# Preregistration document for the color-discrimination project using AEPsych for trial placement and Wishart Process model for ad-hoc model fitting

Fangfang Hong[1], Ruby Bouhassira[1], Jason Chow[2], Craig Sanders[2], Michael Shvartsman[2], Alex Williams[3], Phillip Guan[2], David Brainard[1]

[1]University of Pennsylvania, Department of Psychology
[2]Reality Lab Research, Meta
[3]New York University, Center of Neural Science

## Purpose

This study builds upon our previous experiment to evaluate the efficacy of the Wishart process model combined with the adaptive sampling method AEPsych (v.0.7) for estimating color discrimination thresholds. This method itself will be described elsewhere; in short, it involves fitting a finite-basis Wishart process model to trial-by-trial forced-choice psychophysical data we will collect here. The fitting is based on the likelihood of the data, given the Wishart process parameters, and a Gaussian prior over the weights of the finite basis functions.

Our goal here is to determine the color discrimination thresholds in the nominal isoluminant plane, and in doing so to evaluate the efficacy of the Wishart process approach. In the initial experimental design, we measured the thresholds at a 3 × 3 grid of reference stimuli and fit the Wishart process model to the pilot data (N=5). To validate the model's predictions, we compared them to thresholds obtained using the conventional method of constant stimuli (MOCS). Across all pilot participants, we observed a consistent overestimation of thresholds predicted by the Wishart model fit to the AEPsych trials compared to those predicted by Weibull psychometric functions fit to the MOCS trials. We attributed this discrepancy to the blocked design that we used, and the fact that stimulus uncertainty was different across the two trial types. Specifically, while AEPsych sampled comparison stimuli in many directions around a reference in a block of trials, the MOCS trial blocks were restricted along a single color direction. We think this led to a faster reduction in stimulus uncertainty for MOCS trials, and thus lower estimated thresholds. It is also possible that our 3 × 3 grid of reference stimuli was too coarse.

To address these issues, we have revised the experimental design with the following improvements: (1) increasing the number of reference stimuli tested using MOCS trials, (2) fully interleaving the two trial types within each session, and (3) removing the fixed grid of reference stimulus for AEPsych-based trials. Instead, both the reference and comparison stimuli will now be treated as free parameters and determined adaptively by AEPsych, effectively turning the task into one where we determine performance over a 4D psychometric field. As outlined in the first preregistration document, our long-term objective is to extend this method beyond the isoluminant plane to estimate the full 3D color discrimination ellipsoids across the color gamut.

This document outlines the updated experimental design, including the dual-computer setup that dedicates one system to computing trial placement while the other handles stimulus presentation. We also describe the fallback trial strategy aimed at minimizing trial delays and introduce new analyses for evaluating trial efficiency.

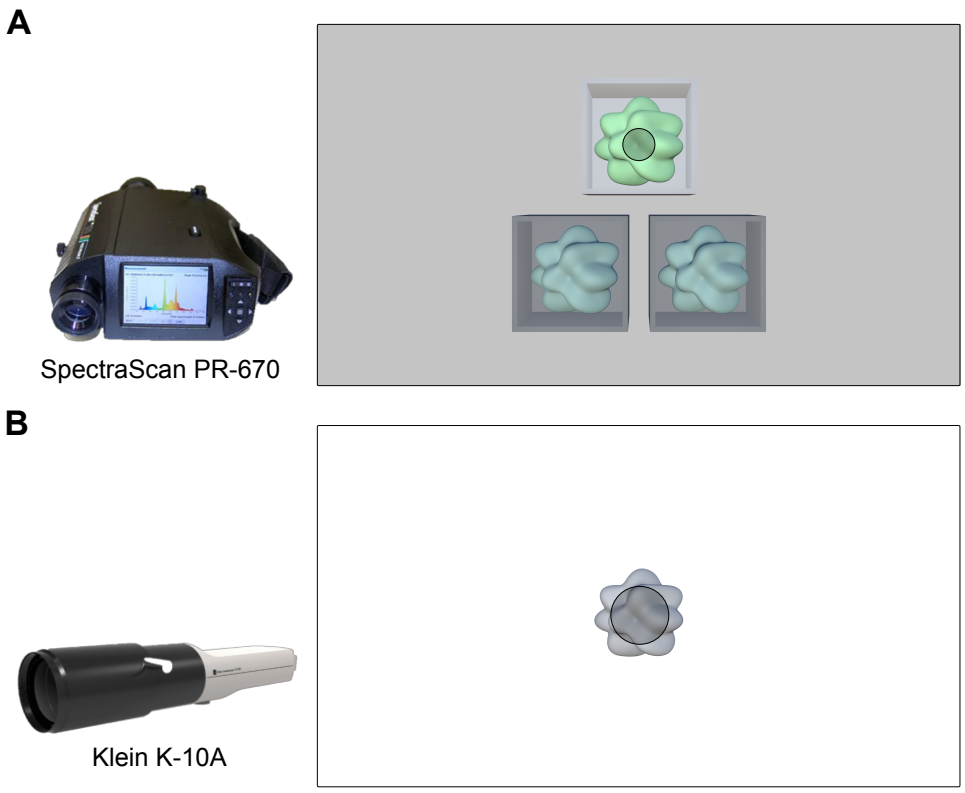


**Figure 1. Stimuli and equipment.** (A) The SpectraScan PR-670 radiometer was positioned at the same distance as the participant's chinrest to replicate experimental viewing conditions. The shaded gray circular region on the object indicates the measurement area captured by the radiometer's lens. Calibration Task 1: The surface color of the cubic room and the blobby stimulus (illustrated as the top stimulus) varied across trials for color calibration. This process was repeated for each blobby stimulus to ensure consistency across screen locations. (B) The Klein K-10A colorimeter was placed in close proximity to the stimulus to measure luminance increments. Calibration Task 2: The stimulus was positioned at the center of the screen, with the cubic room removed and the background set to white. This configuration facilitated the extraction of stimulus pixels from saved images for further analysis.

# Methods

## Apparatus

The experiment will be conducted using an Alienware computer running Windows 10 Enterprise, equipped with Intel® Core™ i7-10700K processor and NVIDIA GeForce RTX 3080 GPU. The display is a DELL U2723QE monitor (59.8 cm width, 33.6 cm height, 3840 x 2160 resolution, 60 Hz refresh rate, achieving 10-bit color depth via 8-bit plus frame rate control). The monitor will be positioned 189 cm from the chinrest, subtending a visual angle of 17.98 x 10.16 degrees of visual angle (dva). Monitor color and luminance measurements were obtained with a Klein K-10A colorimeter and a SpectraScan PR-670 radiometer (**Fig. 1A**). The pixel resolution of the display is approximately 200 pixels/dva, above the typical human foveal resolution limit.

A key change in the apparatus compared to the previous version is the separation of trial placement computation from stimulus display. Instead of using the Alienware computer for both tasks, we will dedicate it solely to displaying stimuli, while trial placement will be handled by a MacStudio running Sequoia 15.3.1, equipped with an Apple M1 Ultra chip with 20 CPU, 50 GPU and 128 GB memory. The two computers will communicate via a shared network disk, using a text file that they both have access to. More specifically, the MacStudio will run AEPsych to

compute trial placement and write the RGB values of the stimulus to the text file, which the Alienware will then read to present the stimuli. This setup allows the MacStudio to take advantage of the 2.9-second non-stimulus-presentation interval between two consecutive trials, significantly reducing wait times and ensuring a smoother experimental flow (see **Procedure** for details).

Additional modifications to the setup include a USB speaker (3 Watts output power, 20k Hz frequency response) for playing auditory feedback, and a gamepad controller (Logitech Gamepad F310) for registering trial-by-trial responses.

**Stimulus**

The stimulus is a blobby 3D object created in Blender with a matte, non-reflective surface. The scene consists of three of these blobby stimuli positioned in a triangular arrangement: one at the top, one at the bottom left, and one at the bottom right (**Fig. 1B**) on a gray background (x = 0.306, y = 0.326, Y = 116.8 cd/m2). Each blobby stimulus (2.49 x 2.49 dva; 203.9k pixels out of 8,294.4k pixels) is centered and floating within its own cubic room (3.27 x 3.27 dva; x = 0.302, y = 0.322, Y = 66.08 cd/m2), which is illuminated by white spot light with a 180 degree beam angle set to maximum intensity. Rendering is performed using Unity's standard shader, with color adjustments applied by modifying the material's texture. Note that in **Fig. 1B**, the cubic rooms surrounding the three stimuli are shown with different colors as part of the calibration procedure (see **Calibration** for details). However, in the actual experiment, all cubic rooms will have identical colors to ensure consistency across stimuli (see **Procedure** for details).

