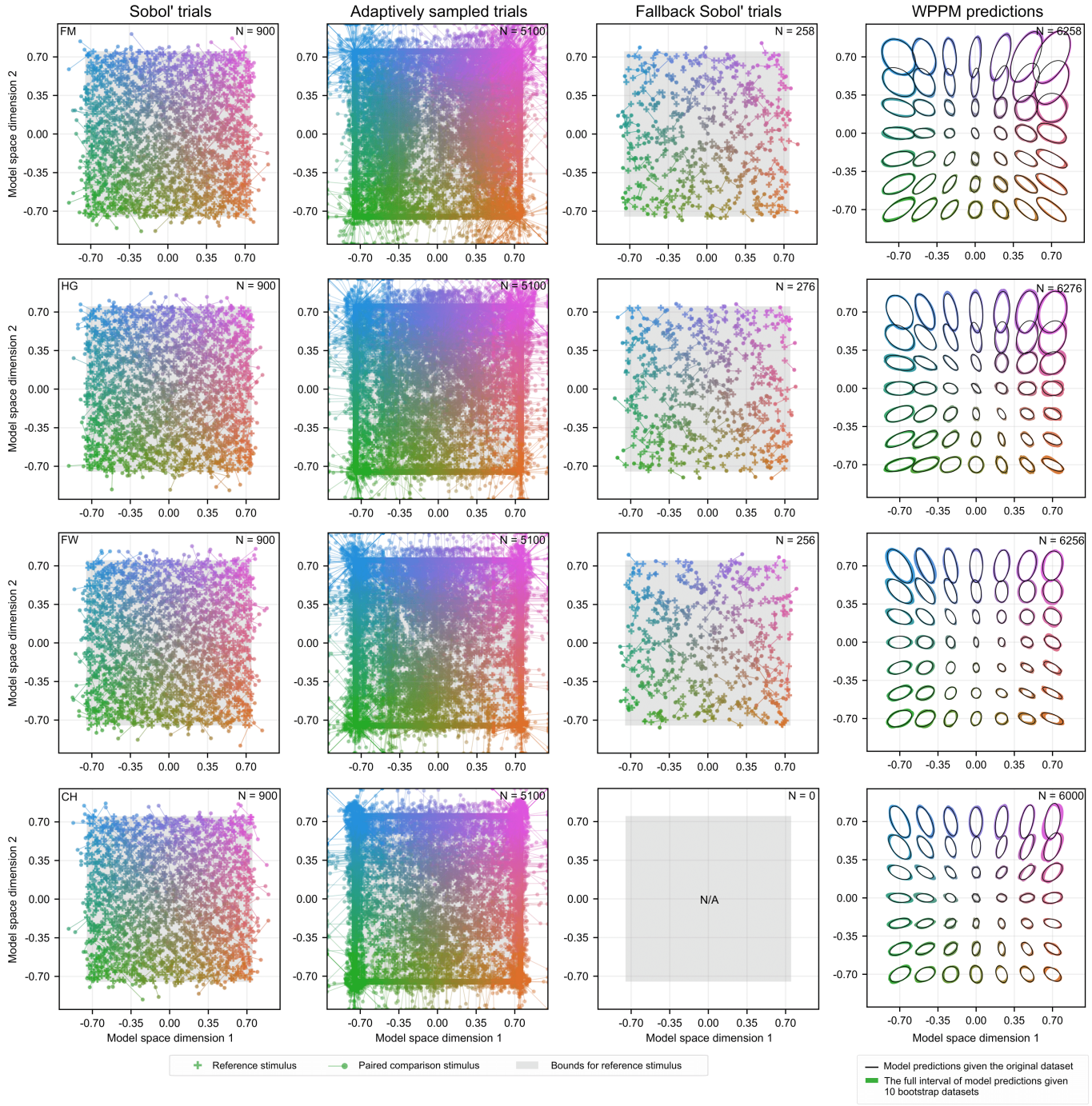

Figure S1. (continued) AEPsych-driven trials (900 Sobol'-sampled and 5,100 adaptively sampled), fallback trials, and WPPM predictions for all participants. Each row represents data from one participant. Note that for participant CH, no pre-generated Sobol' trials were used, as the fallback strategy was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

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

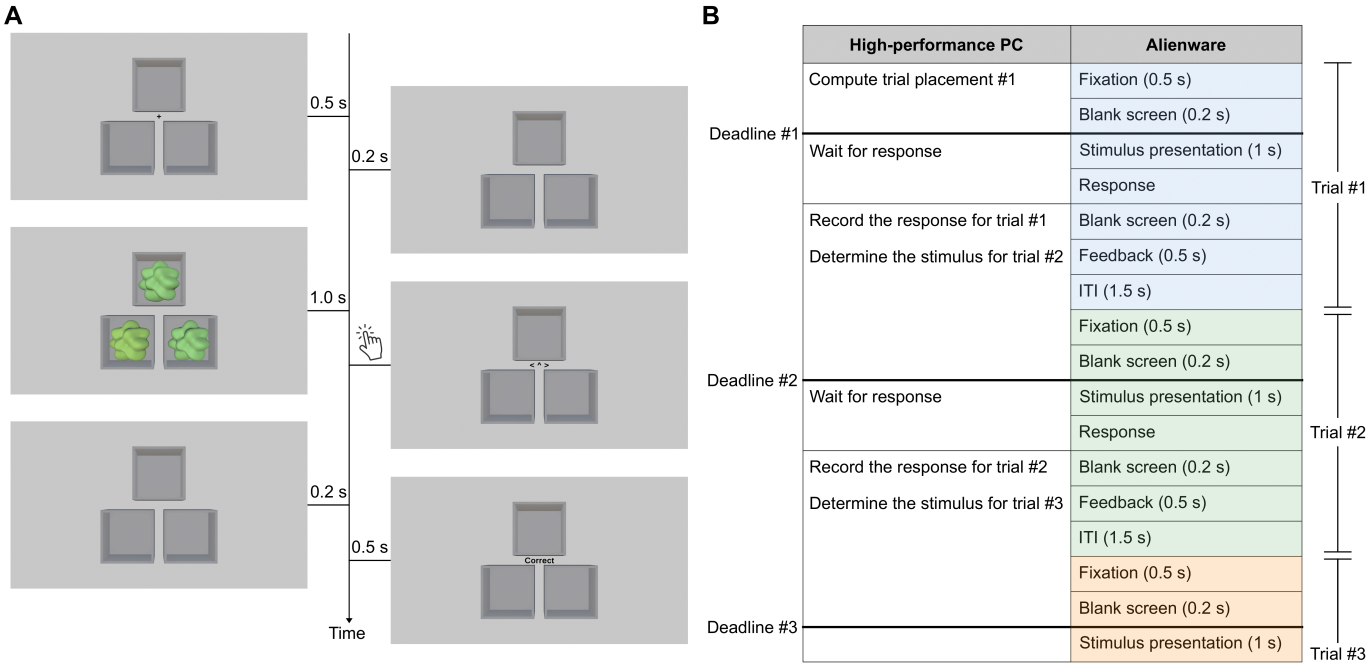

Figure S2. Task timing and real-time trial scheduling. (A) Trial sequence: a 0.5 s fixation cross was followed by a 0.2 s blank interval, then a 1 s presentation of three blobby stimuli. Participants responded at their own pace to identify the odd-one-out, after which a 0.2 s blank screen and a 0.5 s 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 pre-determined 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 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 to handle 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 S2***). 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, included 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

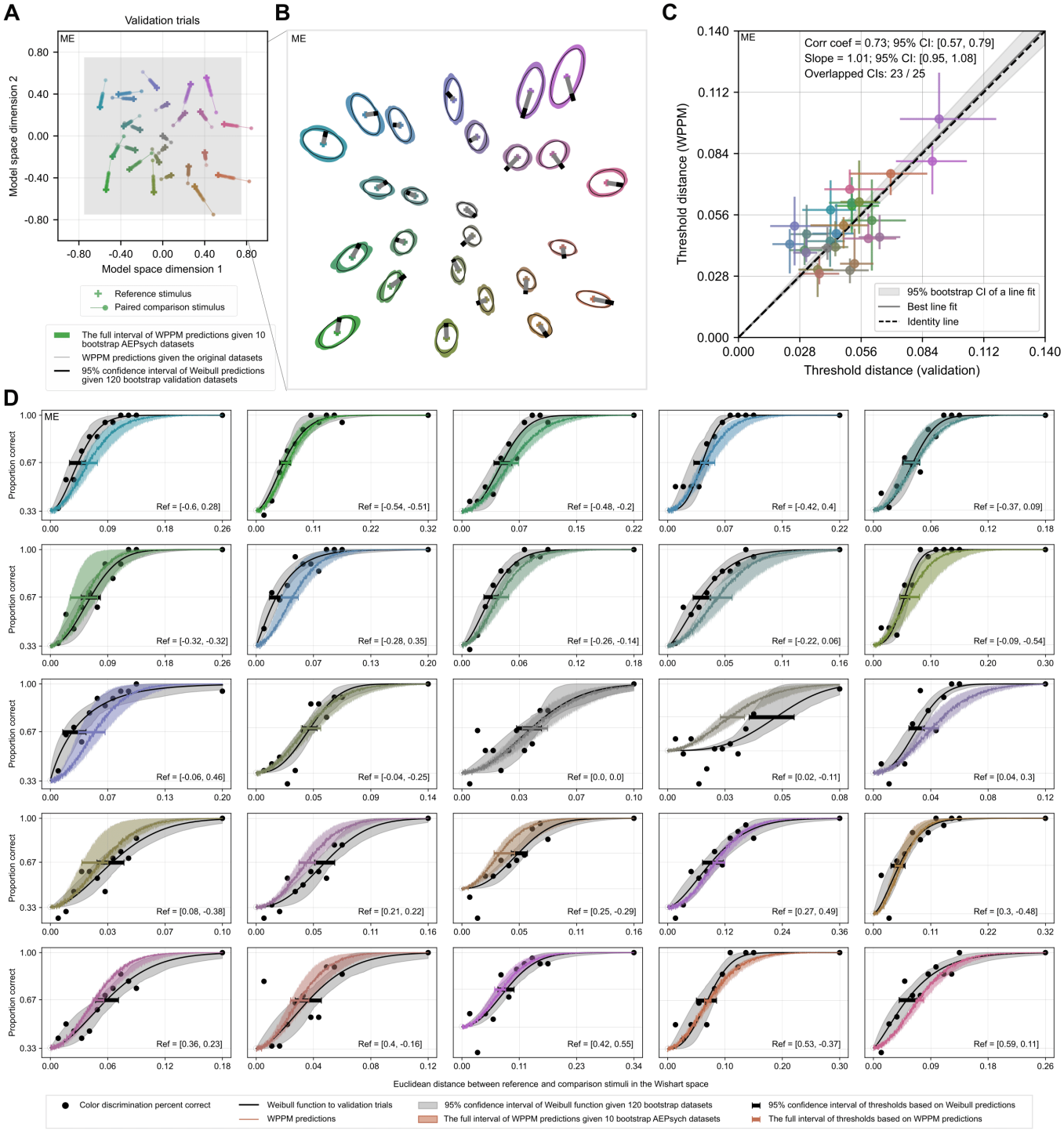

Figure S3. Validation for participant ME. Same format as ***Figure 2***D-G in the main text.

