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], 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 is an initial experiment where we will measure the efficacy of our Wishart process model combined with adaptive sampling method AEPsych, for determining a color discrimination field. This method itself will be described elsewhere, but in short it involves fitting a finite-basis Wishart process model to trial-by-trial forced-choice psychophysical data of the sort 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 a partial field of color discrimination ellipses in the nominal isoluminant plane, and in doing so to evaluate the efficacy of the Wishart process approach. This initial test will be for the central portion (around a white point) of the plane. Although our long-term goal is to extend the method to full isoluminant plane and then to the full field of 3D color discrimination ellipsoids throughout the gamut of color space, we are starting with an initial subset to try out and possibly refine our experimental and analysis methods.

We will proceed by using the Wishart process model to analyze a large set of psychophysical trials, and compare with data collected as interleaved method of constant stimuli (MOCS) trials that will establish a conventional dataset for comparison, for a limited number of reference stimuli and color directions.

In this document, we specify our experimental design, including our initial specification of the number of trials per participant we will need both for the Wishart process model and MOCS.

Since this is our first foray into measurements, we may after examining the data decide to collect more trials to sharpen the estimates for either or both methods. We also note that based on first principles, how well the Wishart process model works may depend on the smoothness of the prior specified, which can be varied. We are likely to explore different choices in the analysis here, after we have examined the data, to get a better understanding of the effect of the prior with real data.

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 3090 GPU. The display is a DELL U2723QE monitor (59.8 cm width, 33.6 cm height, 3480 x 2160 resolution, 60 Hz refresh rate, achieving 10-bit color depth via 8-bit + FRC). The monitor will be positioned 130

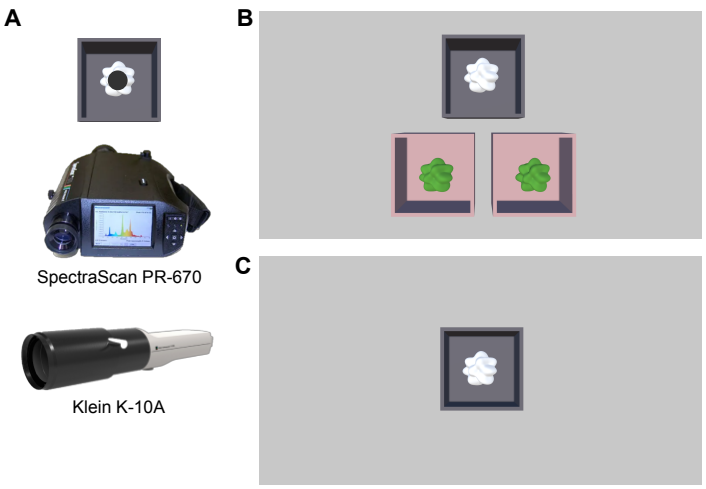

Figure 1. Stimuli and equipments. (A) The SpectraScan PR-670 radiometer was positioned at the same distance as the chinrest to replicate participant viewing conditions. The black region on the object indicates the area measured by the lens of each device. The Klein K-10A colorimeter was placed very close to the stimulus to measure luminance increments. (B) The stimulus for calibration task 1: The surface color of the cubic room and the blobby stimulus (shown here as the top stimulus) varied across trials to allow for color calibration. This task is repeated for each blobby stimulus to ensure consistency in measurements across screen locations. (C) The stimulus for calibration task 2: The stimulus was positioned at the center of the screen.

centimeters from the chinrest, subtending a visual angle of 25.9 x 14.7 degrees of visual angle. Monitor color and luminance measurements were obtained using a Klein K-10A colorimeter and a SpectraScan PR-670 radiometer (**Fig. 1A**). The pixel resolution of the display is approximately 140 pixels/deg, above the typical human foveal resolution limit.

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**). Each blobby stimulus (2.07 x 2.07 dva) is centered and floating within its own cubic room (4.85 x 4.85 dva), which is illuminated by a directional white light set to maximum intensity. The scene is rendered using Unity's standard shader, and color adjustments are applied by modifying the texture of the material. Note that in 1B, the three stimuli are shown with different backgrounds. This is simply to indicate to the reader that we can vary the background, which we may do in future experiments. In the present experiment, the context in which each of the three blobby objects is seen will be the same as each other.