**Calibration**

The stimuli used for calibration matched those used in the experiment (**Fig. 1**). Two calibration tasks were conducted: (1) using the SpectraScan PR-670 to calibrate each blobby stimulus at the location it will be presented (**Fig. 1A**) and (2) using the Klein K-10A to measure the precision (quantization) of our display pipeline (**Fig. 1B**).

In the first calibration task, we verified several key aspects, including the monitor's gamma function as driven through Unity (sampled in 61 evenly spaced steps from 0 to 1), primary spectral power distributions and their stability, primary chromaticities and their stability, linearity, additivity, as well as the effect of background on the spectral power distribution of the target stimulus (**Fig. S1**). We then repeated the calibration task on the other two blobby stimulus locations, verifying the consistency across screen locations (**Fig. S2**). As a result, the same gamma calibration (primary spectra, gamma correction) will be applied to all three objects during the main experiment, using the data from the bottom right blobby stimulus. Specifically, we interpolated gamma table for 4,096 RGB values using a combination of linear and polynomial fits, which were then used to derive the inverse gamma function for gamma correction in Unity (**Fig. 2A**). To validate this gamma correction, we repeated color calibration with the correction applied in Unity; results showed excellent alignment with the identity line (**Fig. 2B**).

In the second calibration task, we used the Klein K-10A to confirm that the output color depth achieved smooth increments via Unity's standard shader with its inherent and implicit spatial dithering. We tested RGB values within the range of 511/1023 to 541/1023, with an increment of 1/1023. Each stimulus was displayed for 5 seconds, and the RGB settings from the first frame of the frame buffer were saved in EXR format. We then compared the average RGB values from the EXR files across the blobby object to the luminance (cd/m²) measured over the

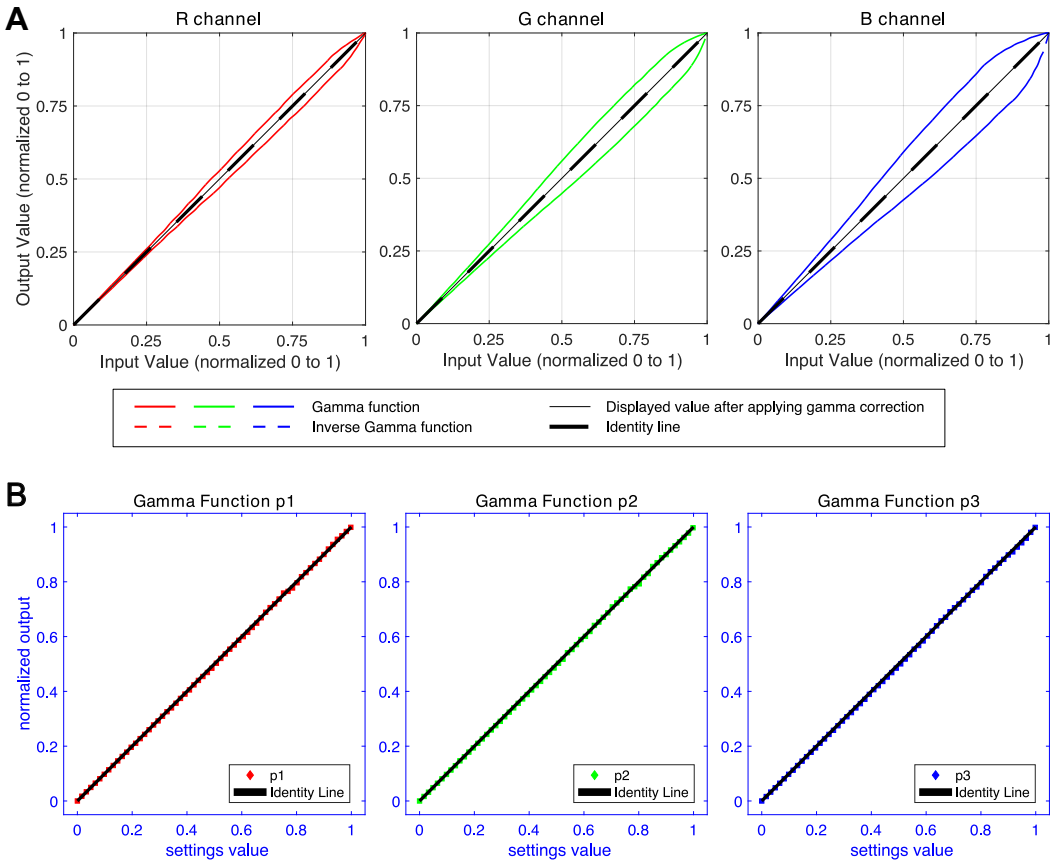


**Figure 2. Gamma correction.** (A) Gamma functions are first interpolated at 4096 levels (solid lines), which are then used to derive inverse gamma functions (dashed lines). The theoretically expected RGB values post-gamma correction align closely with the identity line. (B) Gamma functions measured after applying gamma correction, showing near-perfect alignment with the identity line across all primary colors.

entire stimulus presentation. Results confirmed that Unity achieved at least 10-bit depth (**Fig. 3**). This level of quantization was consistently observed regardless of whether the stimuli's color properties were modified via surface texture (**Fig. 3**) or surface color (**Fig. S3**). For this study, we will modify the material's texture for color adjustments.

To ensure the gamma function remains stable over time, we will repeat this calibration monthly throughout data collection. Calibration results will be compared to the initial calibration, and the gamma correction will be updated only if noticeable shifts in the gamma curves or other device properties are observed.

## Design

This updated experimental design differs from the previous version in three key aspects.

First, reference stimuli are no longer fixed to a 3 × 3 grid, which previously treated the task as an interleaved 2D psychometric field characterization problem, with one such problem for each reference. Instead, the new design dynamically samples reference stimuli bounded between -0.75 and 0.75 in the Wishart space, which itself is bounded between -1 and 1. Specifically, rather than relying on a fixed grid, we use AEPsych to determine both the reference and comparison stimuli adaptively. Comparison stimuli are generated by adding a delta value along each of the two dimensions, with delta values constrained between -0.25 and 0.25 to