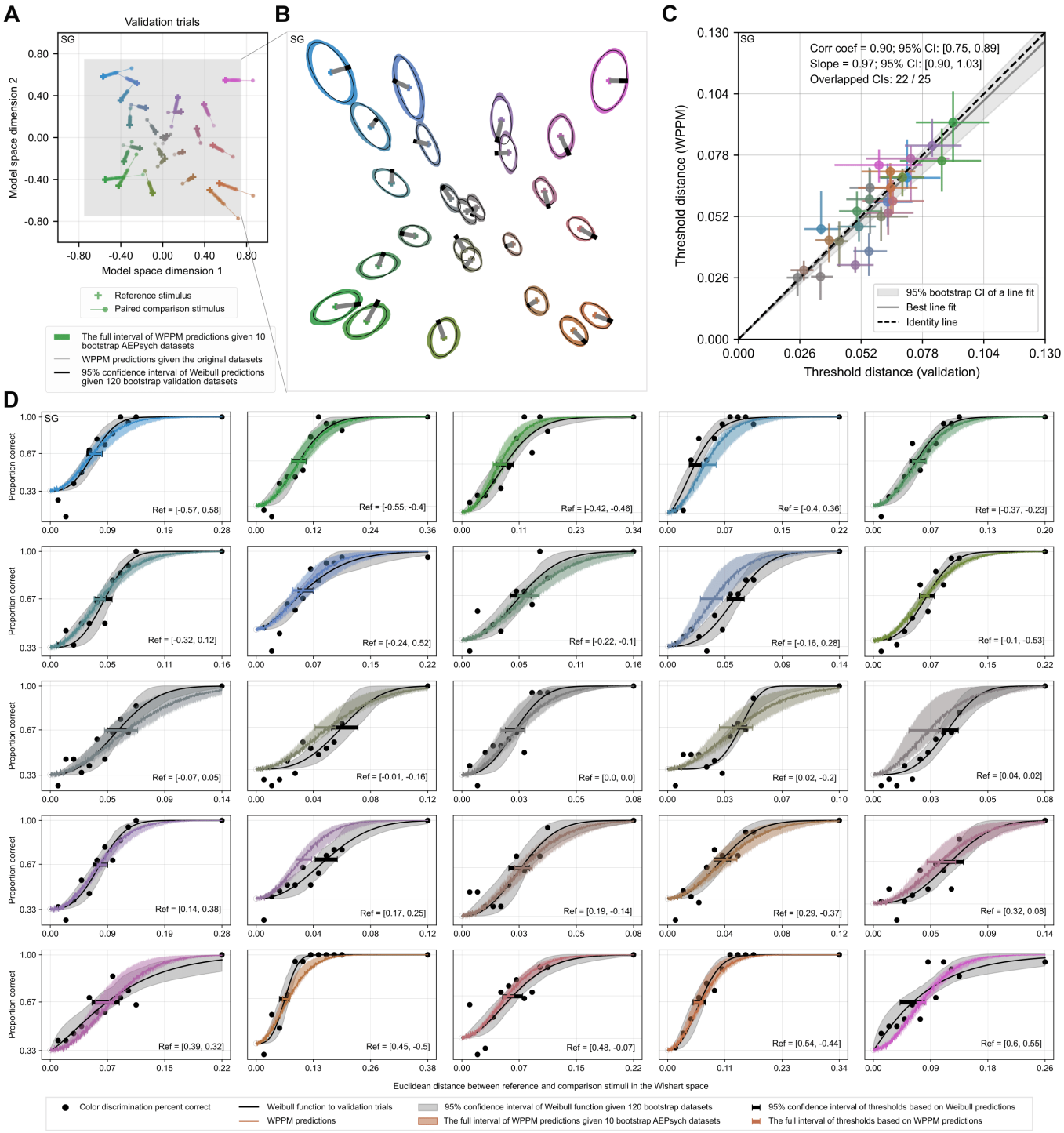

Figure S4. Validation for participant SG.

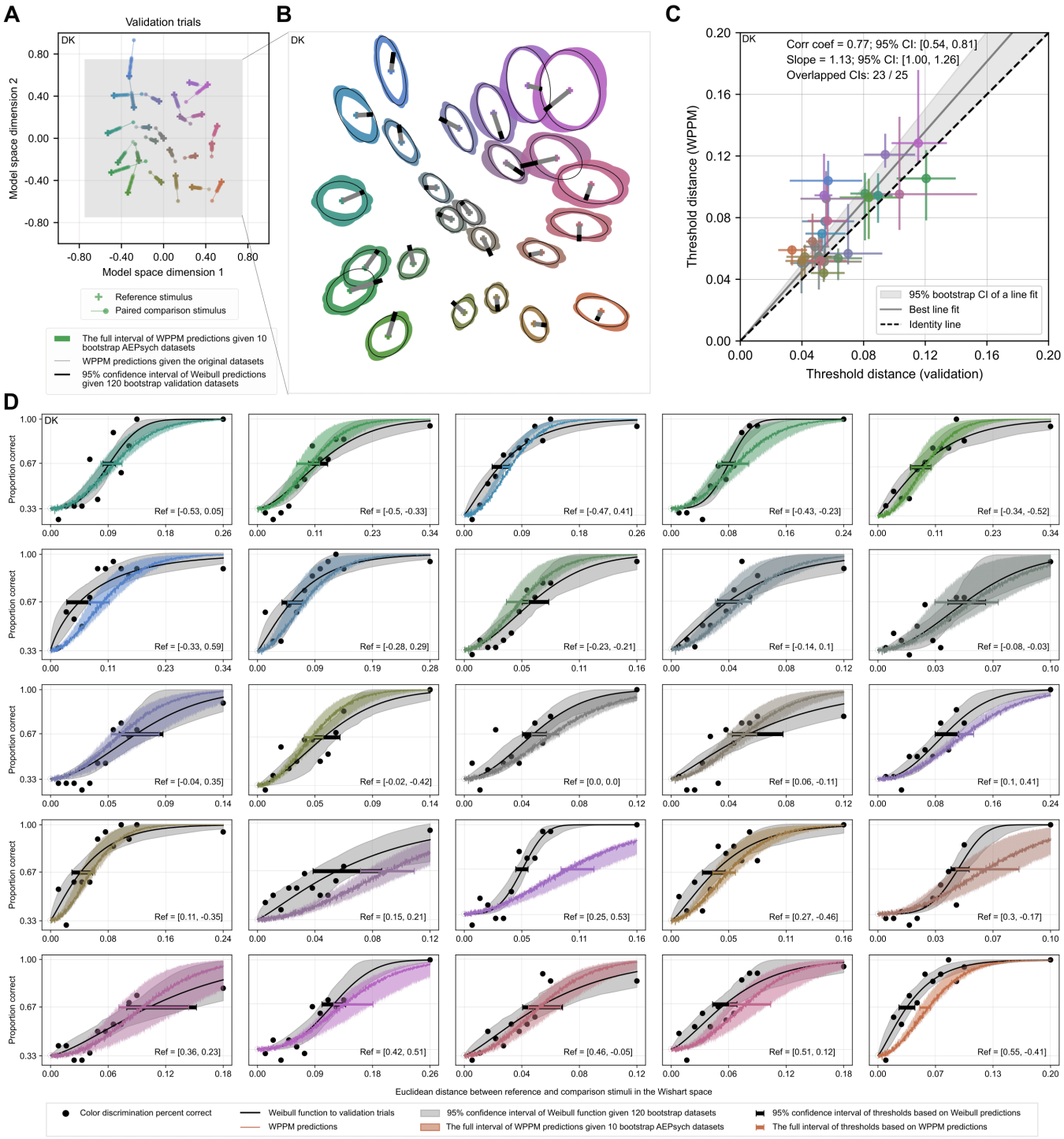

Figure S5. Validation for participant DK.

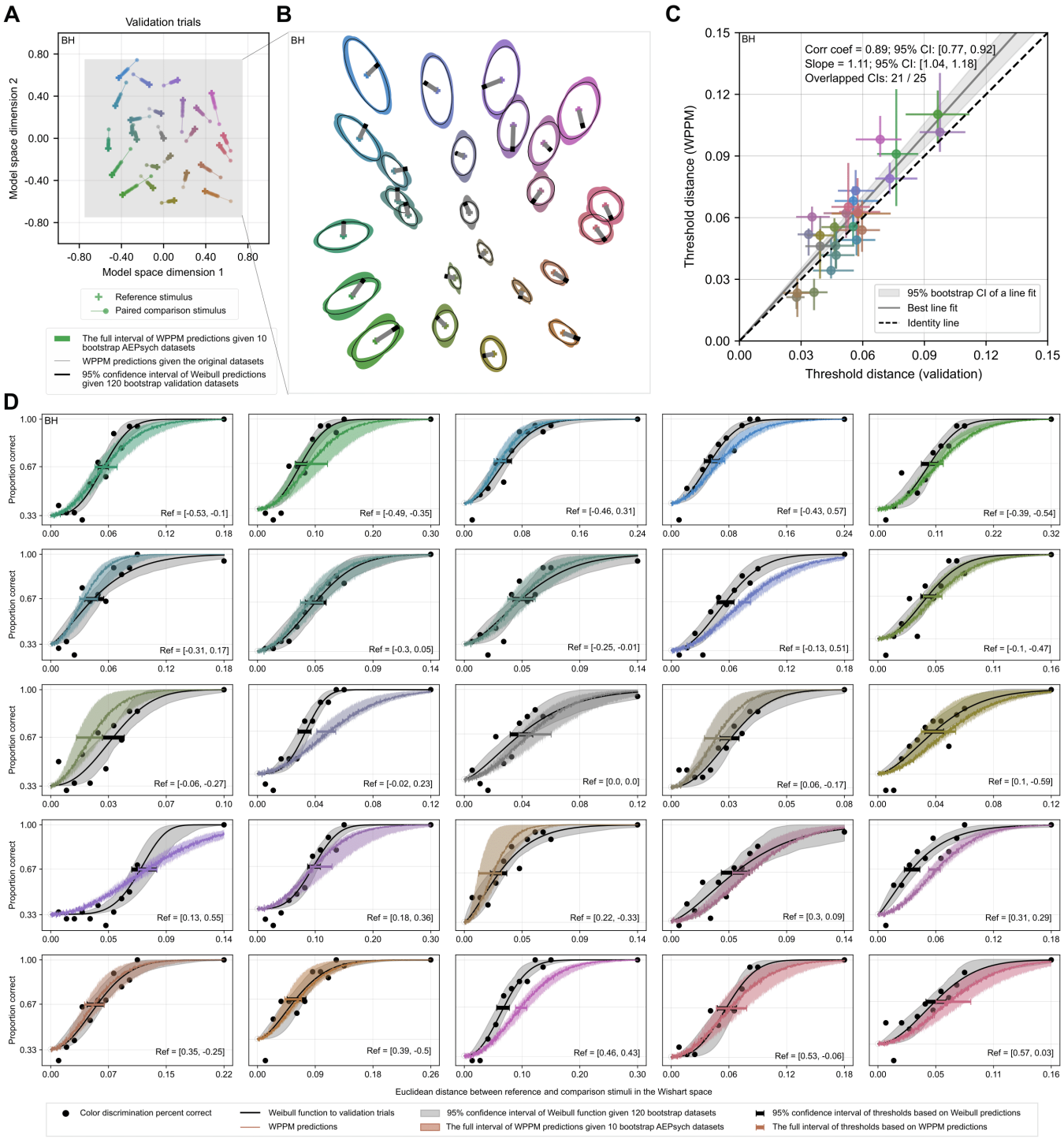

Figure S6. Validation for participant BH.

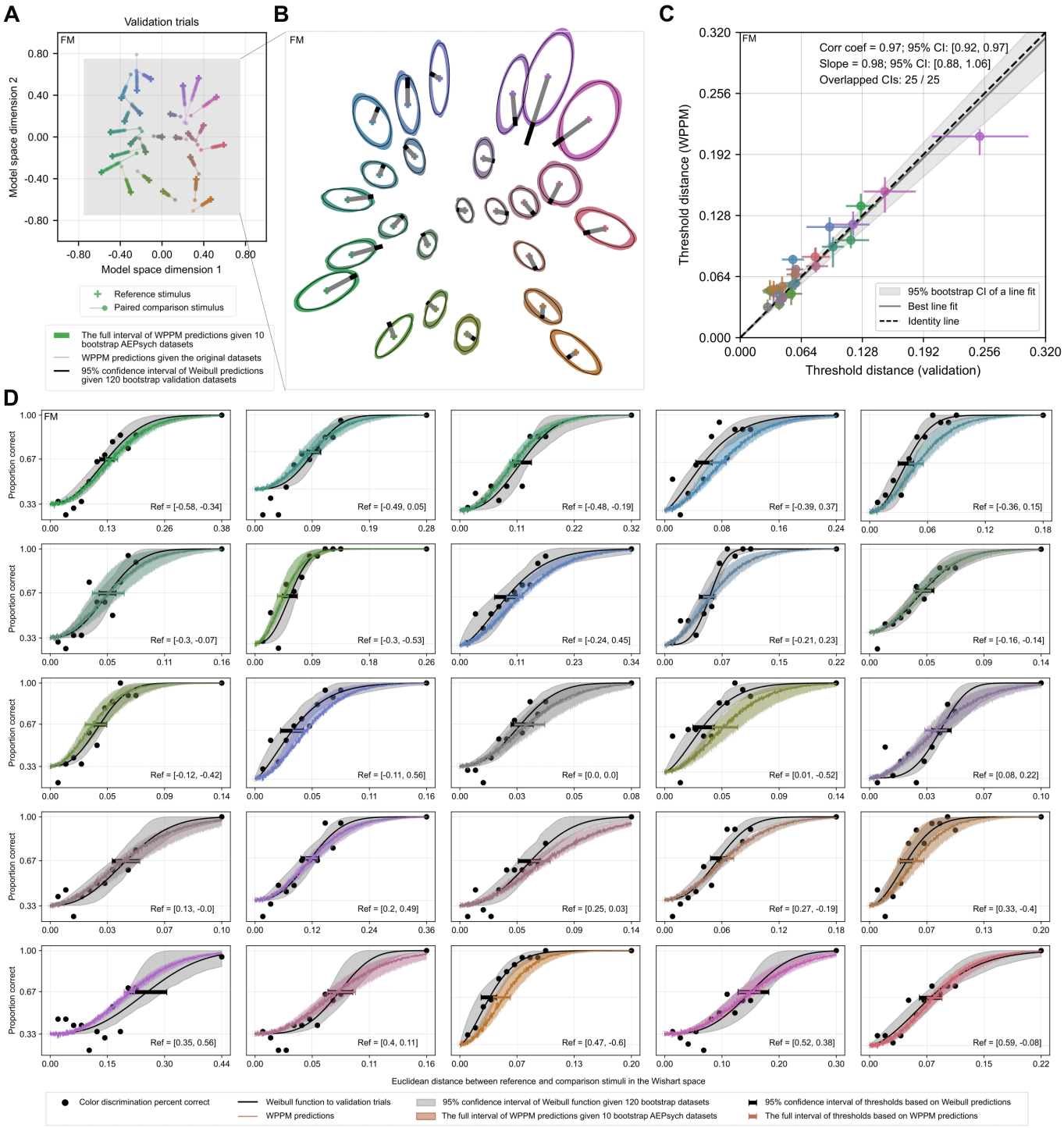

Figure S7. Validation for participant FM.