Calibration

The stimuli used for calibration matched those used in the experiment (**Fig. 1B**). Three calibration tasks were conducted: (1) using the SpectraScan PR-670 to calibrate each blobby stimulus at the location it will be presented, (2) repeating task 1 with a single blobby 3D stimulus positioned at the center of the screen (**Fig. 1C**), and (3) using the Klein K-10A to measure luminance the precision (quantization) of our display pipeline, using the same scene setup as in task 2. In the first of these, the same contextual background was used at all three locations, and this background matched that used for the measurements at the center of the screen.

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. 2**). We then repeated the calibration task on the other two blobby stimulus location, verifying the consistency across screen locations (**Fig. 3**). As a result, the same

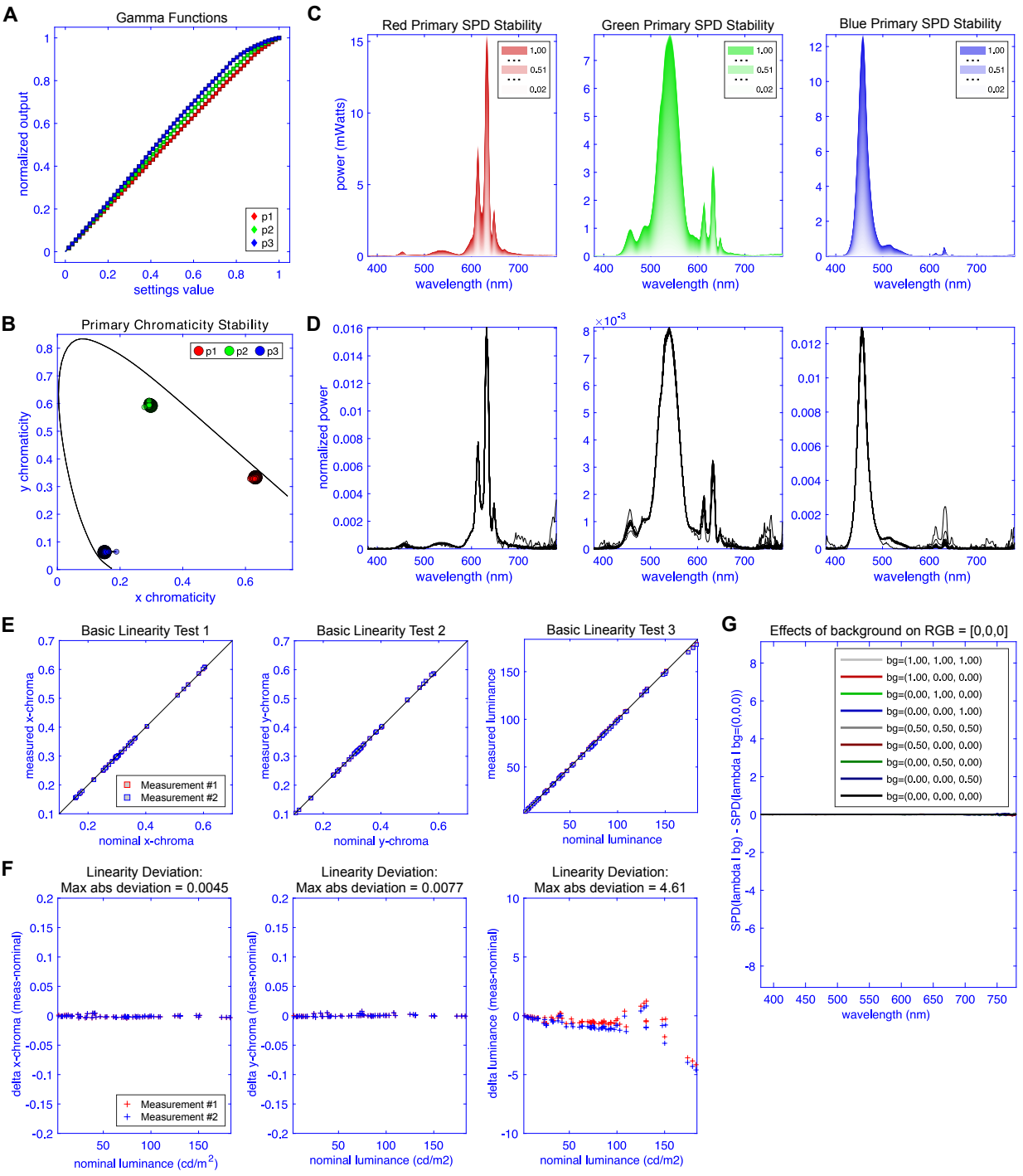

Figure 2. Color calibration results for the blobby stimulus in the top position of the triangular arrangement. (A) Gamma functions for the three primary colors (red, green, blue). (B) Chromaticity coordinates of each primary color in CIE color space at different intensity levels. (C) Spectral power distributions (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.