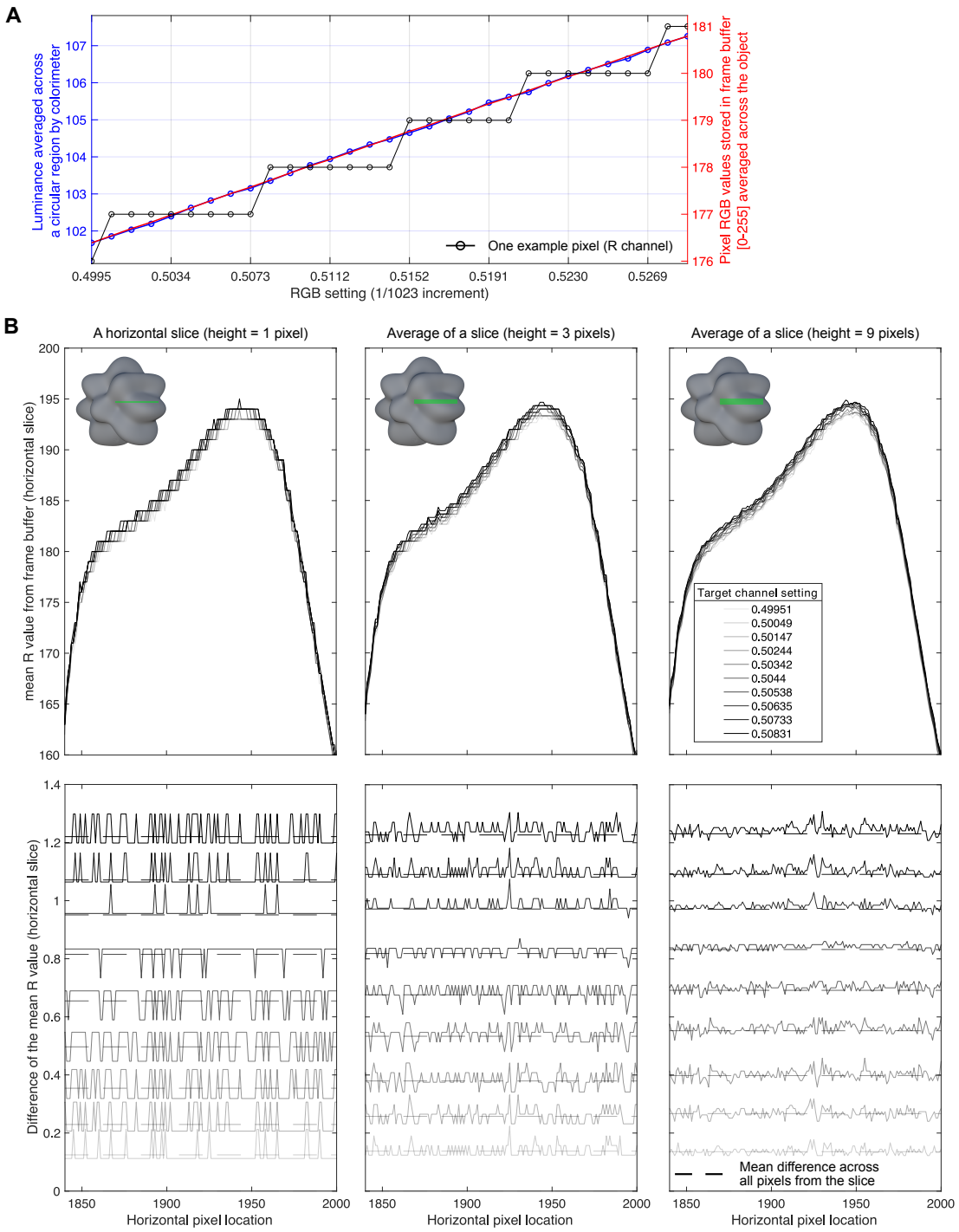


**Figure 3. Evidence of spatial dithering by Unity's standard shader when the surface texture of the stimulus is being modified.** (A) Spatial dithering by Unity's standard shader is suggested by comparing the luminance measurements from the Klein K-10A (averaged across a circular region on the blobby 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.

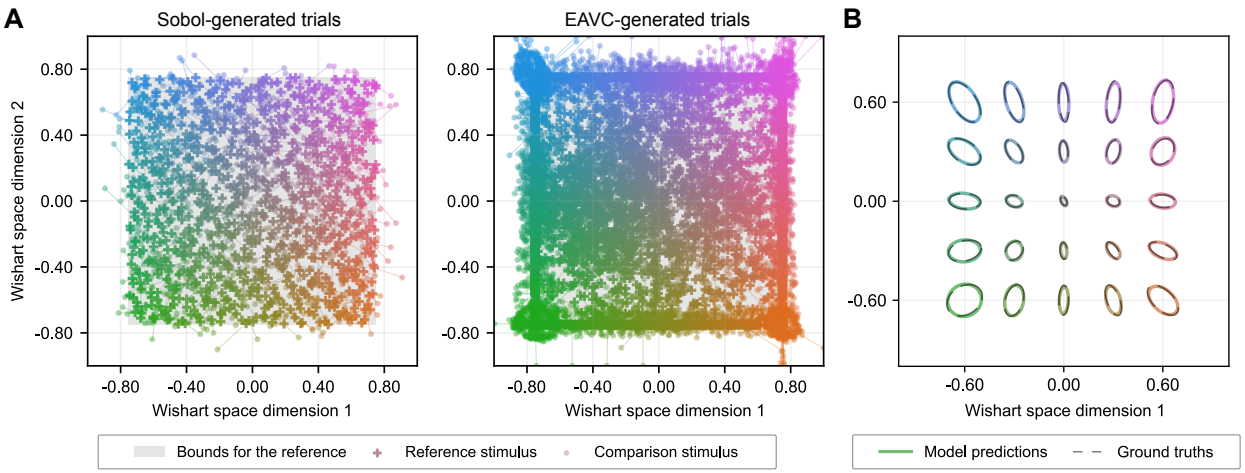


**Figure 4. Trial placement guided by AEPsych and Wishart model predictions.** (A) Sobol-generated trials (N = 900) and EAVC-generated trials (N = 5,100). The reference stimulus is sampled within the range [-0.75, 0.75], while the delta values are constrained within [-0.25, 0.25], ensuring that both reference and comparison stimuli remain within the bounds of [-1, 1]. Trials are simulated based on a ground-truth Wishart model fitted to pilot data from participant S4. (B) Wishart model predictions. Note that the axis scale in B is different from A.

ensure that stimuli remain within the [-1, 1] boundary. By varying both the reference stimuli (2D) and the delta values (2D), the task is now effectively a 4D psychometric field characterization problem. Participants will complete 6,000 AEPsych-based trials, consisting of 900 Sobol-sampled trials and 5,100 trials using expected absolute volume change (EAVC) sampling. The Sobol-sampled trials will be scaled by one of three factors (1/4, 2/4, 3/4), with an equal distribution across these scalers. Smaller scaling factors increase task difficulty, and to counterbalance this effect, the scalers will be pseudo-randomized in the first session. The efficacy of this 4D thresholding task was validated through simulations using a ground-truth Wishart model fit to a pilot dataset, which confirmed that 6,000 trials, with 900-5100 Sobol-EAVC split, are sufficient to reliably recover color discrimination thresholds (**Fig. 4, Fig. S4**).

The second key change is the selection of MOCS conditions. Previously, eight conditions were selected at four fixed reference locations, with each location including two conditions—one aligned with the DKL L-M axis and the other with the DKL S axis. The updated design increases the number of MOCS conditions to 25, with one reference always set at the achromatic color. The other 24 reference locations will be Sobol-generated within the plane bounded between [-0.6, 0.6], while the chromatic direction in which the comparison stimulus will vary along will also be Sobol-generated between 0 and 360 degrees (**Fig. 5A-B**). We will use different random-number-generator seeds unique to each participant (**Table S1**). Each MOCS condition consists of 12 levels, with each level repeated 20 times, leading to 6,000 total MOCS trials. The stimulus levels were determined by identifying the comparison stimulus along the chromatic direction that corresponds to 95% correct performance, as determined by the Wishart fit illustrated in **Fig. 4**. Twelve levels were then evenly spaced along this direction, excluding the level that matches the reference stimulus exactly. Additionally, an easy level was included by selecting a comparison stimulus positioned at twice the distance of the 95% threshold stimulus. This level will serve as catch trials. Simulations confirmed that this design maintains a strong agreement between Wishart model-predicted thresholds and those estimated using Weibull psychometric functions (**Fig. 5C**).