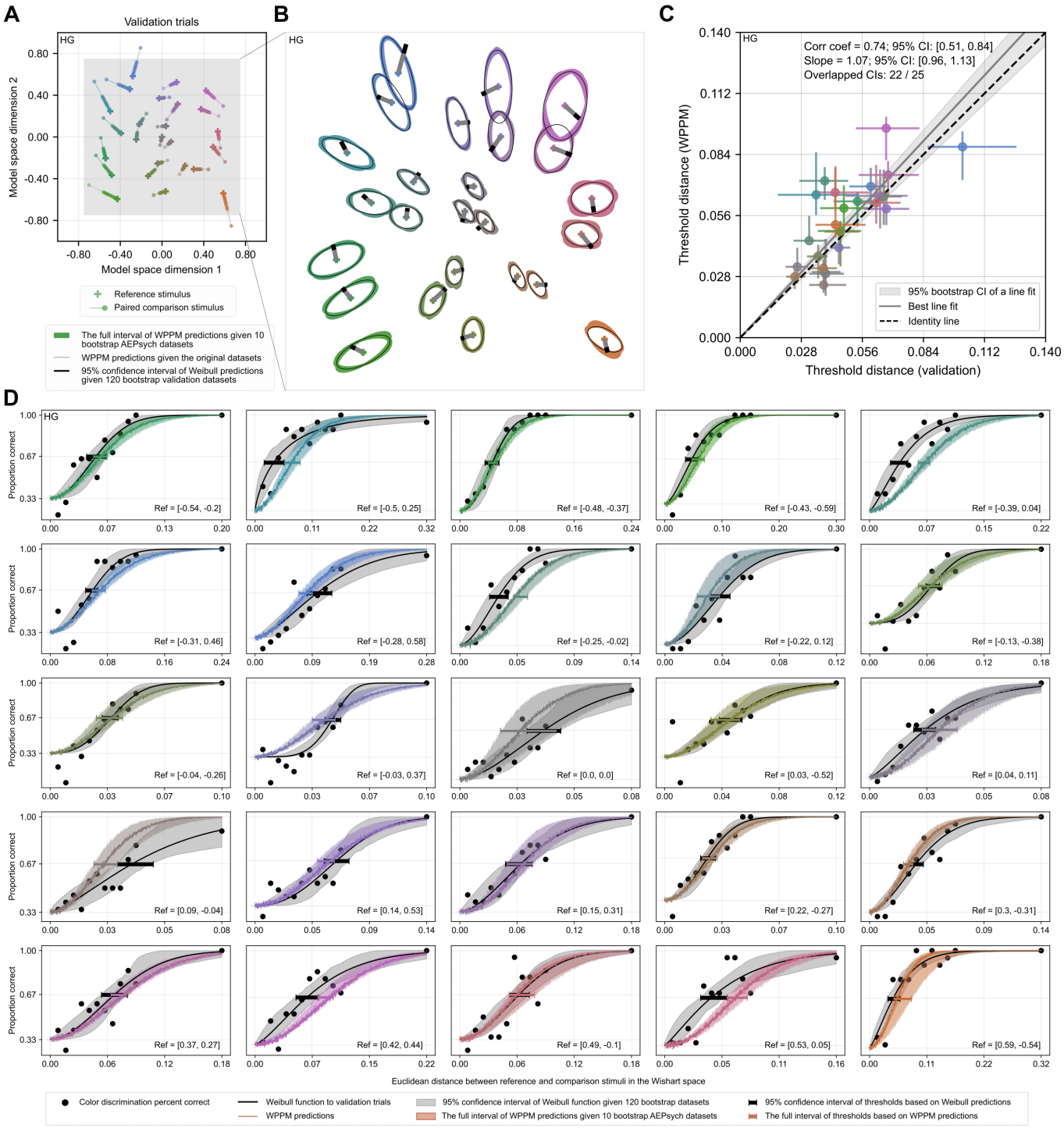

Figure S8. Validation for participant HG.

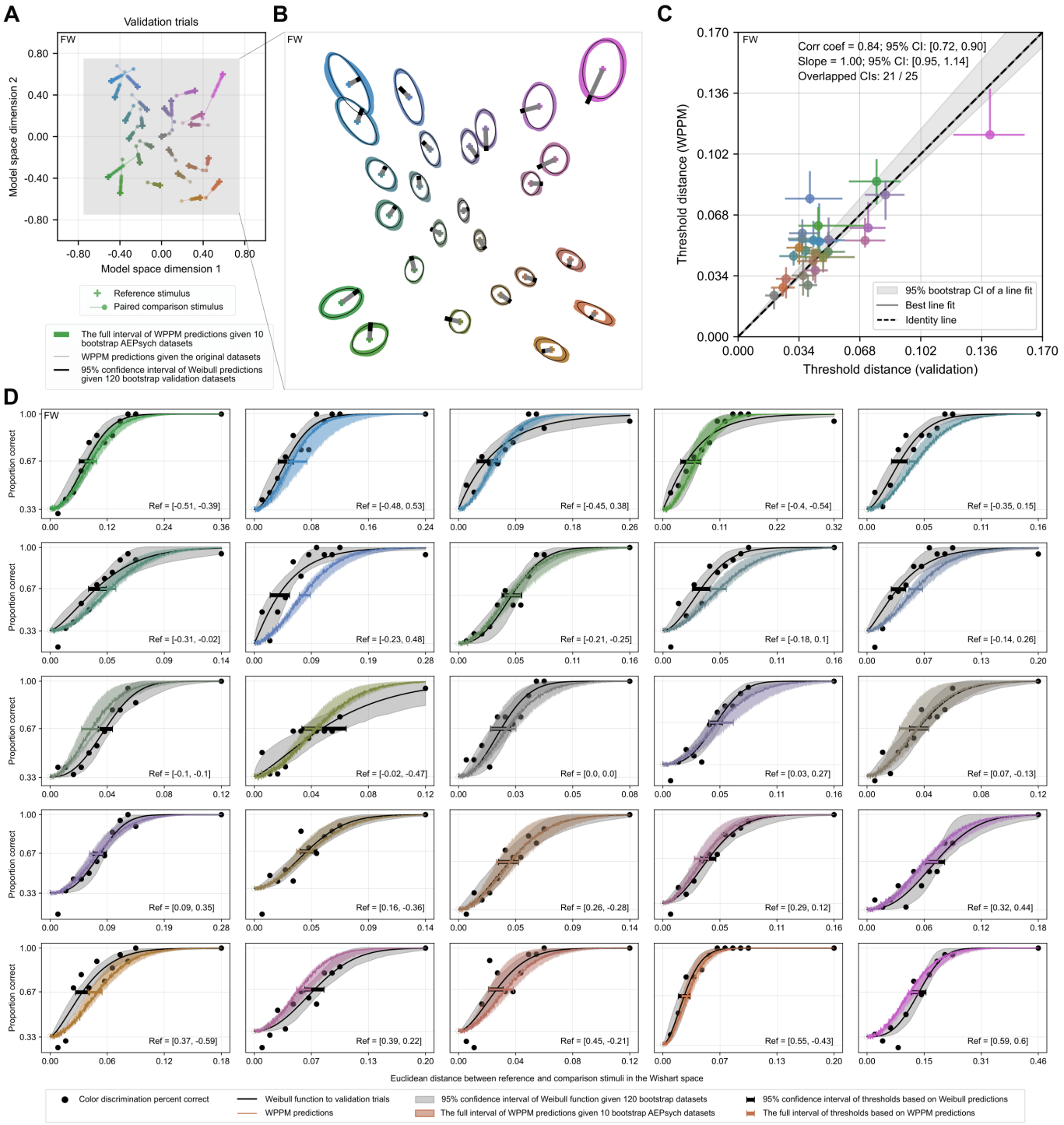

Figure S9. Validation for participant FW.

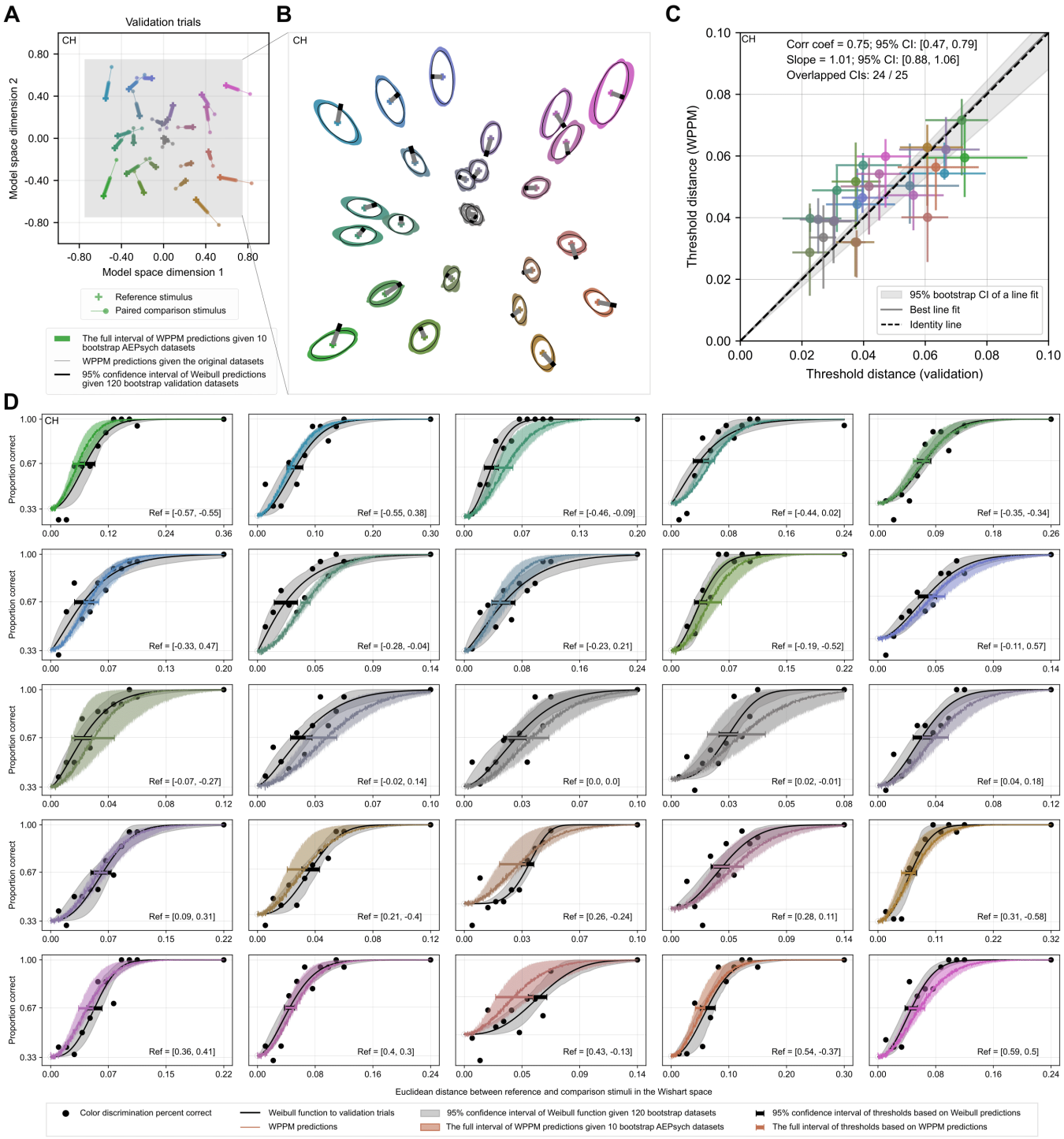

Figure S10. Validation for participant CH.