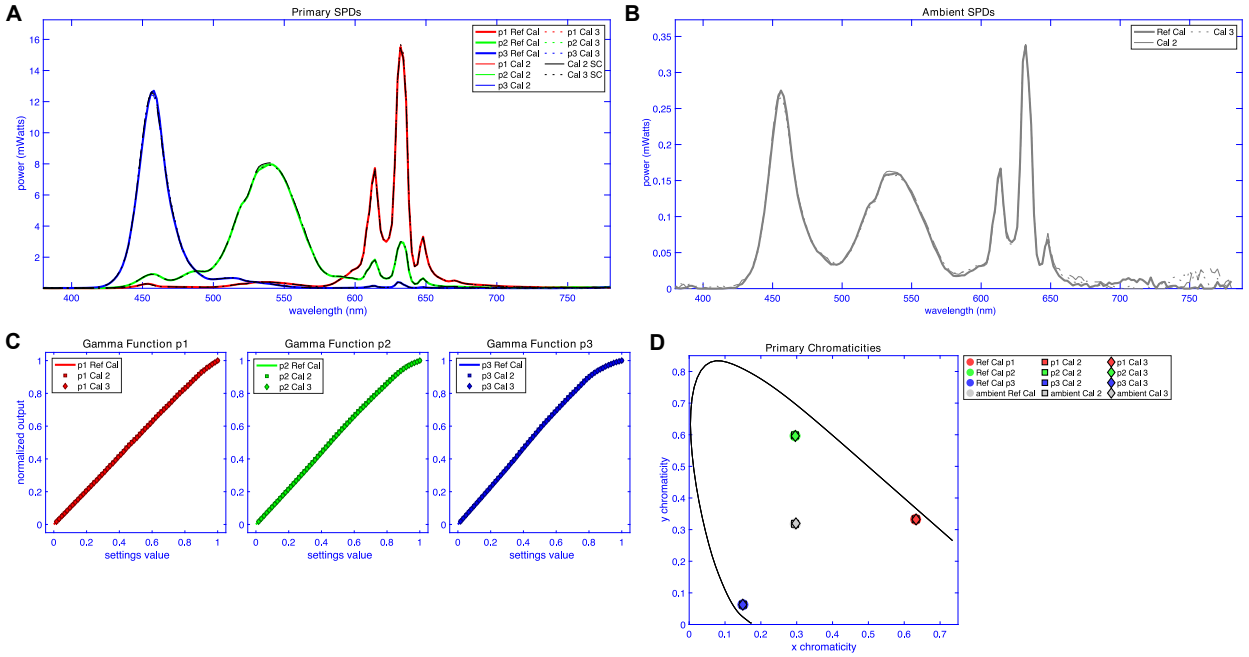

Figure 3. Comparison of color calibration data across the same object positioned at three different locations on the screen. (A) The spectral power distributions (SPDs) of the three primary colors (red, green, and blue) across the three calibration locations (Cal 1: top, Cal 2: left, and Cal 3: right). (B) The ambient SPDs. (C) Gamma functions for each primary. (D) 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.

gamma calibration (primary spectra, gamma correction) will be applied to all three objects during the main experiment, based on the calibration at the center of the screen.

In the second calibration task, we tested whether the gamma function varied based on different measurement areas on the object. A smaller measurement aperture (approximately 0.41 dva) was used for Cal 2 and Cal 3 compared to the one used for the reference calibration (approximately 1.23 dva). Results showed minimal (but observable) differences (**Fig. 4A-B**). These may have to do with Unity's internal gamma correction algorithm, which is not exposed to us. We judged these differences to be small enough to neglect. Moreover, we interpolated gamma table was 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. 4C-D**). Lastly, to validate this gamma correction, we repeated color calibration with the correction applied in Unity; results showed an almost perfect alignment with the identity line (**Fig. 4E**). To ensure the gamma function remains stable over time, we will repeat this calibration every two weeks 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.