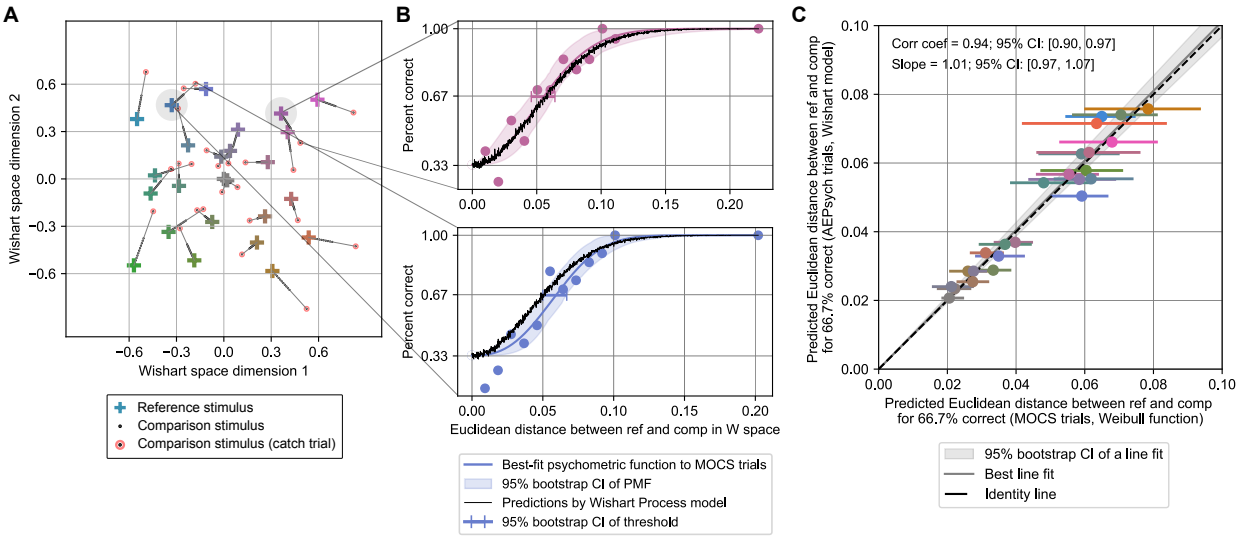


**Figure 5. Selection of MOCS conditions and comparison of thresholds predicted by the Wishart model and Weibull psychometric functions.** (A) Distribution of 25 MOCS conditions, sampled by a Sobol-generated reference stimulus and a corresponding chromatic direction along which comparison stimuli are varied. One of the reference stimuli is forced to be at the achromatic color. (B) Best-fit psychometric functions for two selected MOCS conditions. Horizontal axis: the Euclidean distance between the reference and its corresponding comparison stimuli in the Wishart space; vertical axis: the average percent correct after combining 20 trials per level. Top panel: An example where the threshold predicted by the Wishart model closely matches the threshold estimated from the Weibull psychometric function fit to MOCS trial data. Bottom panel: A case where the two threshold estimates diverge, with the 95% confidence interval of the Weibull-predicted threshold falling outside the Wishart-predicted threshold. (C) Threshold comparison across all 25 MOCS conditions. The solid line represents a linear fit to the data.

The third major change involves the interleaving of AEPsych and MOCS trials. In the previous design, there were 9 AEPsych conditions and 8 MOCS conditions, each with a fixed reference color. While conditions were shuffled across sessions, trial types were not shuffled within a session. In the updated design, AEPsych and MOCS trials will be fully interleaved both across and within sessions. The MOCS trials will be pseudo-randomized, ensuring that every 300 MOCS trials contain all unique MOCS trial types. Since there are equal numbers of AEPsych and MOCS trials (6,000 each), they will be paired and randomly shuffled within each pair. While the experiment is designed to run 12,000 trials in a pre-determined random sequence, AEPsych sometimes takes longer to compute the next trial placement. To avoid keeping participants waiting, we will implement a fallback trial strategy: if AEPsych exceeds the maximum wait time (2.9 seconds; explained in **Procedure**), the next MOCS trial in the list will be bumped up so that stimulus presentation continues. If AEPsych is still not ready, we continue bumping MOCS trials forward until it completes its computation. While this fallback trial strategy may slightly disrupt the intended intermixing of trial types, simulations confirmed that the effect is negligible, as the average number of MOCS trials per session remains within small deviations of the intended count (**Fig. 6**). These results are based on simulations and were generated on a different computer than the one used for the experiments. However, we expect similar or better performance since the experimental computer has more computing power.

We will again restrict our stimuli (both reference and comparison) to be on the isoluminant plane in the DKL space (**Fig. 7A-B**). Stimuli are presented in RGB space (**Fig. 7C-D**), while trial placement and model fitting are conducted in Wishart space for mathematical

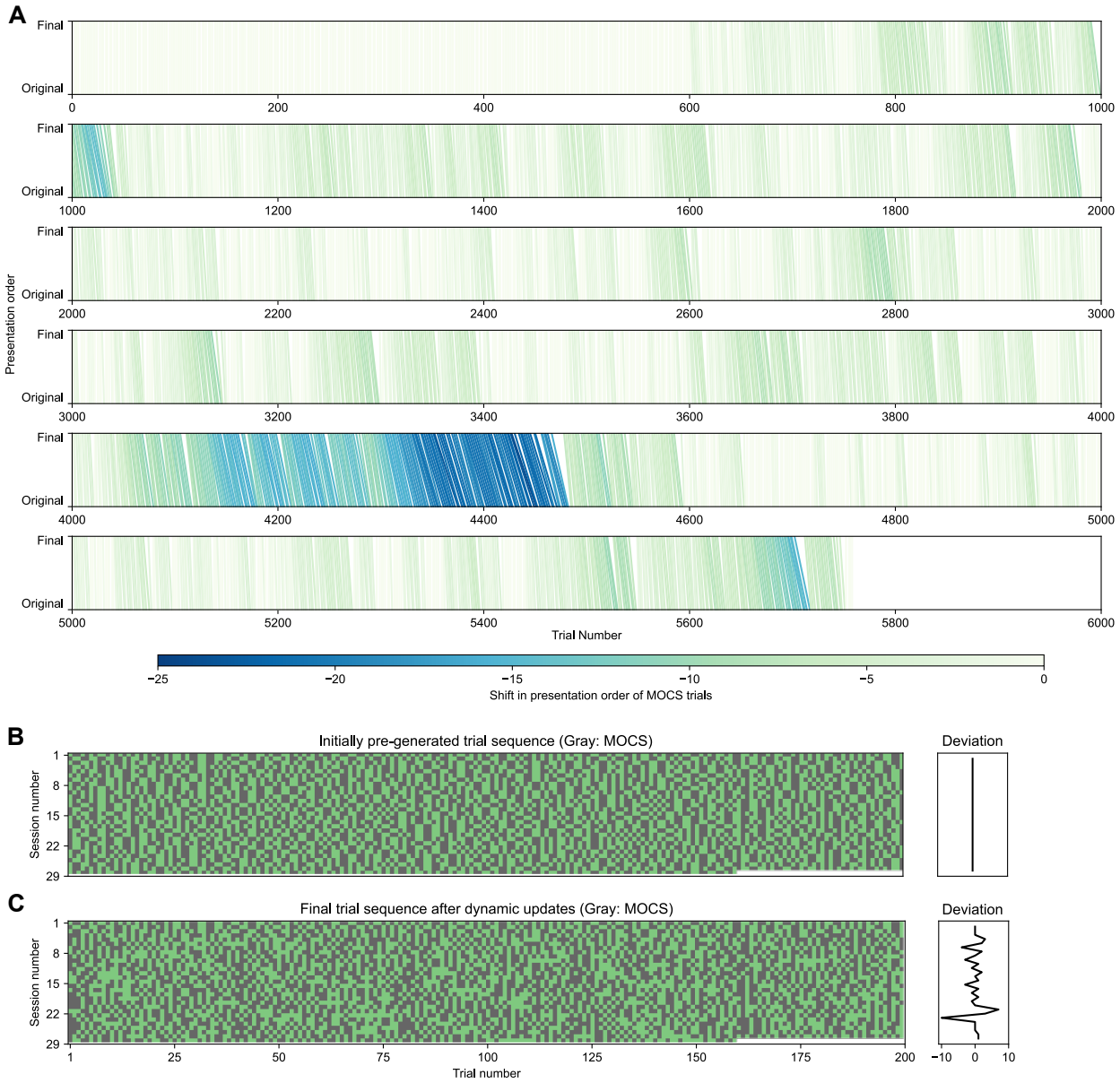


**Figure 6. The impact of the fallback trial strategy on trial intermixing.** (A) Comparison of the original trial placement for MOCS trials versus their final placement after implementing the fallback trial strategy. Vertical lines indicate MOCS trials presented at their pre-determined trial numbers, while left-tilted lines indicate MOCS trials that were shifted earlier due to AEPsych exceeding the 2.9s computation deadline. (B) Pre-determined trial sequence, where all AEPsych and MOCS trials are paired and shuffled within each pair. If AEPsych consistently meets the 2.9s deadline, the total number of MOCS trials presented per session should match the intended count. (C) Actual trial sequence after applying the fallback trial strategy. The deviation between the intended and actual number of MOCS trials per session remains within 10 trials, demonstrating the controlled impact of trial rescheduling.

convenience, as it aligns with the Chebyshev basis functions used in the model (**Fig. 7E**). Consequently, transformation matrices are required to convert between these spaces. To ensure that all stimuli remain within the monitor's gamut, we first determined the boundaries of the isoluminant plane by identifying the maximum extent before exceeding the monitor's color gamut. These boundaries form a 2D parallelogram, corresponding to a sliced plane in RGB space. This allowed us to compute a homography matrix that maps points from the 2D