Appendix 4.2: Analysis of discrepancies between WPPM and validation thresholds

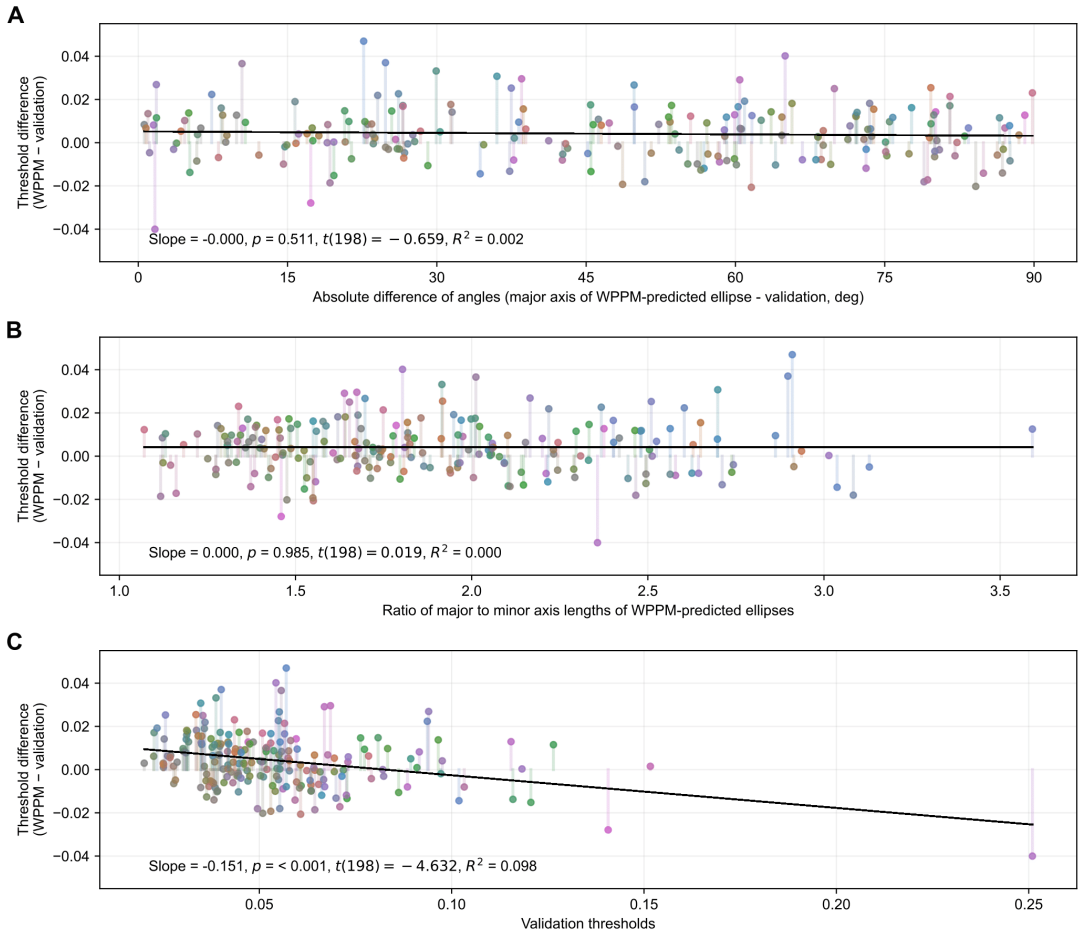

Figure S11. Threshold residuals. Data are pooled across all validation conditions and all participants ($N = 8$). For all panels, color codes for the surface color of the reference stimulus, and the y-axis limits are set to $\pm$ the mean of the validation thresholds. (A) Residuals as a function of the 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 discrepancies 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 S11***A), nor with the aspect ratio of the contours (***Figure S11***B). 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 S11***C; slope = -0.151, $t(198) = -4.632$, $p < 0.001$, $R^2 = 0.098$), 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	0.005	0.002	2.875	0.004	[0.002, 0.009]	0.002
	slope	-0.000	0.000	-0.659	0.511	[-0.000, 0.000]	
Aspect ratio	intercept	0.004	0.004	1.129	0.260	[-0.003, 0.011]	0.000
	slope	0.000	0.002	0.019	0.985	[-0.004, 0.004]	
Validation thresholds	intercept	0.013	0.002	6.269	0.000	[0.009, 0.016]	0.098
	slope	-0.151	0.033	-4.632	0.000	[-0.215, -0.087]	

Table S3. Linear regression results assessing the relationship between WPPM–validation threshold discrepancies 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 S5***), 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 participant

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 ΔE 94 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 S12***A). This procedure was repeated across multiple

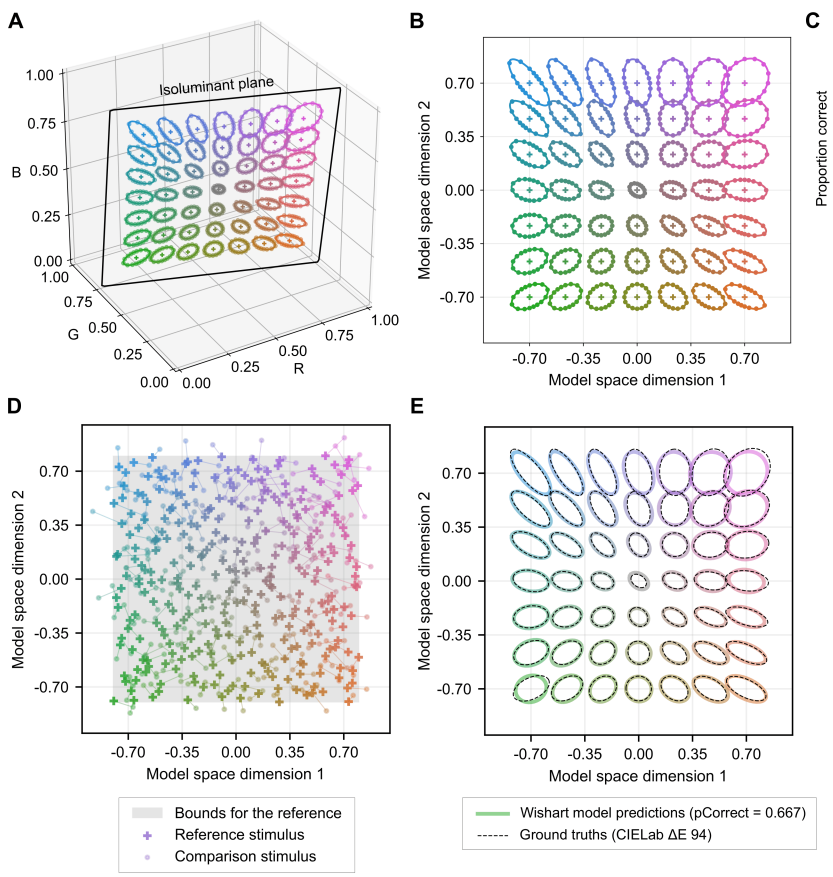

Figure S12. Derivation of the ground-truth Wishart fits based on CIELab ΔE 94. (A–B) Comparison stimuli at the iso-distance contours in the isoluminant plane, shown in both RGB and model spaces. Note that the reference grid and fixed set of directions shown here are for illustration only; the actual 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 ΔE 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 ΔE 94. The WPPM fit was subsequently treated as the ground truth for simulating AEPsych and validation trials.

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 S12***B).

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:

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

We modeled performance using a Weibull psychometric function, which took the ΔE 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 $\Delta E = 2.5$ (***Figure S12***C). 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 S12***D). 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 both AEPsych and validation trials (***Figure S12***E, color lines). We chose to use the WPPM fit as the ground truth—rather than percent-correct performance derived from CIELab ΔE 94 with a Weibull psychometric function—because our goal was to evaluate how well the WPPM can recover the simulated data. Using a ground truth that is itself an instance of the WPPM fit provides a more direct and interpretable comparison, avoiding local irregularities or discontinuities that might be present in CIELab ΔE 94 predictions and thus inherently difficult for the WPPM to characterize.

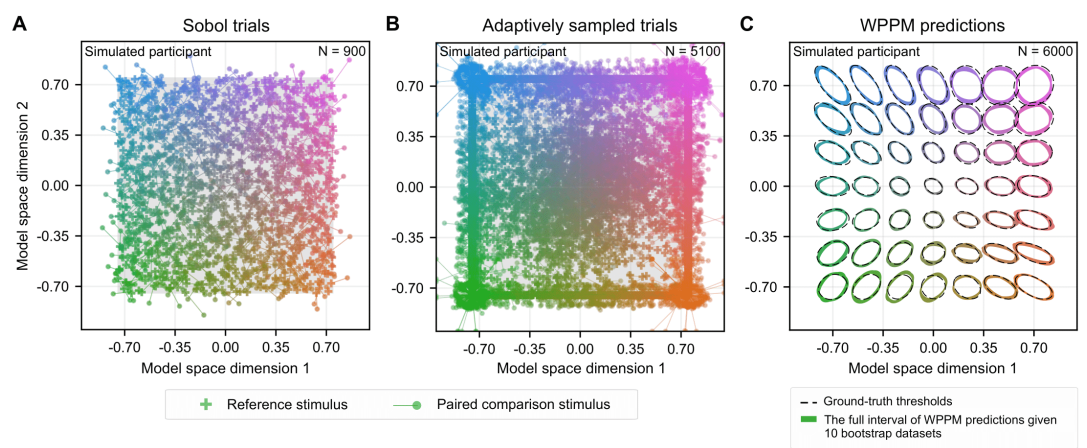

Figure S13. AEPsych-driven trials and WPPM readouts for a simulated participant. Note that the ground-truth thresholds shown in (C) is the same WPPM readouts from ***Figure S12***E.

Appendix 5.4: Fit the WPPM to simulated AEPsych trials

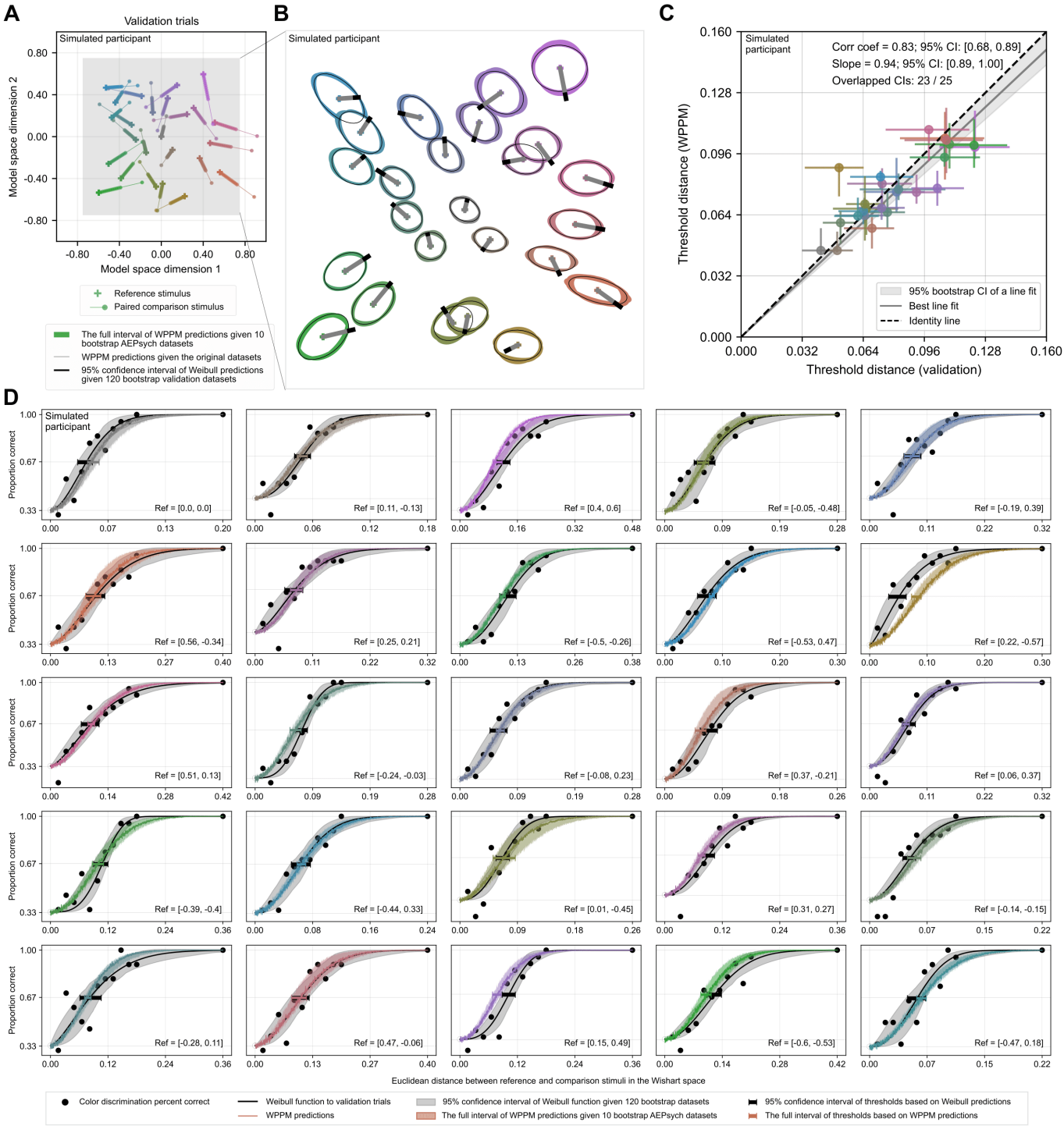

Figure S14. Validation trials and WPPM readouts for a simulated participant.

Based on the ground-truth WPPM fit, we simulated 900 Sobol' trials (***Figure S13***A), followed by 5,100 adaptively sampled trials using AEPsych (***Figure S13***B), 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 10 times, maintaining the same Sobol'-to-adaptive trial ratio within each bootstrapped dataset. As shown in ***Figure S13***C, the WPPM was able to reliably recover the ground-truth model, with some minor local deviations.