In the third 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 511/1024 to 541/1024, with an increment of 1/1024. 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

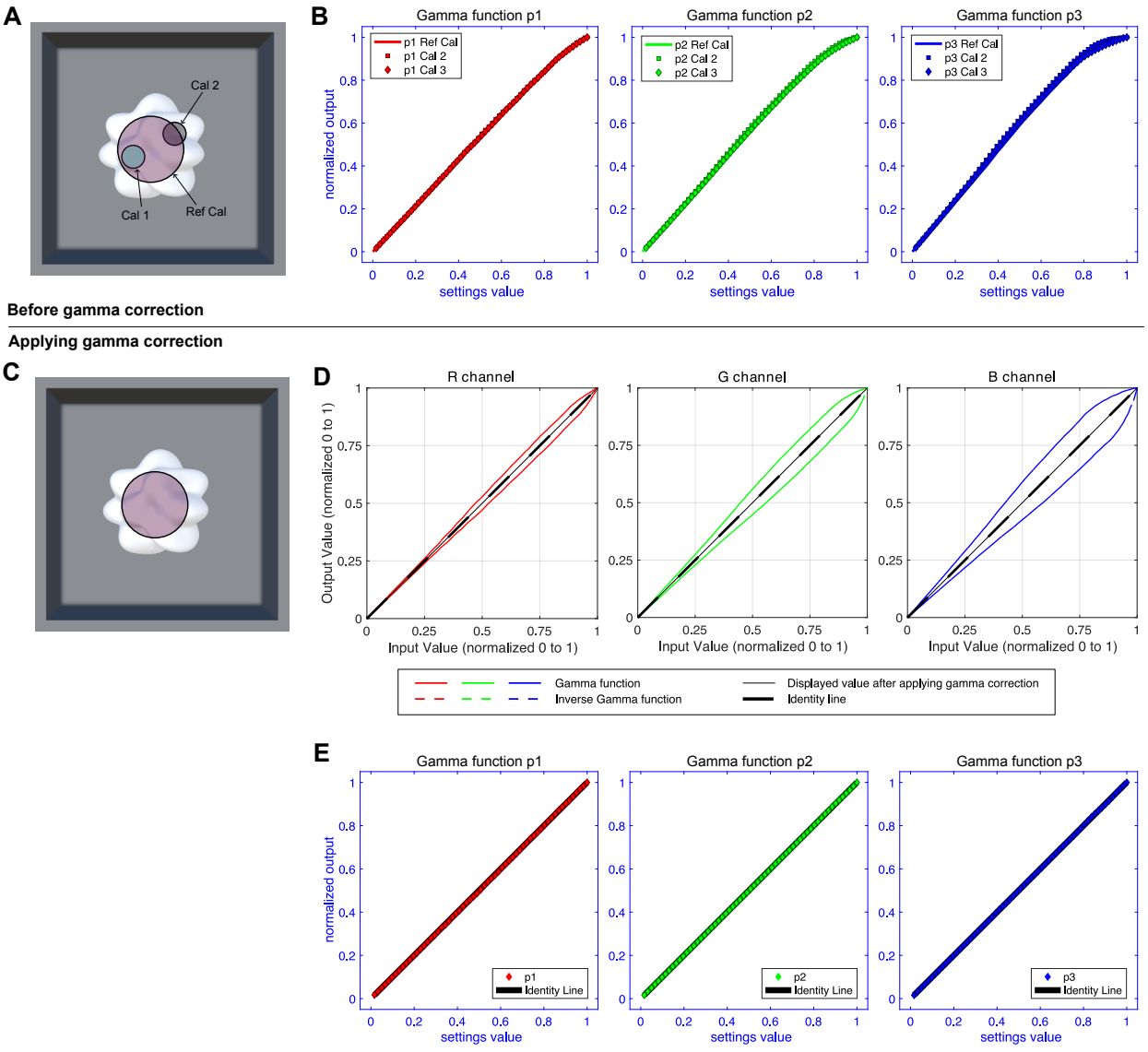

Figure 4. Comparison of color calibration data across different locations on the same object and the effects of gamma correction. (A) The stimulus with marked regions measured by the lens of SpectraScan PR-670. (B) Gamma function comparisons across three measurements for each primary color. (C) Condition with gamma correction applied consistently across all measurement regions. (D) Gamma correction process: 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. (E) Gamma functions measured after applying gamma correction, showing near-perfect alignment with the identity line across all primary colors.

entire stimulus presentation. This confirmed that Unity was able to produce no less than 10-bit depth through what we believe is, in effect, a type of spatial dithering (**Fig. 5**).