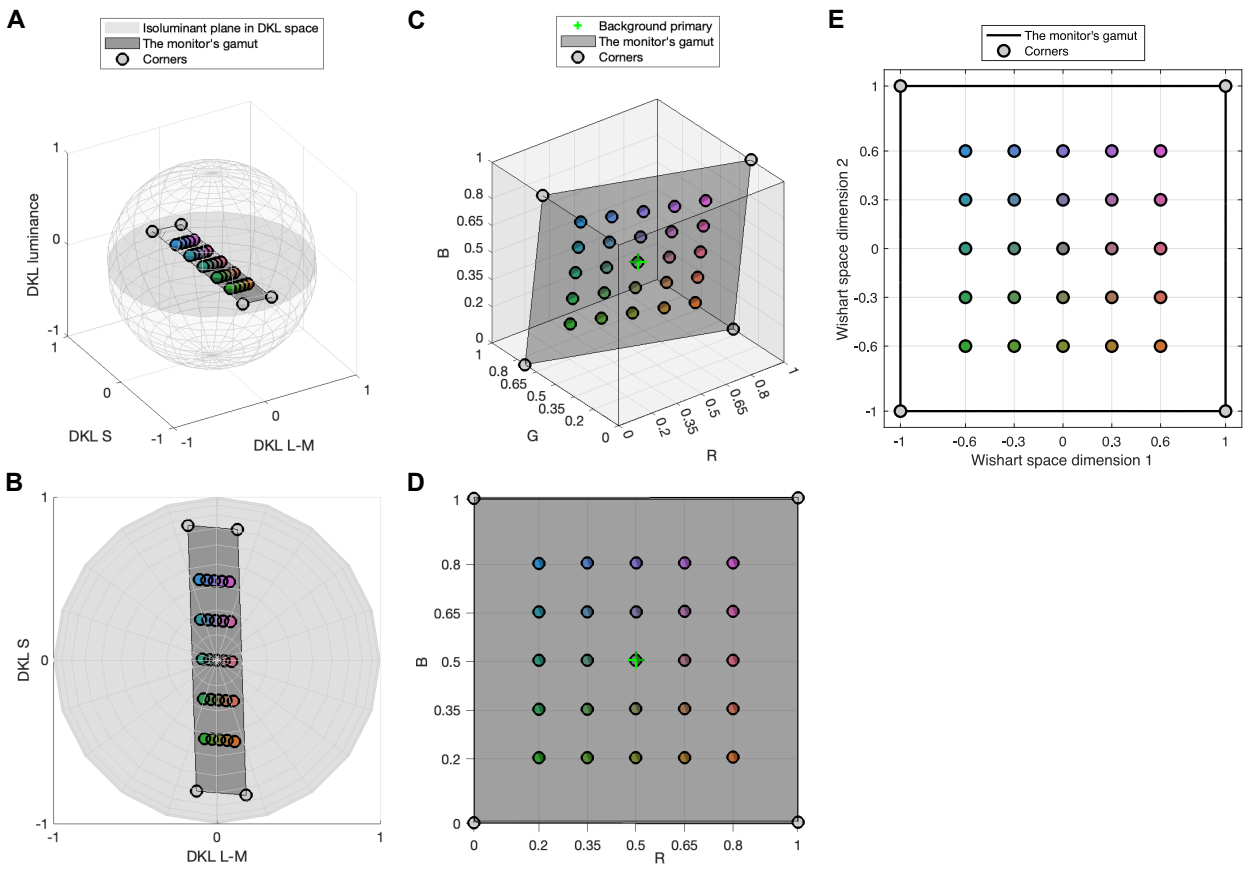


**Figure 7. Mapping stimuli across Wishart, RGB, and DKL color spaces.** A 5 × 5 grid is shown for illustrative purposes; in the actual experiment, reference stimuli are not constrained to a grid. (A) Transformation of stimuli into DKL space, with reference stimuli mapped onto an isoluminant plane in DKL color space. (B) Projection of DKL stimuli onto the isoluminant plane. (C) Stimuli represented in RGB space, with values normalized between 0 and 1. (D) Projection of the RGB stimuli onto the RB plane. (E) Stimuli in Wishart space, bounded between -1 and 1.

isoluminant plane in DKL space to the corresponding 3D parallelogram in RGB space. Additionally, we derived a projection matrix that transforms the 3D parallelogram in RGB space into the 2D Wishart space. The transformation scheme across spaces remains the same as in the previous experimental design. However, the gamut of the isoluminant plane is dependent on the monitor's spectral properties, which we remeasured for this preregistered study. Based on these updated measurements, we recomputed the transformation matrices to ensure accurate color space conversions.

**Procedure**

Participants will perform a three-alternative forced-choice (3AFC) oddity task (**Fig. 8A**). Each trial begins with a fixation cross displayed in the center of the three cubic rooms for 0.5 seconds, followed by a blank screen for 0.2 seconds. Then, three blobby stimuli will appear in the middle of the cubic rooms for 1 second. After the stimulus, a response probe ("< ^ >" indicating the three possible responses) will appear, prompting participants to determine which one is the odd stimulus, with no time constraint. Once participants make a response using a gamepad controller, a blank screen will appear for 0.2 seconds, followed by visual and auditory feedback

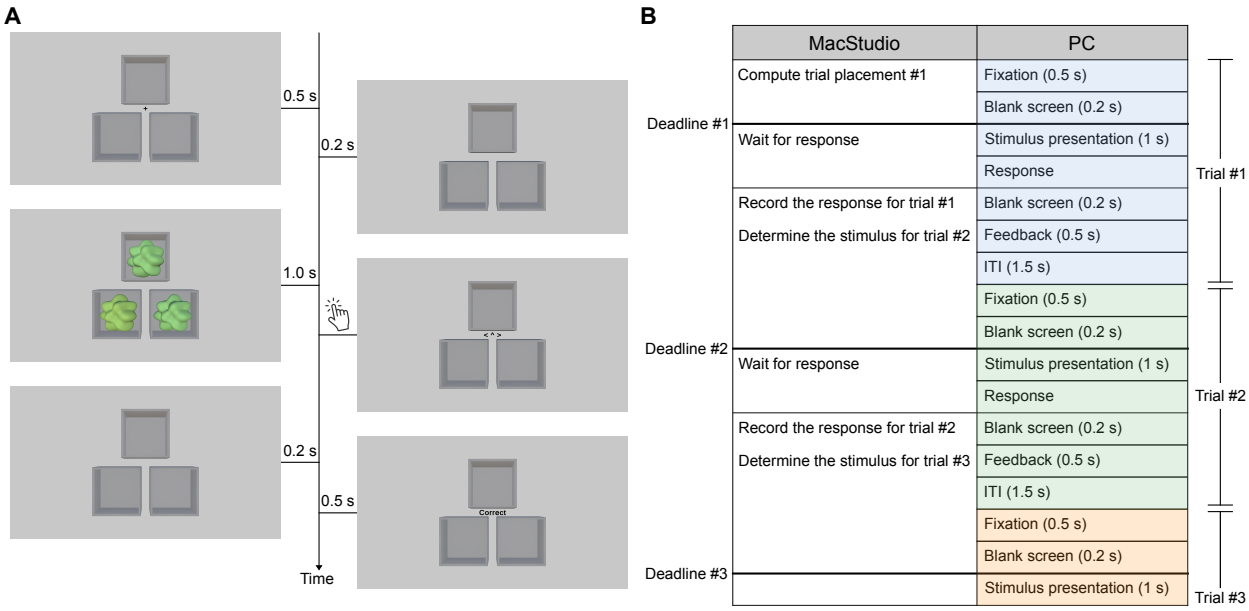


**Figure 8. Task timing and interleaving AEPsych and MOCS.** (A) Each trial begins with a 0.5-second fixation period, marked by a cross at the center of the screen, followed by a 0.2-second blank interval. Next, three blobby stimuli are presented in the cubic rooms for 1 second. After the stimuli presentation, the screen shows a response prompt, instructing participants to identify the odd-one-out stimulus. Participants have unlimited time to respond by selecting the correct option. Once a response is made, a brief 0.2-second blank screen is displayed, followed by a 0.5-second feedback. (B) A schematic representation of the trial timing and computational responsibilities of the two computers. The Mac Studio determines the next trial placement before a 2.9-second deadline (0.7s for trial #1). If it misses this deadline, it will continue to compute and try to meet the next available one. Meanwhile, the PC listens for the RGB values assigned by the Mac Studio and displays the corresponding stimuli.

on accuracy—“correct” or “incorrect,” accompanied by a beep or buzz tone. The inter-trial interval will be 1.5 seconds. Participants will be instructed to move their eyes during stimulus presentation and try to fixate on each object.