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 S14***). 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 discrepancies between WPPM readouts and validation thresholds

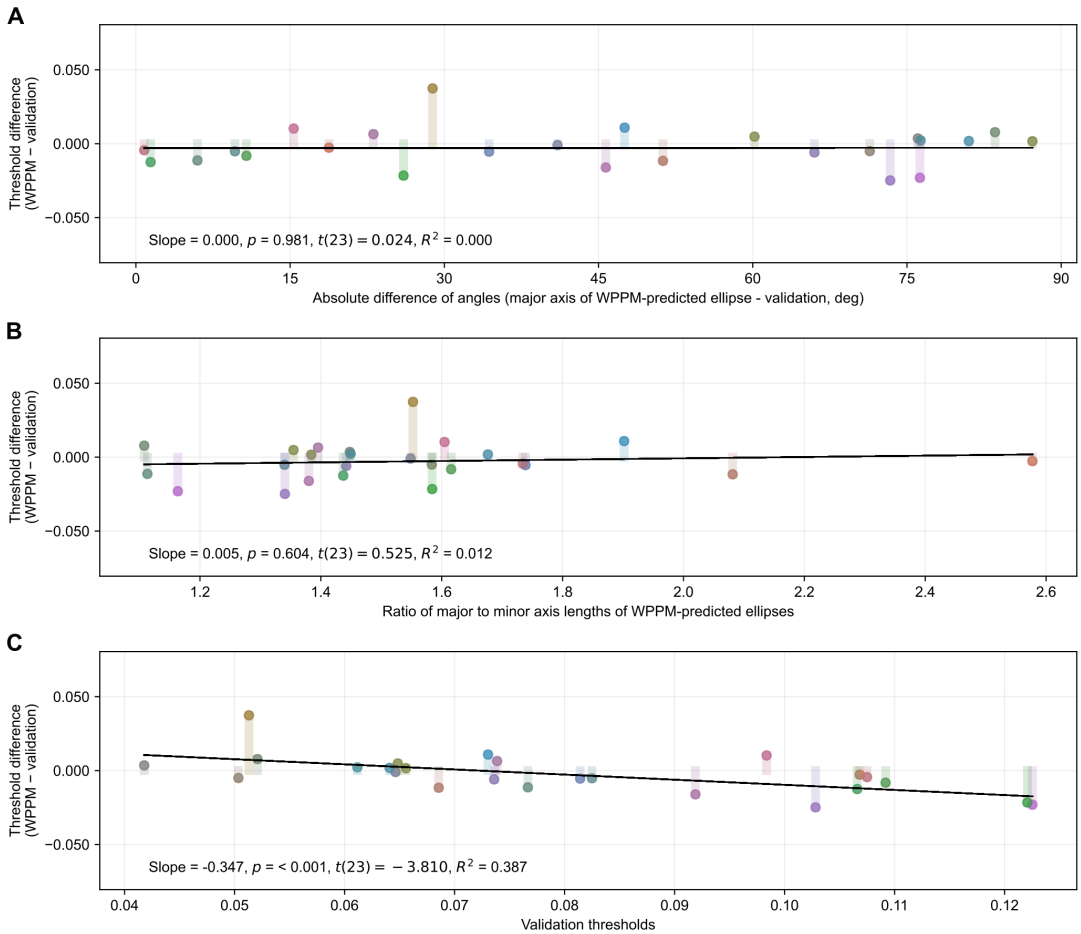

Figure S15. Threshold residuals for a simulated dataset. For all panels, color codes for the surface color of the reference stimulus, and the y-axis limits are set to ± the mean of the validation thresholds. (A) Residuals as a function of the absolute angular difference between the major axis of the 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.

We applied the same statistical analysis to the simulated data as we did for the human data (***subsection***). Consistent with the human results (***Figure S11***), 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 = -0.347, $t(23) = -3.810$, $p < 0.001$, $R^2 = 0.387$; ***Figure S15***; ***Table S5***). As noted earlier, the size of this bias is small compared to the overall range of validation thresholds.

Predictor	Term	Coef	Std Err	t	p	[0.025, 0.975] CI	R^2
Absolute difference of angles	intercept	-0.003	0.005	-0.603	0.553	[-0.013, 0.007]	0.000
	slope	0.000	0.000	0.024	0.981	[-0.000, 0.000]	
Aspect ratio	intercept	-0.010	0.014	-0.727	0.475	[-0.038, 0.018]	0.012
	slope	0.005	0.009	0.525	0.604	[-0.013, 0.022]	
Validation thresholds	intercept	0.025	0.008	3.287	0.003	[0.009, 0.041]	0.387
	slope	-0.347	0.091	-3.810	0.001	[-0.535, -0.158]	

Table S5. Linear regression results for the simulated dataset.

Appendix 5.7: Comparison between the WPPM estimates and the ground truths

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

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

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 S16***A), 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 S16***A). Overall, the mismatches observed in our simulations were modest—well below the level expected if the model were fundamentally mischaracterizing the threshold shapes.

We also examined discrepancies 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 S16***B; also obvious in ***Figure S13***C). 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.

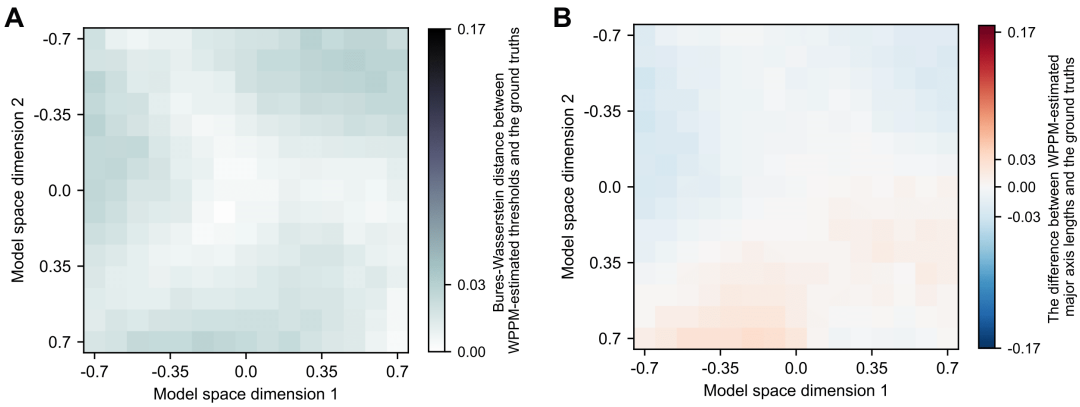

Figure S16. Deviation of WPPM estimates from the ground truth. (A) BW distance between WPPM-estimated thresholds and the ground-truth ellipses. The upper limit of the color map (0.17) corresponds to the maximum BW distance between each ground-truth ellipse and a reference circle whose radius equals the 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.

Appendix 6
Comparison with MacAdam ellipses (1942)

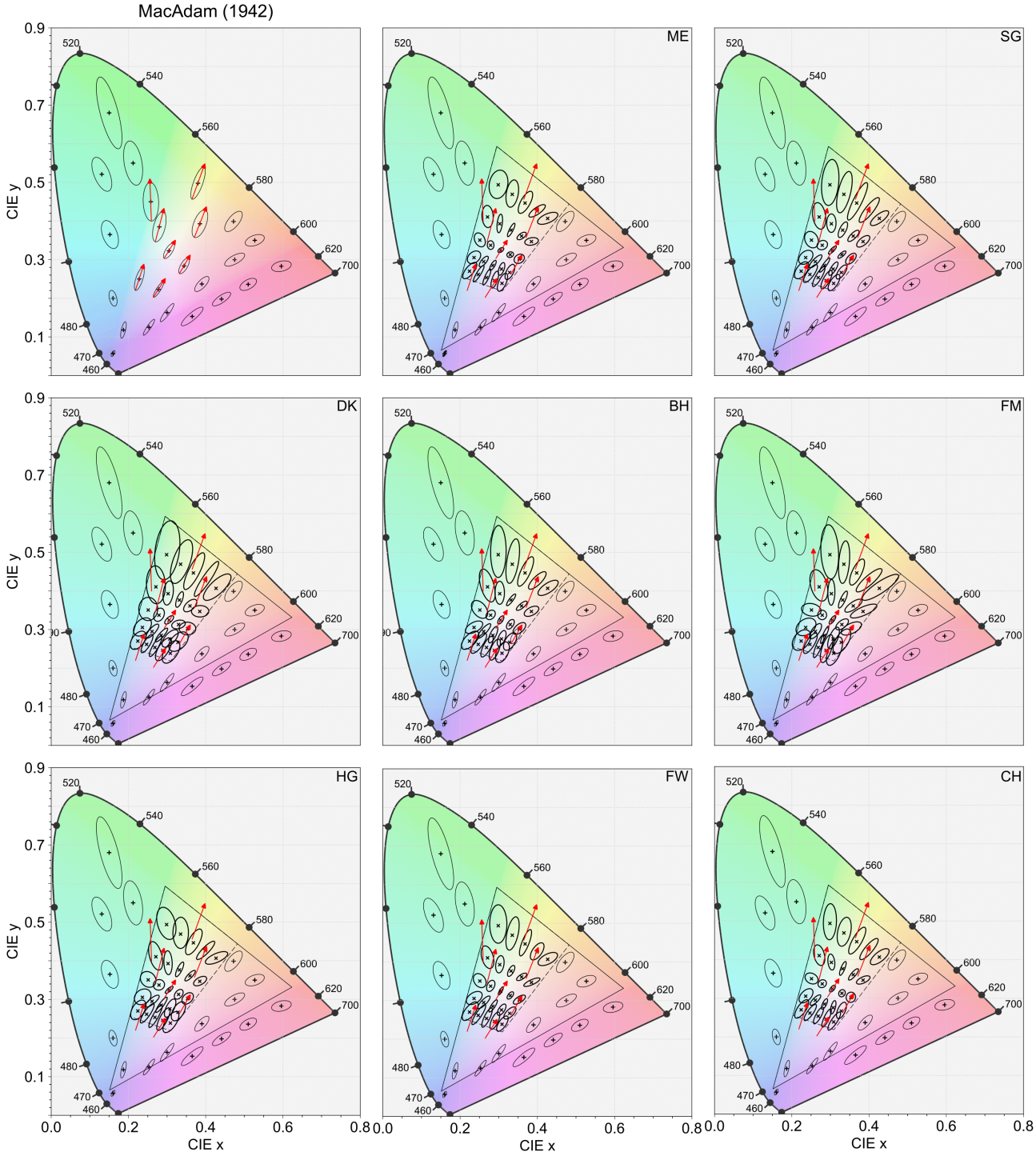

Figure S17. Comparison with ***MacAdam 1942***. Left: MacAdam's original ellipses, enlarged $10\times$ for visualization. Red arrows indicate inferred major axis directions at unsampled reference locations, guessed from nearby ellipses. Right: Threshold ellipses from our measurements, also magnified by $2\times$ for visualization. Triangle: gamut of our monitor; parallelogram: gamut of our isoluminant plane.

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} → $\Delta\mathrm{LMS}_{us}$ → DKL → $\Delta\mathrm{LMS}_{dm}$ → 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 a fair visual comparison of the threshold contours. The comparison reveals a general qualitative agreement between their measurements and ours (***Figure S18***).

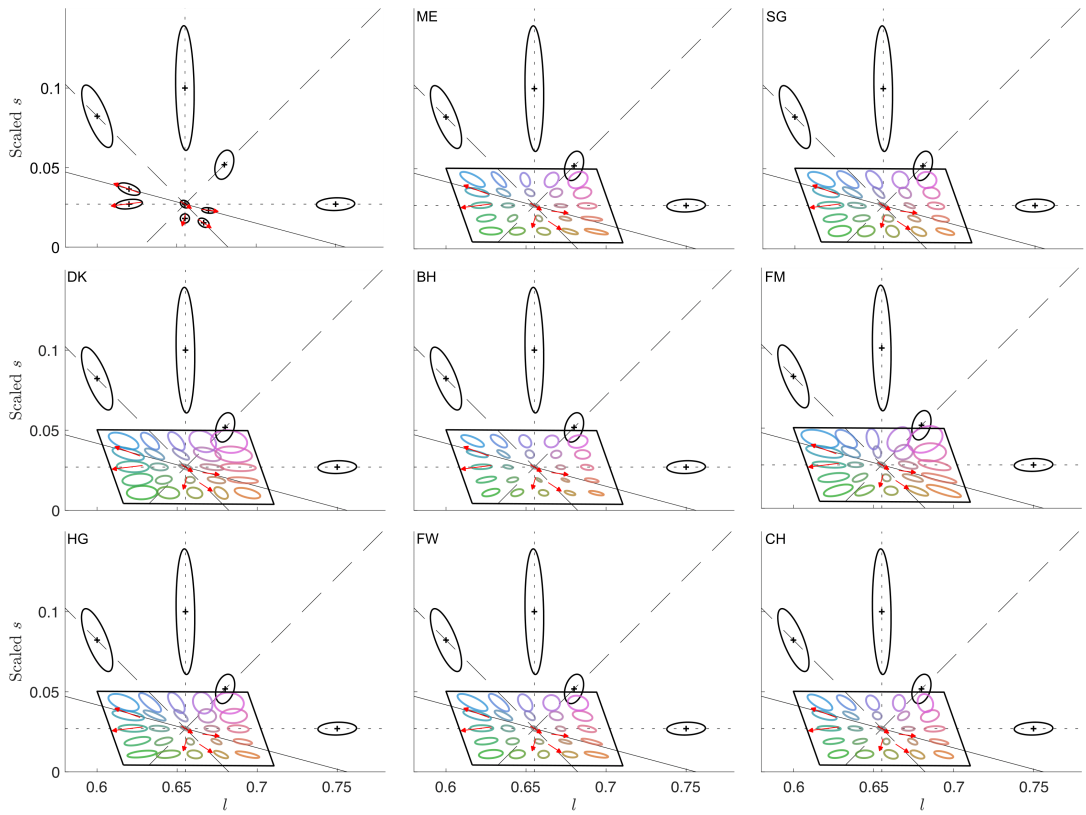

Figure S18. Comparison with ***Danilova and Mollon 2025*** in the scaled MacLeod–Boynton space. Top left: threshold contours from their study (black ellipses), enlarged by $4\times$. Remaining panels: threshold contours from all participants in our study (colored ellipses; $N = 8$). We sampled a grid 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 $1.5\times$ 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 Karl 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 Karl 1992*** conducted their measurements. Finally, we transformed a set of elliptical threshold contours at other reference locations from our original model space into this stretched DKL space to enable direct comparison (***Figure S19*** shows the transformation and comparison for a representative participant; ***Figure S20*** shows results for the remaining participants).

