# Preregistration document 3 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

## Completed analysis and finalized plans

While continuing to collect data from seven additional subjects, we have completed a comprehensive analysis of the data from Subject #1. This analysis has helped us finalize our plans for analyzing the full dataset. Here we summarize the key findings and decisions. As a framing comment, we note that many of the analysis are computationally expensive and that it is easy to think of more things we might do. None-the-less, we are making a set of analysis decisions based on what we know now, so that we have time to run the analysis over our entire dataset in advance of our presentation at the 2025 meeting of the Vision Sciences Society.

**Trial number sufficiency**. We considered whether 6,000 trials is sufficient to reliably estimate the full field of color discrimination thresholds. To evaluate this, we fit the Wishart model to subsets of the data—specifically the first 60%, 70%, 80%, and 90% of trials—and examined how the model predictions evolved with increasing data (**Fig. 1A**). If the predictions were highly variable across subsets, it would indicate that more data were needed. However, the results showed that the model predictions stabilized quickly: even the model fit using only 60% of the data closely resembled the predictions based on the full dataset (**Fig. 1B**). This finding suggests that 6,000 AEPsych trials are sufficient for reliable estimation and that additional data collection beyond this number is unnecessary. This was done with a smoothing parameter value of 0.4, our initial estimate of a reasonable value. We have not repeated the analysis with the 0.5 value that we currently favor (see below), although we may do so in the future. See more below on the definition and choice of smoothing parameter.

**Comparison of thresholds estimated separately for each trial type.** We investigated the consistency of threshold estimates derived from the Wishart model fit to the 6000 AEPsych chosen trials and from the separate 6000 validation trials collected using the Method of Constant Stimuli (MOCS). We conducted bootstrap resampling of the AEPsych trials 10 times. In each resampled dataset, we preserved the original ratio between Sobol-generated AEPsych trials and adaptively placed AEPsych trials. We then refit the Wishart model to each of the 10 bootstrapepd datasets to assess variability in the predicted thresholds (**Fig. 2A**, left panel). We compared the resulting 100% confidence intervals (full bootstrapped range) from these bootstrap-based predictions to those obtained from the MOCS validation trials, which were

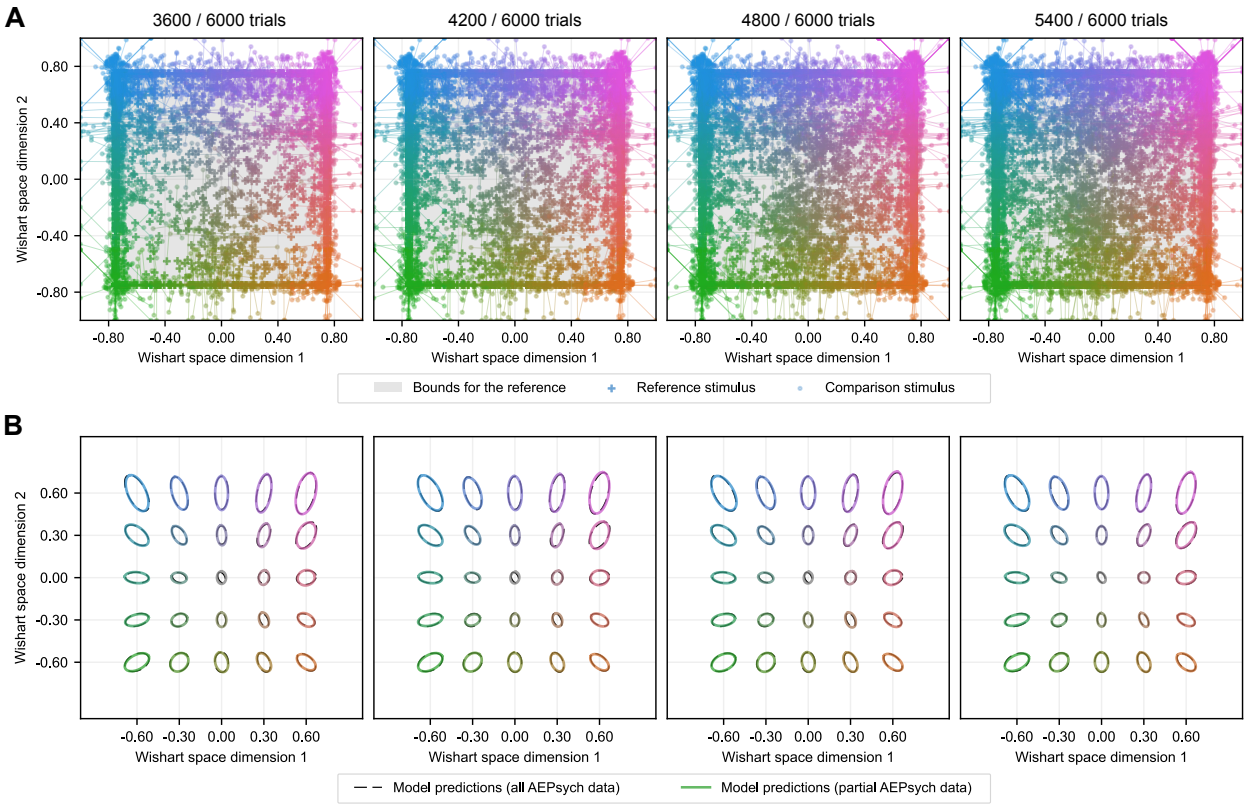


**Figure 1. Update of trial placement and Wishart model predictions with accumulating data.** (A) Trial placement guided by AEPsych adaptive sampling as more data are collected. Each panel shows the distribution of reference (crosses) and comparison (dots) stimuli in Wishart space after 3600, 4200, 4800, and 5400 of the total 6000 trials. The initial 900 Sobol-seeded trials are excluded for clarity. (B) Comparison of Wishart model predictions based on partial datasets (colored ellipses) with those based on the full dataset (black dashed ellipses).

derived using Weibull psychometric function fits. As noted in the text, our current use of 100% confidence intervals is due to the computational time currently required to bootstrap the Wishart Process model fits. We used 95% bootstrapped intervals for the MOCS validation trials.