The main experiment will consist of 10 sessions. In each session, participants will begin with 40 practice trials to familiarize themselves with the task. Practice trials will not be analyzed. Each session will consist 1,200 trials, and participants will take a break every 200 trials, resulting in a total of 5 breaks during each session. Each session is expected to take approximately 1.5 hours to complete. Data from complete blocks (200 trials) in a session will be retained. Data from any partial blocks will be discarded.

As described above, we implemented a procedure to minimize subject wait time between trials when AEPsych calculations take longer on a particular trial. A detailed unpacking of the trial timing sequence is shown in **Fig. 8B**. The design allows at least 2.9 seconds for AEPsych to compute the next trial once the Mac Studio receives the participant’s response from the PC (the Alienware computer). This interval includes a post-stimulus period from the current trial (0.2-second blank screen, 0.5-second feedback display, 1.5-second inter-trial interval) and a pre-stimulus period from the next trial (0.5-second fixation, and another 0.2-second blank screen). If AEPsych does not complete its computation within this window, the next available MOCS trial from the pre-determined trial sequence will be presented instead. However, AEPsych’s computation will not be interrupted; it will continue running and attempt to meet the next available deadline. If the first deadline (2.9 seconds) is missed, the second deadline extends to approximately 7 seconds, factoring in the previous interval, a 1-second stimulus

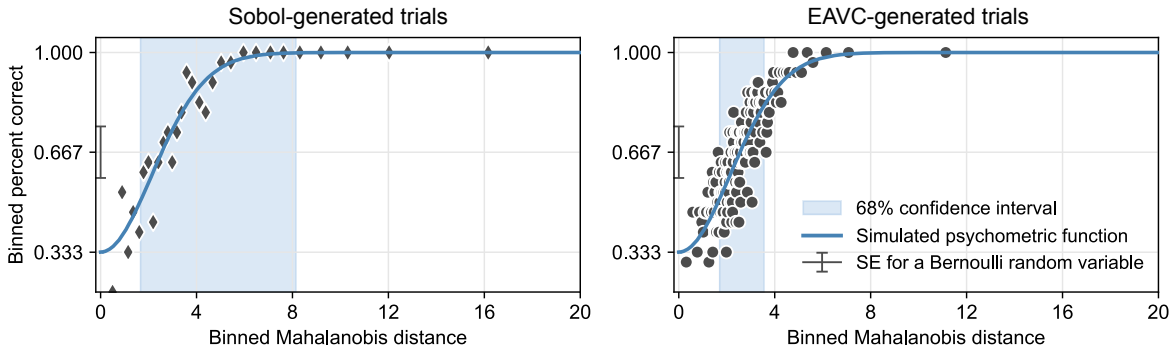


**Figure 9. Evaluation of the efficiency of trial placement determined by AEpsych.** Percent correct as a function of Mahalanobis distance. Trials are first sorted by Mahalanobis distance and then averaged in bins of 30 trials. The Mahalanobis distance is computed based on the Wishart model fit for each reference-comparison stimulus pair. Shaded region: the central 68% of data points; blue solid line: predictions derived using unit circles.

presentation, and an estimated 0.2-second response time. If AEPsych also misses this deadline, the next opportunity arises at around 11.1 seconds, incorporating an additional stimulus presentation and response time. This staggered scheduling ensures that trials continue smoothly while allowing AEPsych sufficient time to compute adaptive placements when possible.

## Data analysis

All analyses of color calibration data were conducted in MATLAB 2023b. The calibration scenes and the experiment will be created and run in Unity 2022.3.24f1, programmed in C#, and behavioral data will be analyzed in Python 3.11.

During color calibration, gamma functions were measured to compute the inverse gamma lookup table and the transformation matrix needed for converting Wishart space values into RGB values. These tables are then exported in CSV format to Unity, enabling gamma correction and the conversion of AEPsych-determined value pairs into RGB values.

For behavioral data analysis, we will first separate each participant's data into AEPsych trials and MOCS trials. For evaluation of the Wishart process model approach, the Wishart process model will be fitted exclusively to the AEPsych trials. For each set of MOCS trials, we will fit a Weibull psychometric function with a fixed guess rate of 1/3, estimate the stimulus level corresponding to the threshold (defined as 66.7% correct performance), and perform 120 group-wise bootstraps to compute a 95% confidence interval for the threshold level (**Fig. 5B**). This procedure will be repeated for all MOCS conditions, after which we will fit a straight line fixed at the origin to each bootstrapped dataset. This will yield a confidence interval on the slope (**Fig. 5C**). If the confidence interval includes 1, it suggests good agreement between the thresholds predicted by the Wishart model (fitted to AEPsych trials) and those derived from Weibull psychometric functions (fitted to MOCS trials).

To assess the efficiency of AEPsych for trial placement, we will compute the Mahalanobis distance between the reference and comparison stimuli for each trial, sort the distances, and compute the average across every 30 trials (**Fig. 9**). As a reference, we will compare these distances to those of two unit circles separated by varying distances. Efficient trial placement should result in all binned Mahalanobis distances clustering around the rising part of the reference curve. To evaluate model performance, we will compare the Wishart model

with an alternative model that assumes independent threshold contours across reference stimuli, which does not assume a smoothly varying threshold contour in the behavioral data. Model comparison will be conducted using 10-fold cross-validation, and Bayes factors will be computed as a performance metric.

We will likely perform additional analyses on the dataset. These may include estimates of the precision of the Wishart model fits, including but not limited to consideration of how many trials are needed for good Wishart model fit reliability, comparison of data across subjects, comparison to published color discrimination data and color metrics, and computation of geodesics within the isoluminant plane. For some of these purposes, we may fit the Wishart process model to all data collected, both AEPsych and MOCS trials.

**Participant eligibility, enrollment and exclusion criteria**

The study will recruit paid volunteers, compensated at a rate of $20 per hour. Participants will be drawn primarily from the University community and obtained through world-of-mouth or other informal advertisement. The principal investigator, post-docs, graduate students and research specialists in our lab may also participant as unpaid volunteers. The experiment requires that participants have visual acuity of at least 20/40 in each eye and normal color vision. Visual acuity will be tested using a Snellen eye chart, and color vision will be tested using Ishihara color plates. Subjects will be age 18 or older. One or more of the subjects may have participated in the first experiment of the study.

All participants will consent to participate in the study by reviewing and then signing an IRB approved consent form at the time they are initially enrolled in any of its experiments. Consent will be obtained by lab personnel. The experimental procedures and general purpose of the experiments will be explained to the subject before participation begins, as well as the expected duration of their participation. The experimental procedures do not involve any known risks, and participants will be free to withdraw from the experiment at any time.

To identify participants who are not attentive to the stimuli, we will monitor their performance on easy trials. Easy trials will be the MOCS trials with the RGB values that are most far away from the corresponding reference stimulus. Each session will include 50 such trials. Subjects who get such trials consistently incorrect will be encouraged to pay more attention during the experiment. If a participant consistently gives more than 5 incorrect responses on these trials across sessions, they will be gently withdrawn from the study.

**Order of data collection**

We will begin by collecting a full dataset for one subject, and analyze this data before enrolling additional subjects. Our hope is that data from the first subject will not reveal the need for any major modifications to the experimental protocol, in which case we will include the first subject's data in reports of this experiment. We will post a follow-up preregistration document listing any changes made before running additional subjects, indicating whether the first subject's data will be retained, and specifying the number of subjects to be run in the remainder of the experiment.