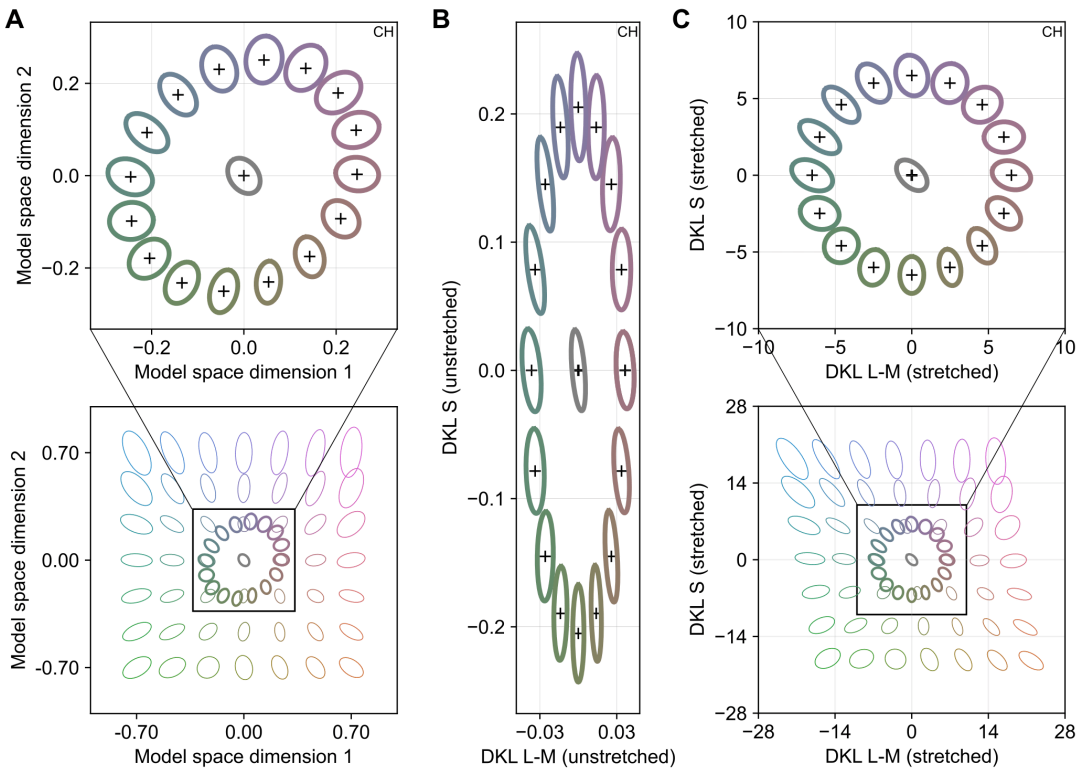

Figure S19. Transformation from the model space to a stretched DKL space used in ***Krauskopf and Karl 1992*** for participant CH. (A) Model space. Threshold contours were read out in this space based on each participant's WPPM fit. Notably, our data were collected on a much larger region of the isoluminant 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 to one.

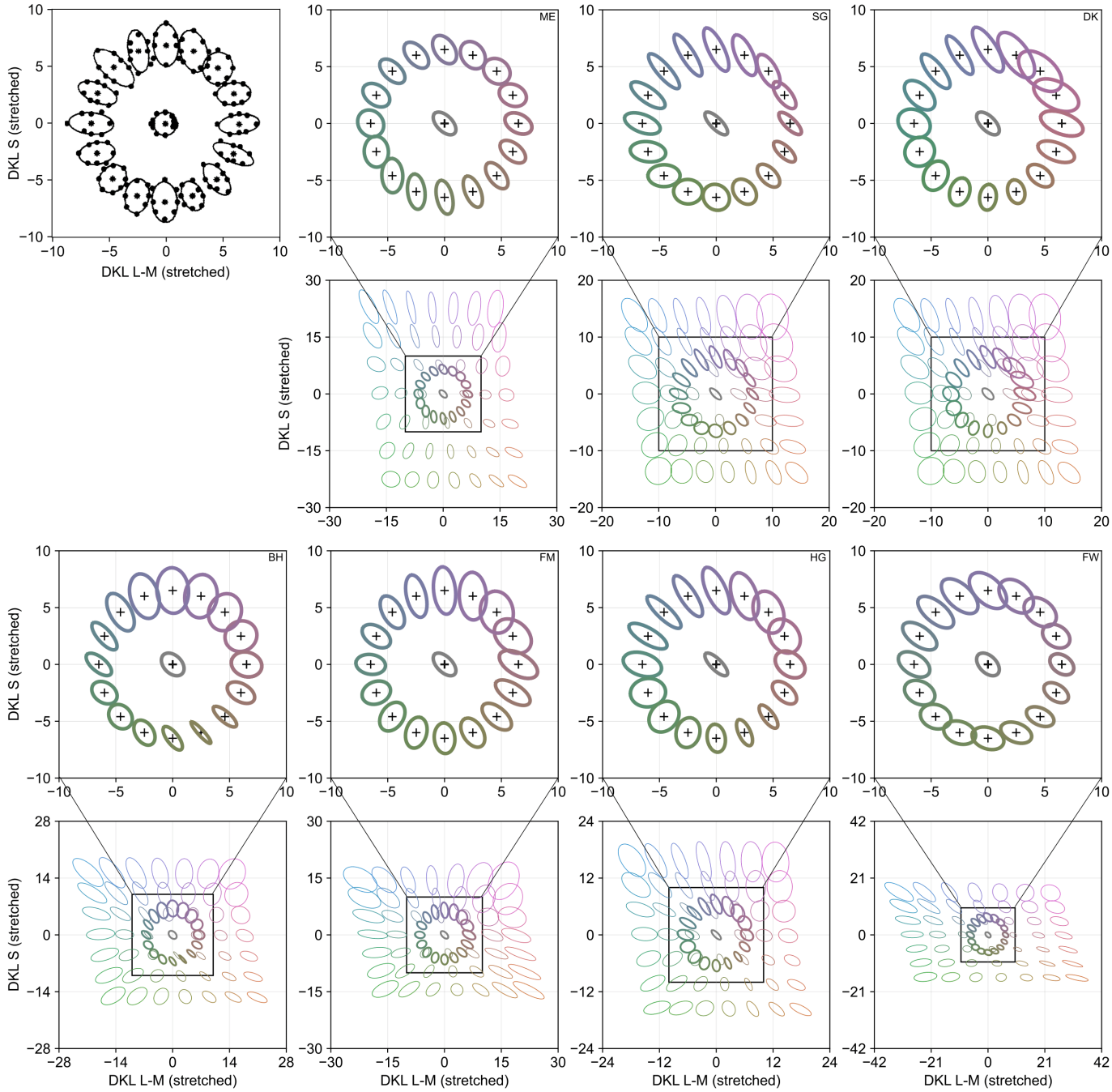

Figure S20. Comparison with ***Krauskopf and Karl 1992*** across participants. Top left: original threshold contours reported by ***Krauskopf and Karl 1992***, reproduced under Creative Commons CC BY-NC-ND 4.0). Remaining panels: threshold contours for the remaining participants, transformed into the stretched DKL space using participant-specific scaling of the cardinal axes. All contours are plotted at their original sizes.

Appendix 9

Comparison with CIELab ΔE 76, 94, 2000

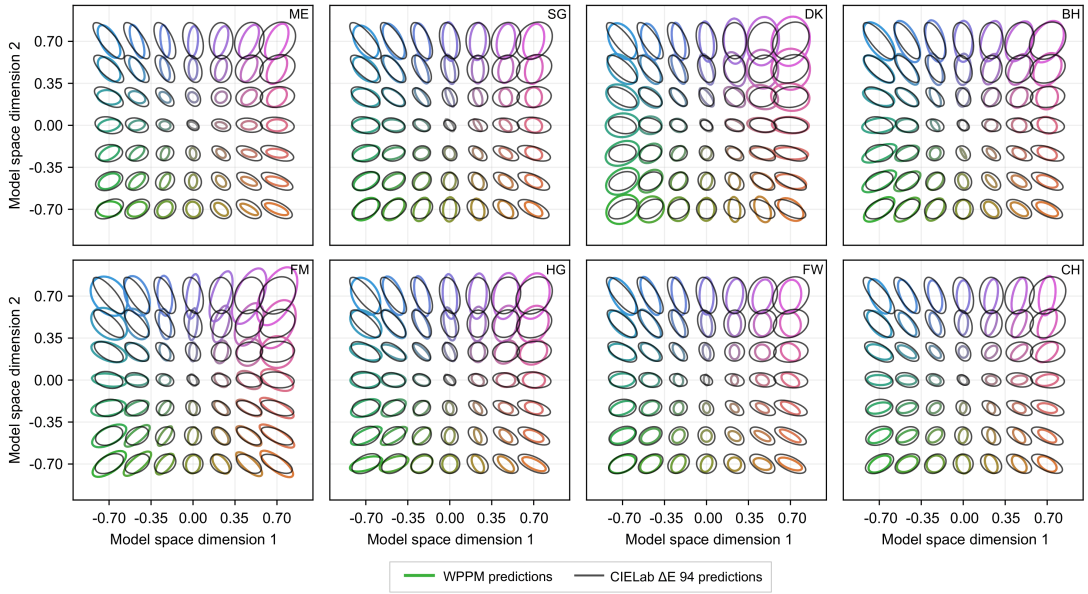

Figure S21. Comparison with CIELab ΔE 94 (***McDonald and Smith, 1995***) predictions. These are scaled by a factor of $2.5\times$ to approximately match the scale of the measured thresholds in our study, which are shown at their original scale.

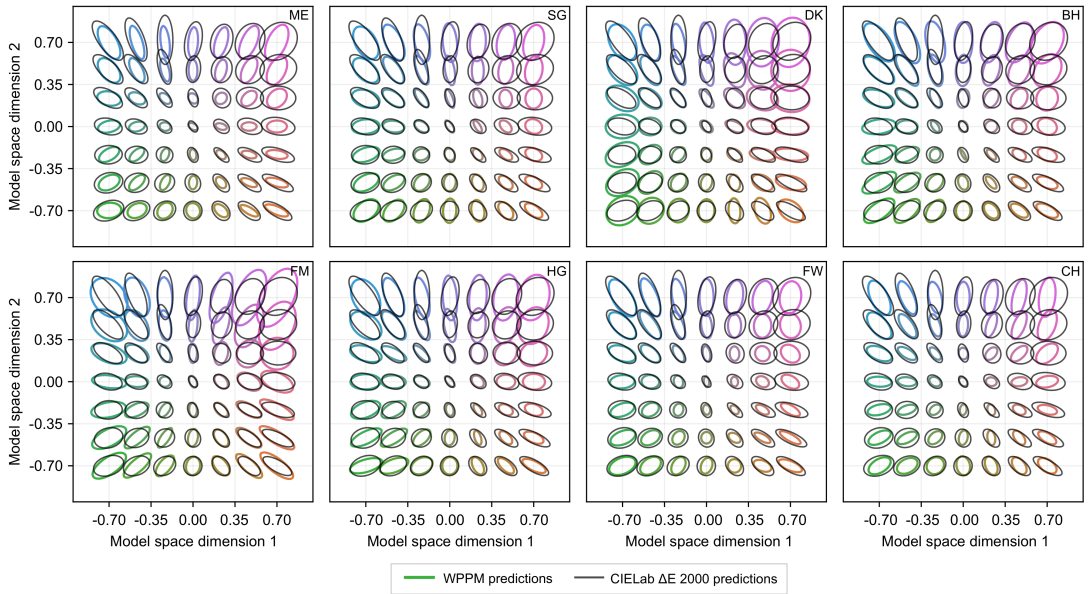

Figure S22. Comparison with CIELab ΔE 2000 (***Sharma et al., 2005***) predictions. These are scaled by a factor of $2.5\times$ to approximately match the scale of the measured thresholds in our study, which are shown at their original scale.