Design

The reference stimuli will be sampled as a 3 x 3 grid within the 2D Wishart space, which is bounded between -1 and 1. Sampling will occur evenly between -0.6 to 0.6 with 3 steps along each Wishart dimension, which will also serve as the space where AEPsych determines trial

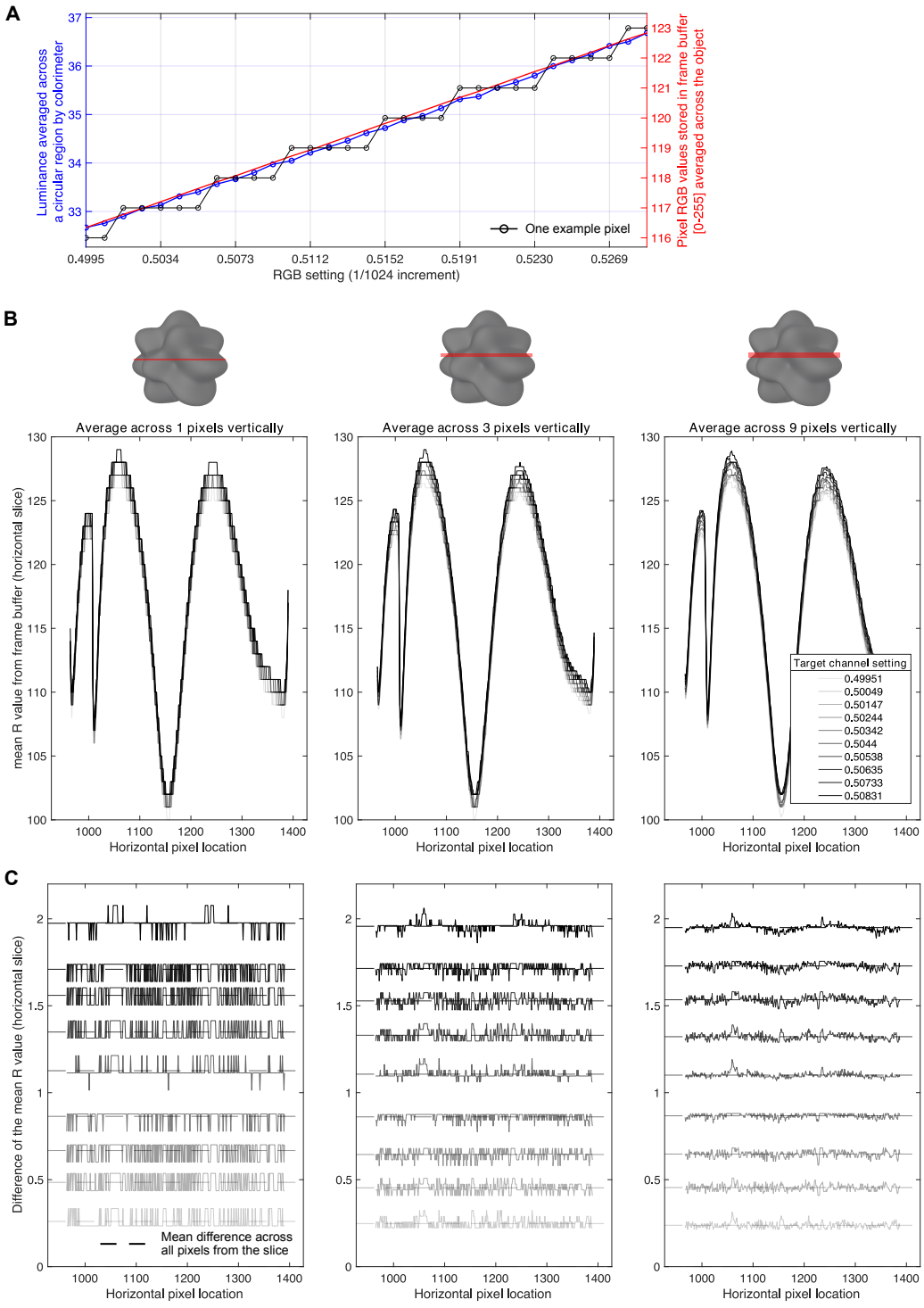

Figure 5. Evidence of spatial dithering by Unity's standard shader. (A) Spatial dithering by Unity's standard shader is suggested by comparing the luminance measurements from the Klein K-10A (averaged across a circular region on the blobby object) with the RGB values stored in the frame buffer. The measured luminance shows small incremental changes as the RGB settings increase in steps of 1/1024. 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). Individual pixel values show 8-bit quantization. (B) Mean R channel values averaged vertically within a horizontal slice of the blobby object. Different shades of gray lines: different target R channel settings. (C) Differences in the R channel values between the minimum target R channel setting and each of the rest settings. For illustration purpose, the solid lines are scaled by a factor of 0.1. Dashed lines: the mean difference averaged across all pixels within each slice.