# Appendix

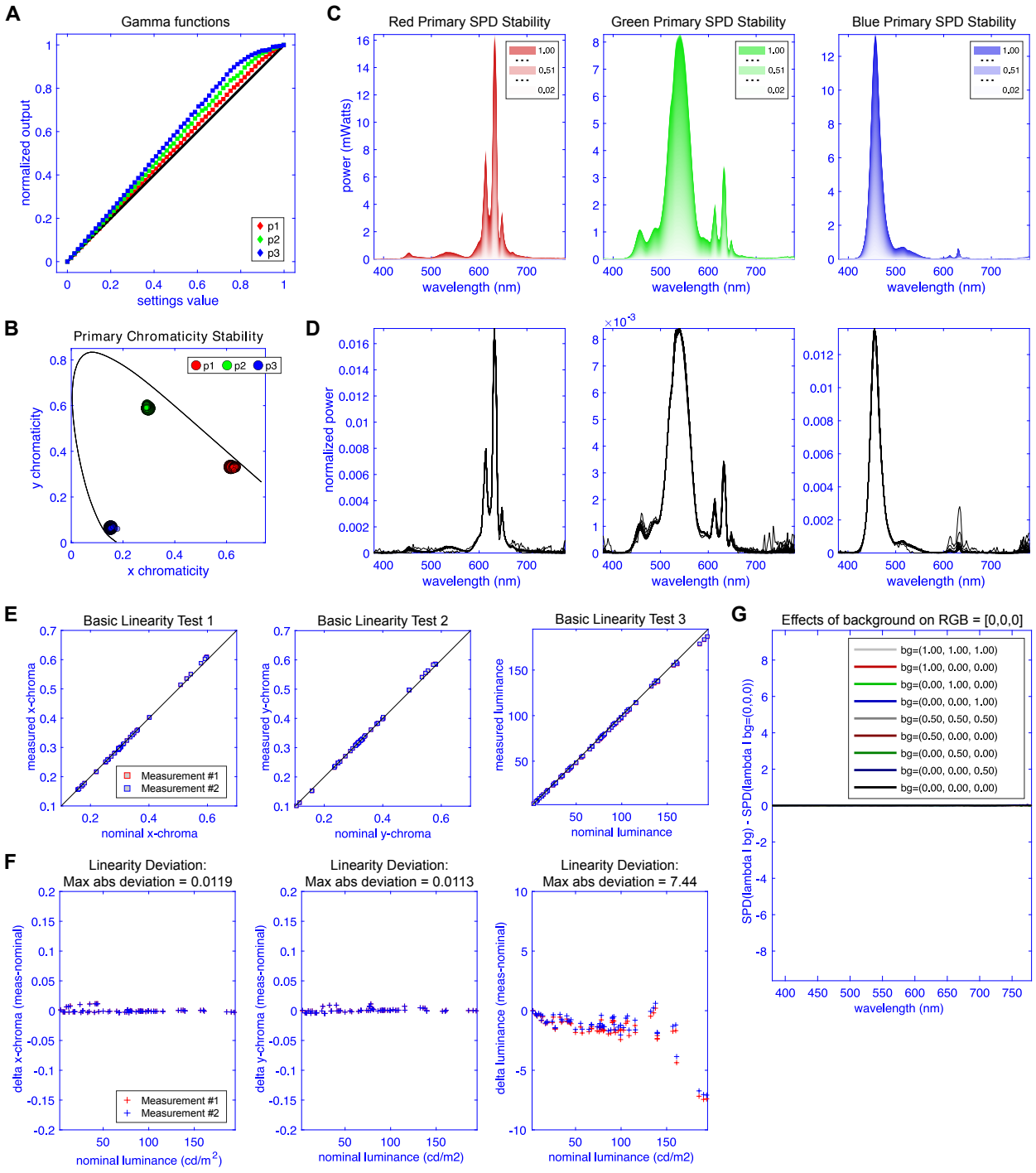


**Figure S1. Color calibration results for the blobby stimulus in the top position of the triangular arrangement**. (A) Gamma functions for the three primary colors (red, green, blue). (B) Chromaticity coordinates of each primary color in CIE color space at different intensity levels. (C) Spectral power distributions (SPDs) of the three primary colors at different intensity levels. (D) Normalized SPDs for each primary. (E) Linearity tests for chromaticity and luminance: comparison of nominal (predicted) and measured chromaticity (x and y) and luminance values across two separate measurements (F) Linearity deviation. (G) The effects of background (the surface color of the cubic room) on the SPD of the blobby stimulus.

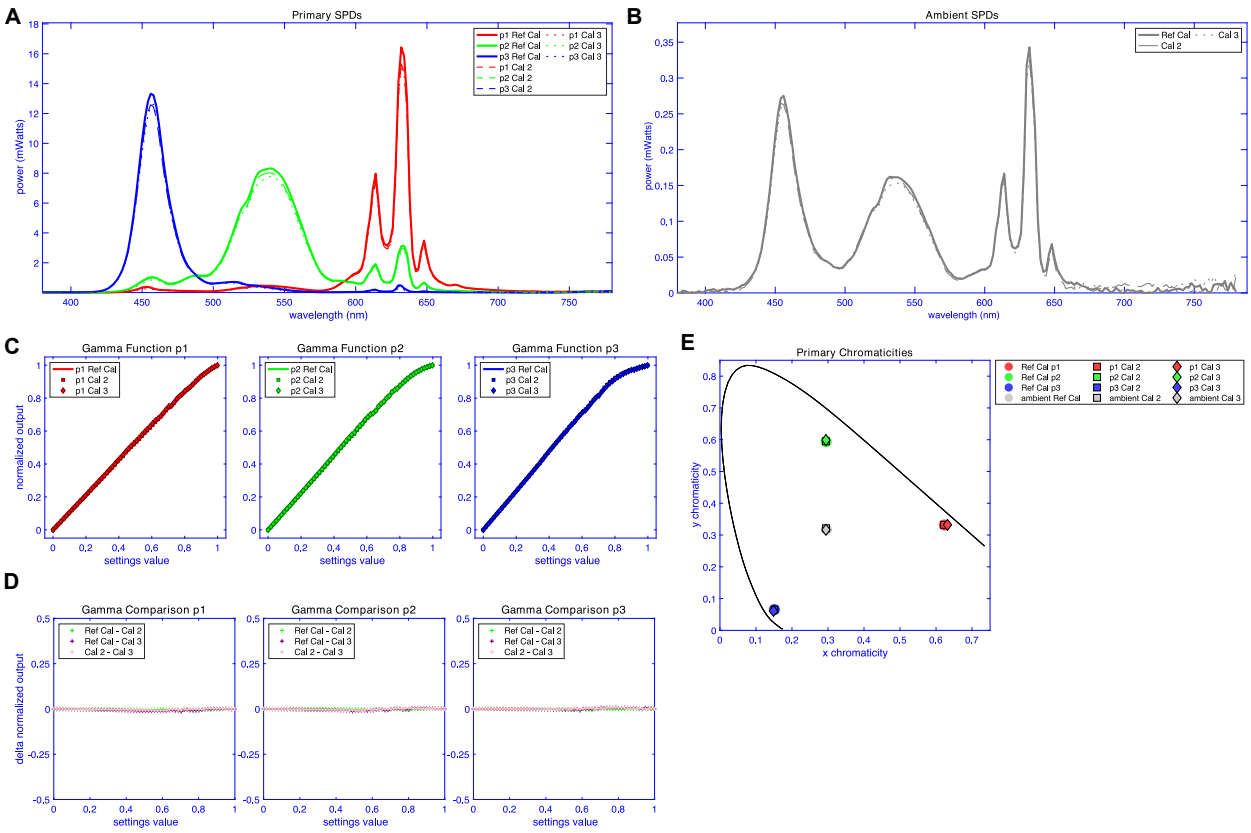


**Figure S2. Comparison of color calibration data across the same object positioned at three different locations on the screen.** (A) The spectral power distributions (SPDs) of the three primary colors (red, green, and blue) across the three calibration locations (Cal 1: right, Cal 2: left, and Cal 3: top). (B) The ambient SPDs. (C) Gamma functions for each primary. (D) The difference of normalized output for three pairwise comparison and for each channel. (E) The chromaticity coordinates for each primary in CIE space. Overall, these results show minimal variation in SPD, gamma functions, and chromaticity across different screen locations, indicating consistent color behavior of the monitor.