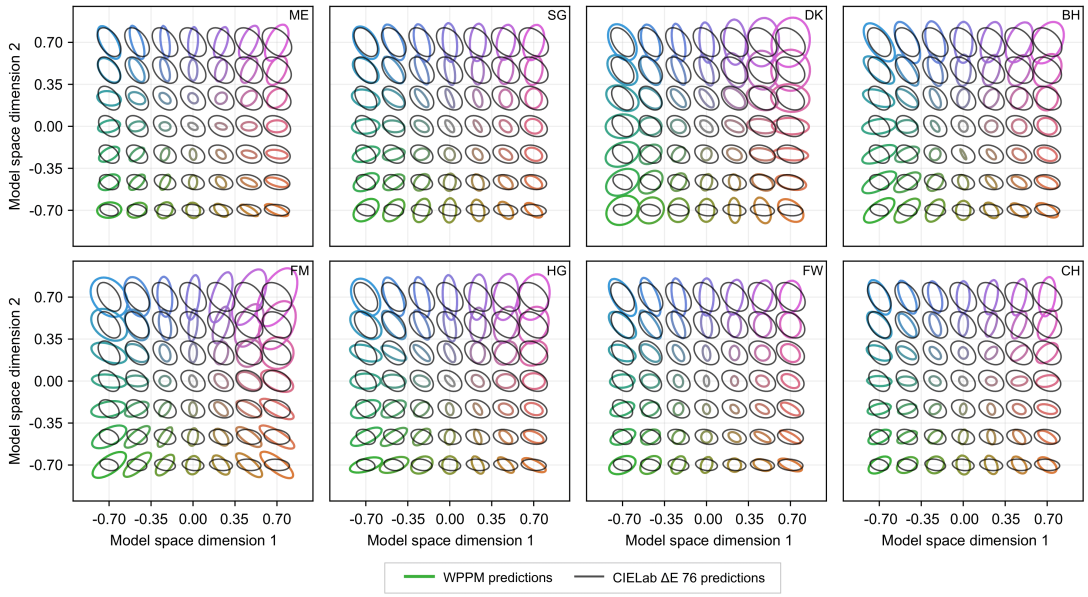

Figure S23. Comparison with CIELab ΔE 76 (***Robertson et al., 1977***) predictions. These are scaled by a factor of $5\times$ to approximately match the scale of the measured thresholds in our study, which are shown at their original scale.

As briefly described in Appendix 5, we derived threshold contours using CIELab Δ metrics as the ground truth (***Robertson et al., 1977***; ***McDonald and Smith, 1995***; ***Sharma et al., 2005***). For each reference color and chromatic direction, we applied an affine transformation from model space to RGB space using the transformation matrix described in Appendix 1.3. The resulting RGB values were then converted to CIELab coordinates, using the monitor gray point ($R = G = B = 0.5$) as the reference white in the XYZ-to-Lab transformation. Threshold points were defined in RGB space as those corresponding to a fixed perceptual distance of $\Delta E = 2.5$. These points were then transformed back from RGB to model space, where ellipses were fit to the resulting threshold contours.

Comparisons revealed that the iso-distance contours from ΔE 94 and ΔE 2000 provided reasonable approximations to our model-predicted thresholds (***Figure S21*** – ***Figure S22***), with only modest deviations. In contrast, the ΔE 76 contours—despite their continued widespread use—diverged substantially from our measured thresholds (***Figure S23***).

Appendix 10
Display characterization
Appendix 10.1: Calibration of monitor output

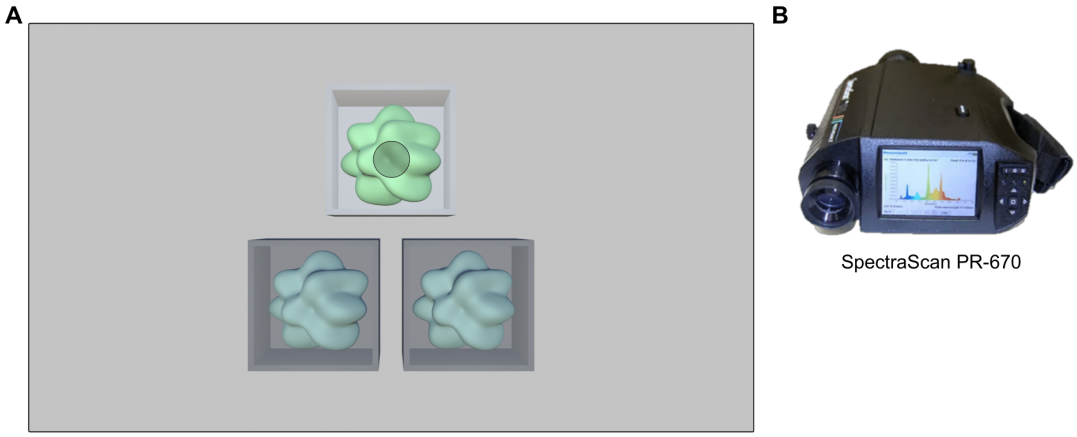

Figure S24. Stimuli and equipment used for calibration. (A) The stimulus setup during calibration was identical to that used in the main experiment. The surface color of both the cubic room and the blobby stimulus (shown here as the top-position stimulus) was varied during the calibration procedure. 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 S24***A). A SpectraScan PR-670 radiometer (***Figure S24***B), 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 S25***A). 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 S25***B), and examined the stability of the primaries' chromaticity in the CIE diagram (***Figure S25***C). There was almost no drift in the chromaticities, indicating the monitor's color output remained stable across intensity levels (***Figure S25***D). To evaluate linearity and repeatability, we compared nominal (predicted) versus measured luminance and chromaticity for two independent measurement runs (***Figure S25***E). Deviations from linearity were minimal and nearly identical across repeats, confirming reliable reproduction (***Figure S25***F). Finally, we tested whether the cubic room's background color affected the stimulus SPD; no measurable influence was detected (***Figure S25***G).

To assess consistency across different locations on the screen, we conducted the same calibration procedure on all three blobby stimuli, one at a time, and compared the primaries and chromaticities. The results showed consistent color behavior of the monitor (***Figure S26***), 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 S27***A). To validate this correction, we repeated the calibration with the gamma correction applied in Unity. The measured output closely aligned with the identity line across all three primaries, indicating accurate correction (***Figure S27***B).

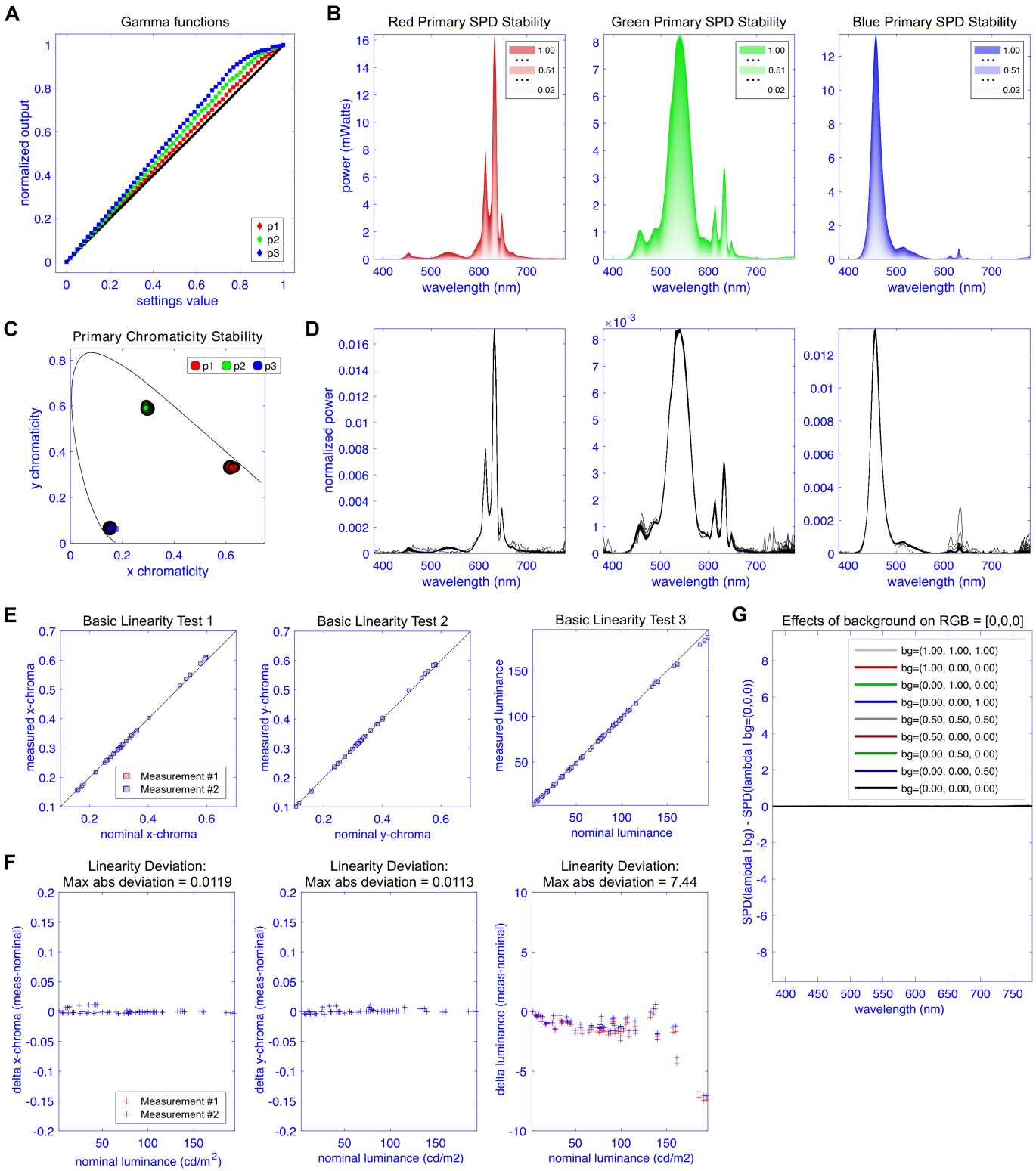

Figure S25. Calibration results. (A) Gamma functions for red, green and blue primaries. Note that Unity's internal correction places them above the identity line. (B) Spectral power distributions (SPDs) of the three primaries across a range of intensity levels. (C) The chromaticity of each primaries in the CIE chromaticity diagram at different intensity levels. (D) Normalized SPDs for each primary, showing spectral shape stability 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.

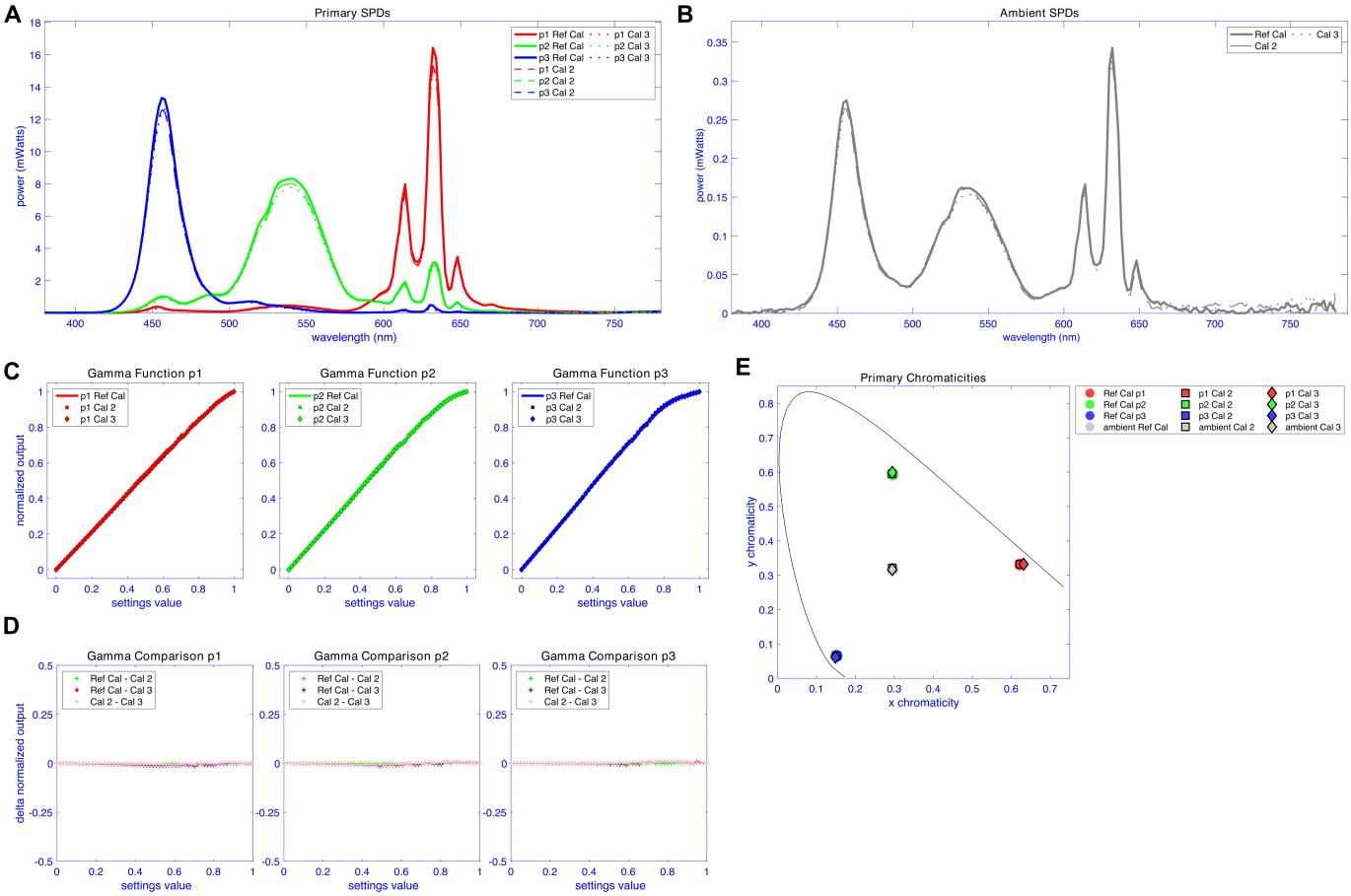

Figure S26. Comparison of calibration results across the three blobby stimuli. (A) Spectral power distributions (SPDs) for each stimulus location: Ref Cal (bottom right), Cal 2 (bottom left), and Cal 3 (top). (B) Ambient light SPDs measured during calibration. (C) Gamma functions for each primary (red, 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 calibration on the bottom-right stimulus with gamma correction applied approximately one month after data collection. The results confirmed that the correction remained accurate and consistent (***Figure S28***).