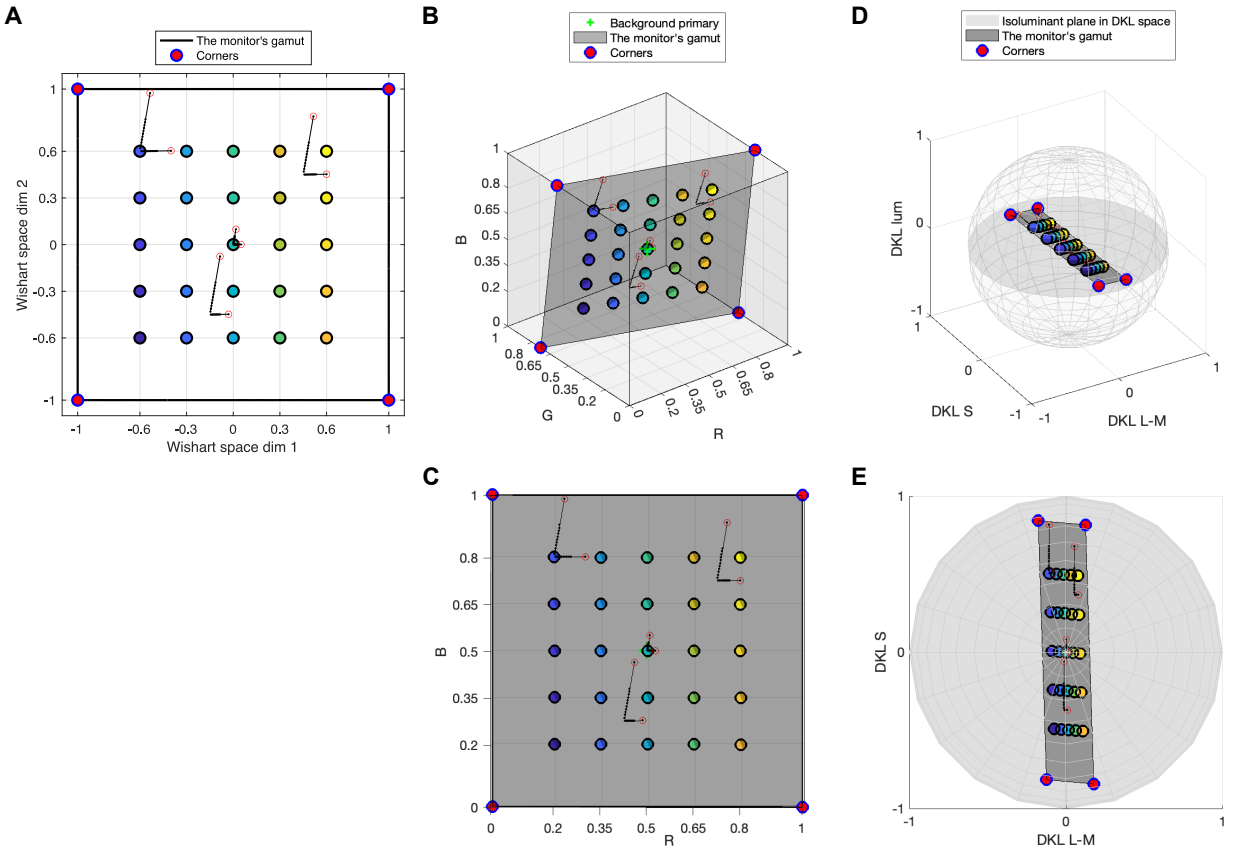

Figure 6. Mapping stimuli across Wishart, RGB, and DKL color spaces. (A) Stimuli in Wishart space, displayed as a 5 x 5 grid of colored dots representing reference stimuli. Black dotted lines indicate 8 sets of MOCS trials, while red open circles mark easy trials used to compute lapse rate. (B) Stimuli represented in RGB space. (C) Projection of the RGB stimuli onto the RB plane. (D) Transformation of stimuli into DKL space, with reference stimuli mapped onto an isoluminant plane in DKL color space. (E) Projection of DKL stimuli onto the isoluminant plane. A 5 x 5 grid is shown for illustrative purposes. As described in the text, the initial experiment will be 3 x 3.

placement. The stimuli will then be transformed to DKL color space, which comprises three dimensions: L+M, L-M, and S. Those values will be further transformed to linear RGB color space, bounded between 0 and 1, used to set the stimulus' surface color in Unity (**Fig. 6**). Gamma correction will be applied when setting the stimuli. After the transformation to linear RGB, the smallest and largest linear RGB values for the reference stimuli will be approximately 0.2 and 0.8.