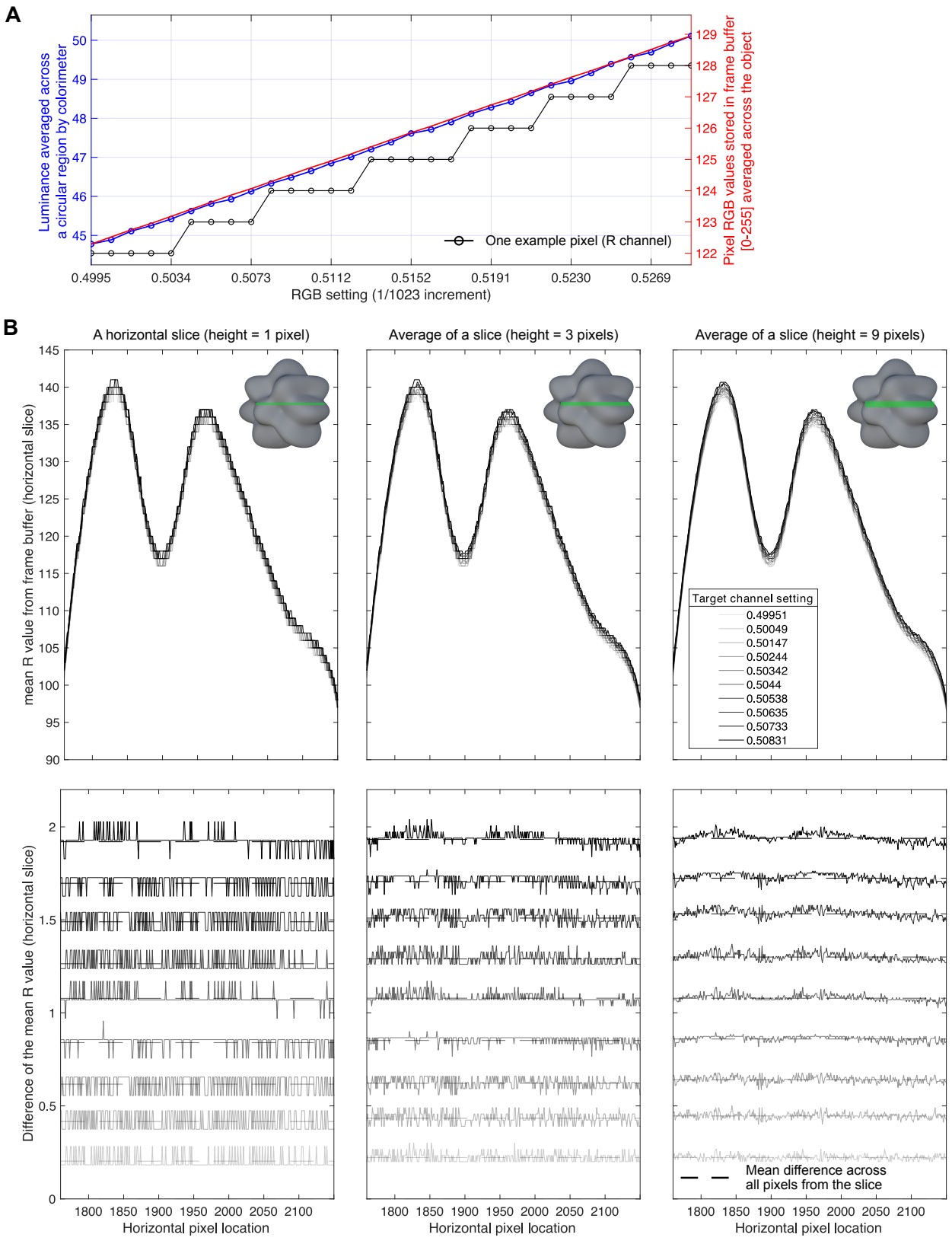


**Figure S3. Evidence of spatial dithering by Unity's standard shader when the surface color of the stimulus is being modified.**

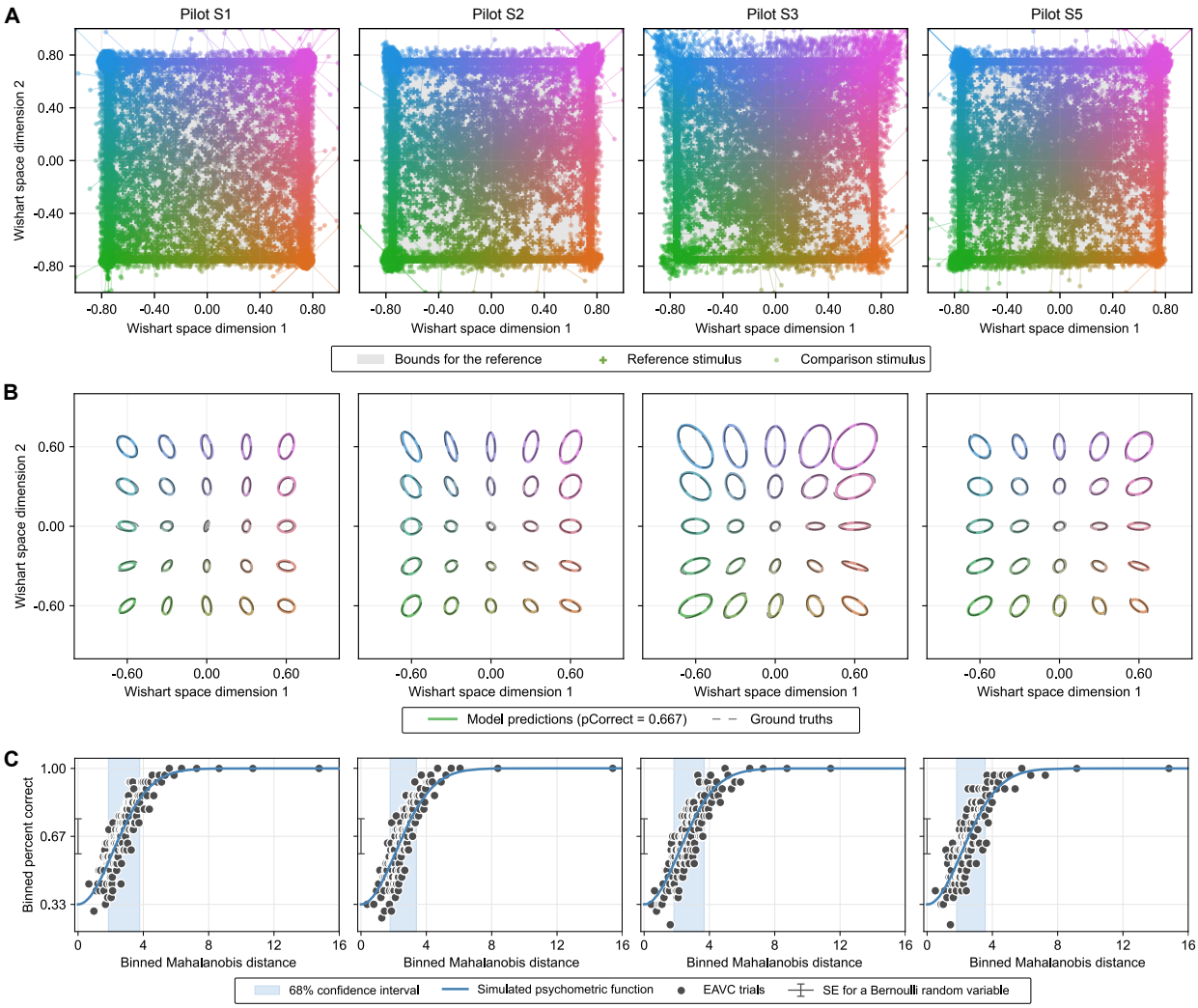


**Figure S4. Trial placement guided by AEPsych and Wishart model predictions**. (A) EAVC-generated trials (N = 5,100) simulated using a ground-truth Wishart model fitted to pilot data from participants S1–S5, except S4. (B) Predicted thresholds based on the Wishart model. Note that the axis scale differs in B from that in A. (C) Percent correct as a function of Mahalanobis distance, computed for each reference-comparison stimulus pair sampled by

| | MOCS | Sobol scaler | Interleaving AEPsych and MOCS trials | Location of the odd stimulus |
|---|---|---|---|---|
| Source | Python | Python | Python | C# |
| Practice | sub# x 1000 | sub# x 1000 | sub# x 1000 + session# | sub# x 1000 + session# |
| Experiment | sub# x 100 | sub# x 100 | sub# x 100 + session# | sub# x 100 + session# |

**Table S1. Scheme for selecting seeds for different subjects from random-number generators.** Four aspects of the experiment require shuffling: (1) the presentation order of the MOCS trials, which consist of 25 conditions, 12 levels, and 20 repetitions per level; (2) the order of Sobol scalers (1/4, 2/4, 3/4) applied to Sobol-generated trials to balance task difficulty; (3) the shuffling sequence between AEPsych and MOCS trials; and (4) the location of the odd stimulus, which can appear at the top, bottom left, or bottom right. The first two aspects are determined only once during the first session, after which subsequent sessions reference the pre-determined indices. In contrast, the shuffling order between AEPsych and MOCS trials, as well as the odd stimulus location, are generated separately for each session.