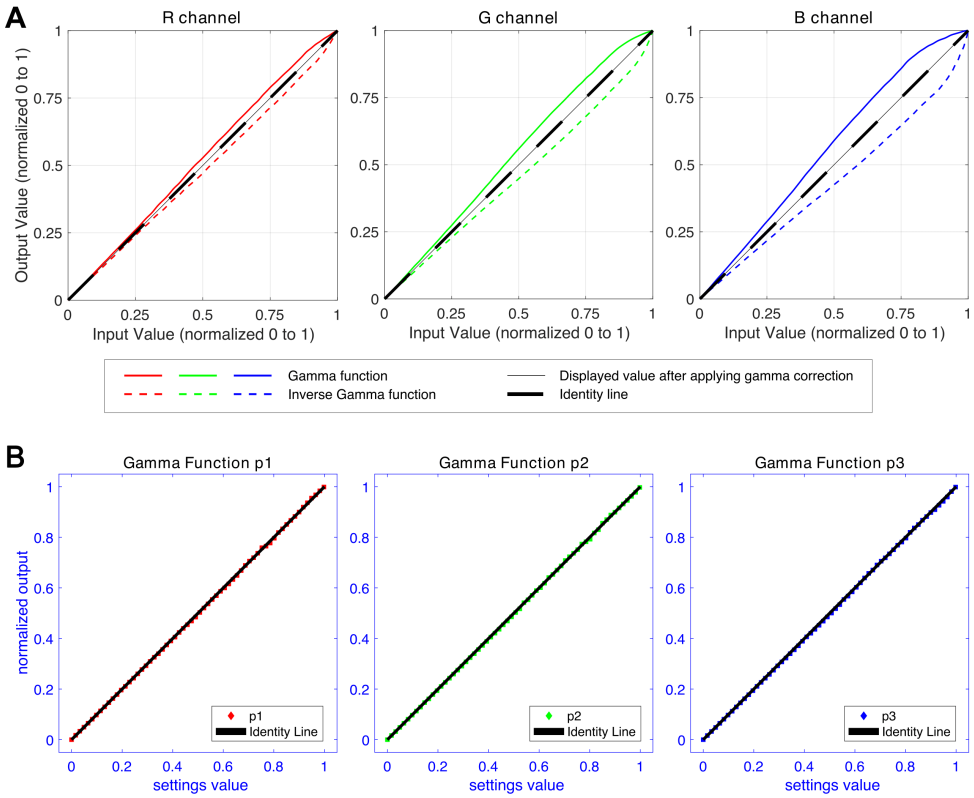

Figure S27. Gamma correction. (A) Measured gamma functions and their corresponding inverse functions for the red, green, and blue primaries, used to construct the gamma correction lookup table. (B) Gamma functions re-measured after applying the correction in Unity, showing close alignment with the identity line for all three primaries.

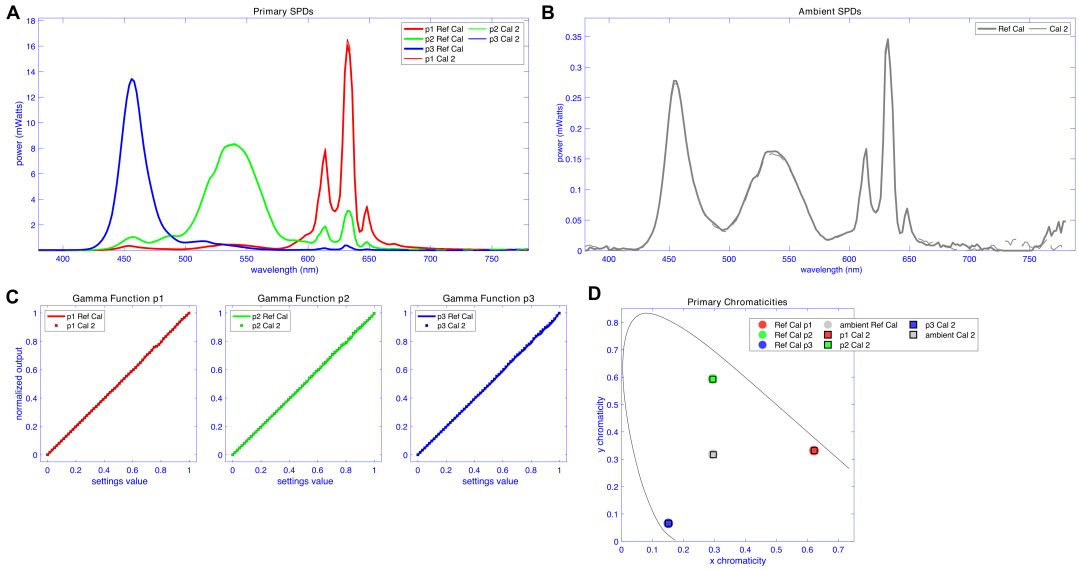

Figure S28. Comparison between the initial and repeated calibration one month into data collection. (A) Spectral power distributions (SPDs) from two calibration sessions at the bottom-right blobby stimulus location: Ref Cal (initial calibration prior to the experiment) and Cal 2 (follow-up calibration). (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.

Appendix 10.2: Assessment of color depth

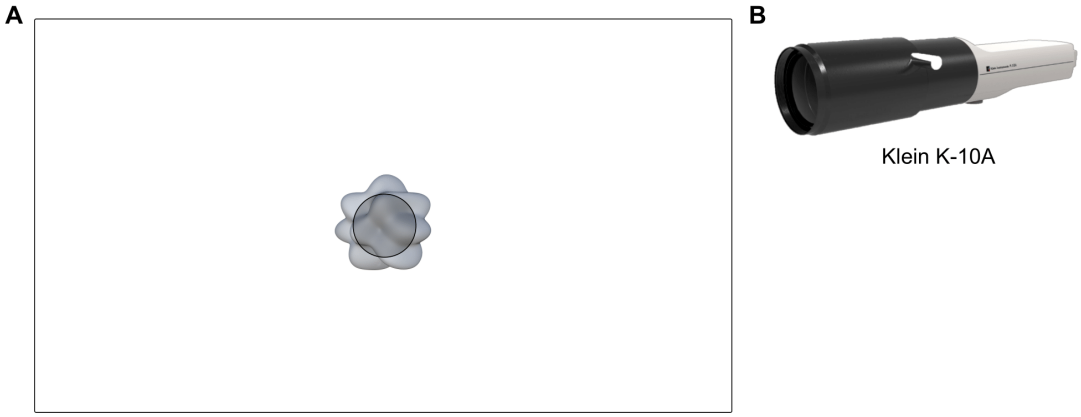

Figure S29. Stimuli and equipment used for calibration. (A) The stimulus setup during calibration was identical to that used in the main experiment. The surface color of both the cubic room and the blobby stimulus (shown here as the top-position stimulus) was varied across trials during the calibration 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.

Color depth measurements were conducted using a single blobby stimulus positioned at the center of the screen (***Figure S29***A). 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 S29***B), placed directly in front of the monitor display without any distance, was used to make the measurements.

Specifically, we tested RGB values ranging from 511/1023 to 541/1023, in increments of 1/1023. 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 throughout the full stimulus presentation. Although individual pixels exhibited quantization below 10-bit precision, the mean luminance increased with each 1/1023 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 S30***A.

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 10-bit color depth through internal spatial dithering (***Figure S30***B).

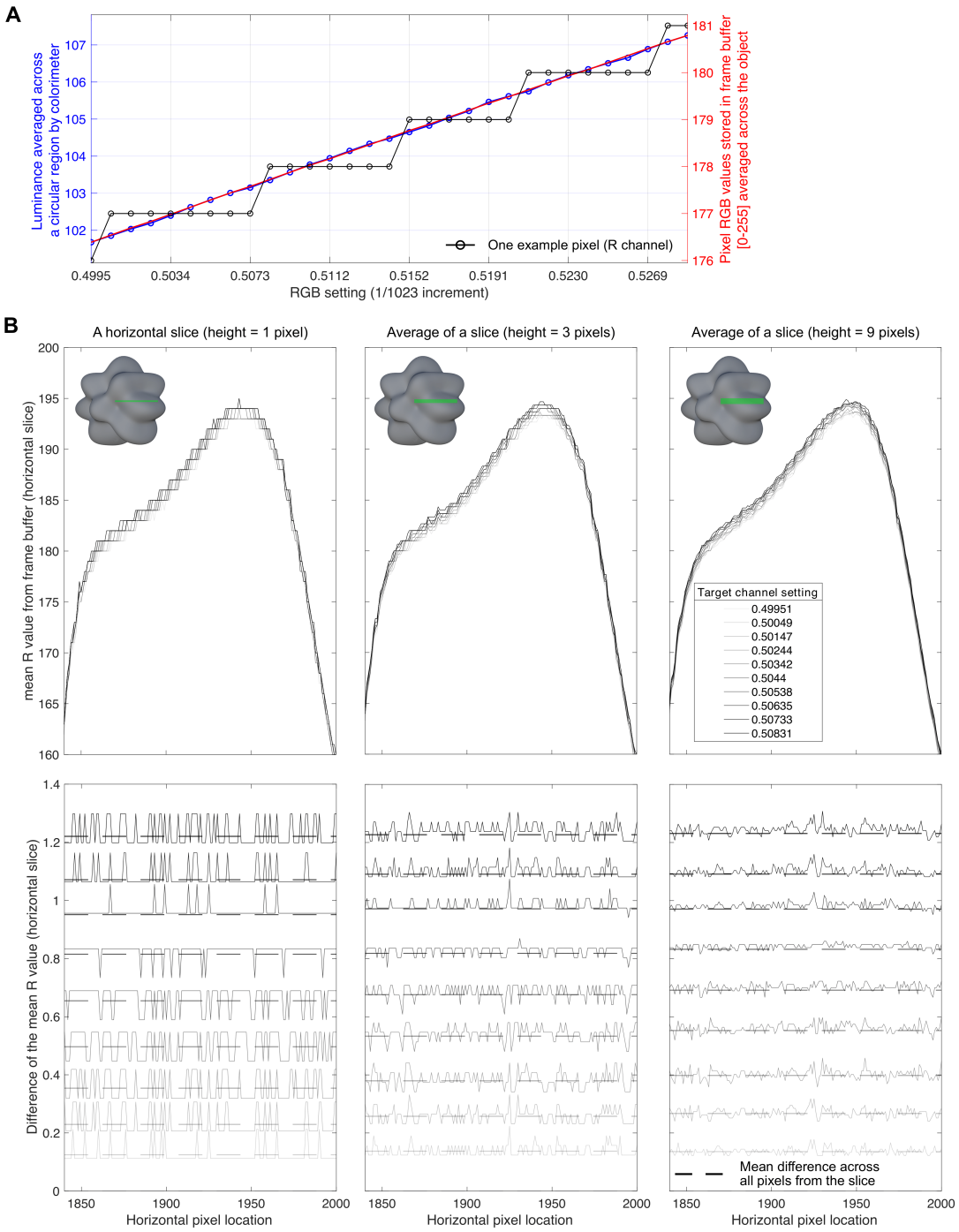

Figure S30. Evidence of spatial dithering by Unity's standard shader when the surface texture of the stimulus is being modified. (A) Spatial dithering by Unity's standard shader is suggested by comparing the luminance measurements from the Klein K-10A (averaged across a circular region on the blobby 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 1/1023. 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 10-bit quantization even though individual pixel values exhibit 8-bit quantization. (B) 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 11
Differentiable Monte Carlo Scheme

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

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

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:

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

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 auto-differentiation routines even when h is chosen to be small. Another feature is that the integrated density is the well-known *logistic function*:

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

which is familiar and easy to compute.