Each reference stimulus used in the Wishart process fits will be studied with 360 trials, including 60 Sobol-generated trials and 300 model-based trials. For the Sobol-generated trials, we will use scalers of 0.01, 0.5, and 1, repeating each value 20 times and then randomizing the full set of 60 values to form the trial sequence. This provides useful information for AEPsych regarding which directions should result in perfect accuracy and which should lead to chance-level accuracy. For the model-based trials, they are governed by AEPsych, designed to place trials around the 66.7% correct performance level.

Additionally, we will include MOCS trials for 4 reference stimuli positioned at [-0.6, 0.6], [0, 0], [0.45, 0.45], and [-0.15, -0.45] in the Wishart space. For each reference stimulus of the MOCS trials, comparison stimuli will be selected along two chromatic directions, one along L-M axis and the other along S axis in DKL space, resulting in a total of 8 sets of MOCS trials. Each

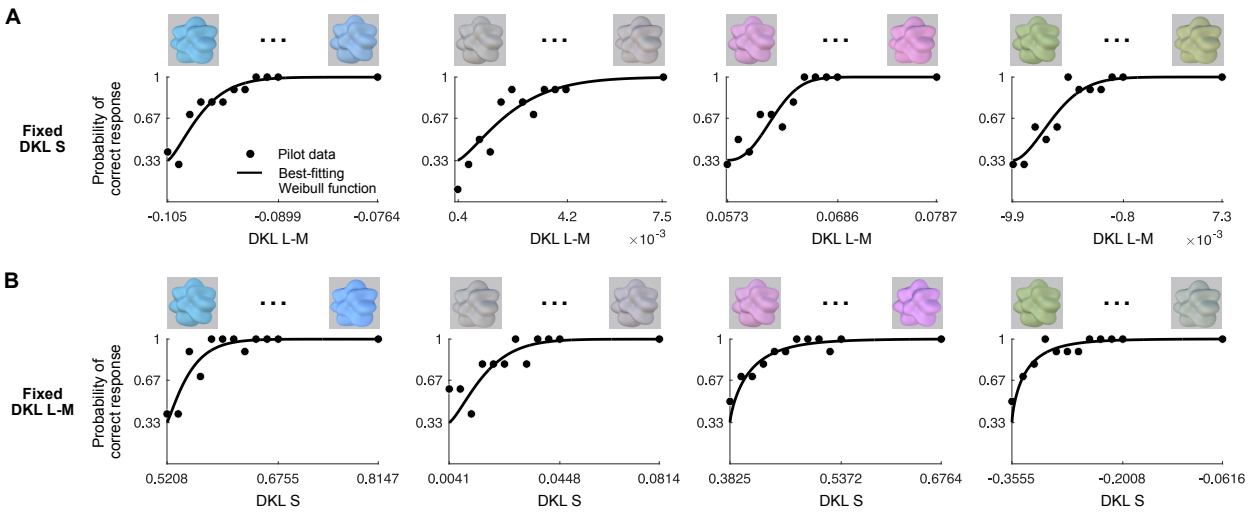

Figure 7. Pilot data used to define the range of MOCS trials. (A) Four sets of MOCS trials conducted with a fixed DKL S axis and varying DKL L-M axis. Each black dot represents the average probability of a correct response across 10 trials for a specific stimulus intensity. (B) Four sets of MOCS trials conducted with a fixed DKL L-M axis and varying DKL S axis. The images above each panel depict the lowest and highest stimulus intensities tested along the varying axis.

direction will be studied with 12 intensity levels, with each level repeated across 30 trials. The minimum and maximum intensity levels were based on results from a pilot experiment with one subject (**Fig. 7**). The MOCS trials will follow a block design, with each block containing all 12 intensity levels shuffled. We pre-generated the trial sequence for each participant by repeated shuffling across blocks.

The experiment will consist of 17 sets, each comprising 360 trials, for a total of 6,120 trials. These sets include 9 from the 3 × 3 grid of AEPsych-based trials and 8 from MOCS-based trials, interleaved in a pseudorandom order. In the future, we may decide to collect additional data at other reference points to combine with these measurements.

Procedure

Participants will perform a three-alternative forced-choice (3AFC) oddity task (**Fig. 8**). 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. Feedback on accuracy will be provided immediately after each response, displaying "correct" or "incorrect" for 0.5 seconds. The inter-trial interval varies, as it depends on the time AEPsych requires to determine the next trial placement. Subjects may move their eyes during stimulus presentation.

The main experiment will consist of ~17 sessions. In each session, participants will begin with practice trials (at least 40 on the first session, and 20 on the following sessions) to familiarize themselves with the task, and may complete as many practice trials as needed. Practice trials will not be analyzed. Each session will consist 720 trials, and participants will take a break every 120 trials, resulting in a total of 6 breaks during each session. Each session is expected to take approximately 2 hours to complete. Data from complete blocks in a session will be retained. Data from any partial blocks will be discarded. Although we will aim for 720 trials

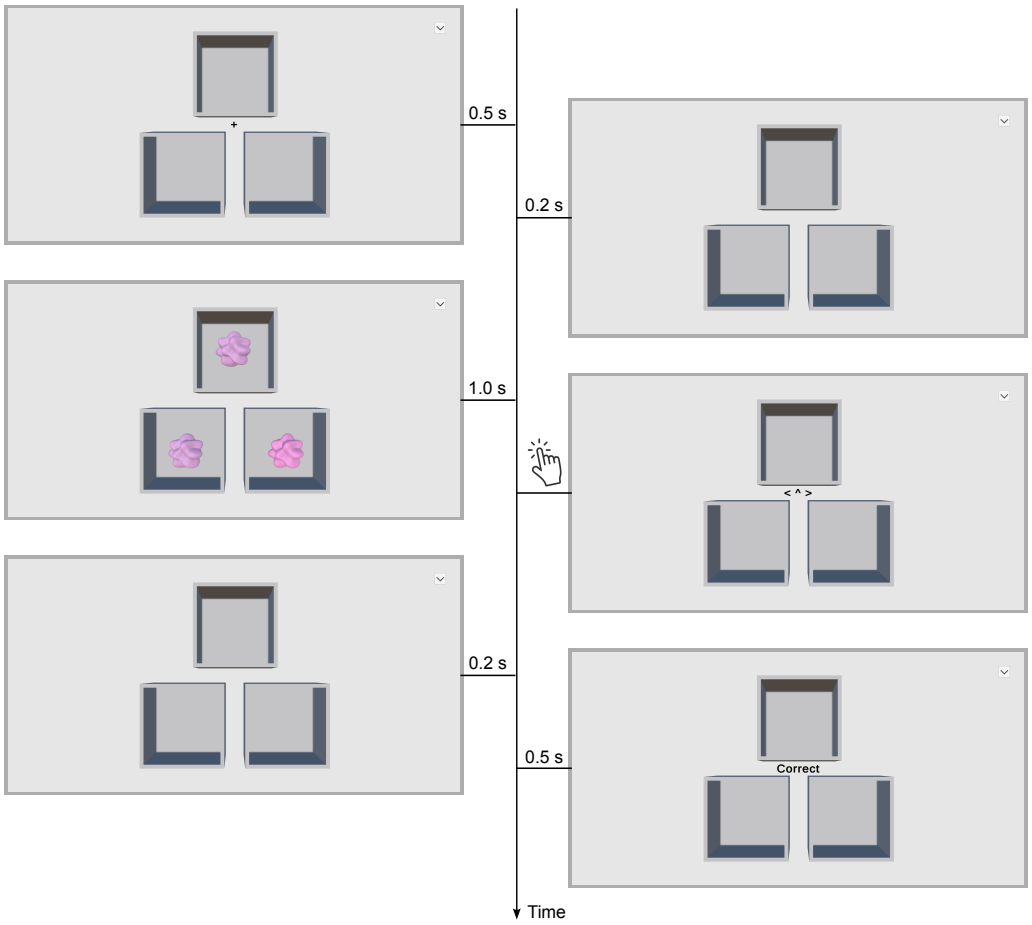

Figure 8. Task timing. Each trial begins with a 0.5-second fixation period, marked by a cross at the center of the screen, followed by a 0.2-second blank interval. Next, three blobby stimuli are presented in the cubic rooms for 1 second. After the stimuli presentation, the screen shows a response prompt, 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, feedback on accuracy is displayed after a 0.2-second delay.

per session, more or fewer may be collected in any particular session depending on how long the trials actually take. Data collection will continue for each subject until all 33 sets are collected, unless the subject withdraws.

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.

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. 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 30 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.

Data analysis

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

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 analyzing behavioral data, we will first separate each participant's data into AEPsych trials and MOCS trials. The Wishart process model will then be fitted exclusively to the AEPsych trials. For each set of MOCS trials, we will fit a Weibull psychometric function to allow comparison with the Wishart model predictions. To evaluate model performance, we will compare the Wishart model with an alternative model that assumes independent threshold contours across reference stimuli, without 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.

Additional analyses are likely to be performed once we have done a first pass on data analysis.