The threshold estimates from the two trial types showed good agreement, with overlapping confidence intervals in nearly all conditions (**Fig. 2A**, middle panel). We further quantified this agreement by fitting a straight line (constrained to pass through the origin) to the paired threshold estimates across the two trial types (**Fig. 2A**, right panel). The correlation was high (r = 0.73), and the slope was close to unity (1.02), indicating minimal systematic bias. To estimate a 95% confidence interval on the slope, we paired the thresholds from the Wishart model fit to the original AEPsych dataset with those derived from 120 bootstrapped MOCS datasets. This approach accommodates the unequal number of bootstraps conducted for the two trial types. This is an approximation, but one good enough to provide us with a sense of the variability. Together, these results support the efficacy of our approach—using AEPsych for adaptive trial placement and the Wishart Process model for fitting—in accurately deriving the full field of color discrimination thresholds.

Notably, the Wishart Process model is Bayesian, incorporating a prior over the weights applied to Chebyshev basis functions. This prior assumes that the variance of each weight, denoted $\eta$,

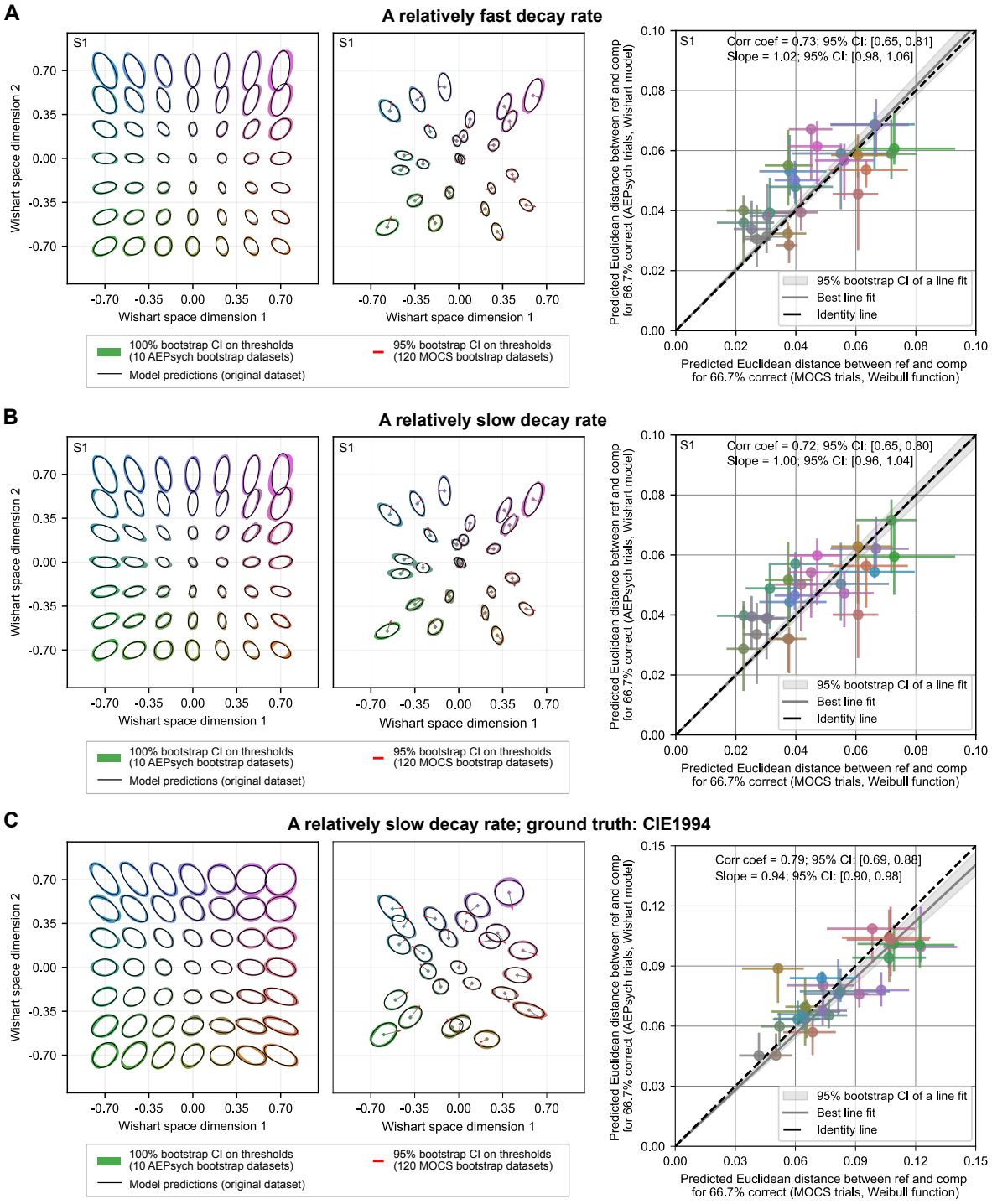


**Figure 2. Comparison of color discrimination thresholds estimated from AEPsych and validation (MOCS) trials.** (A, left panel) Threshold ellipses predicted by the Wishart model based on AEPsych trials, at a 7×7 grid of reference locations. We generated 10 bootstrap resamples of the AEPsych dataset and refit the Wishart model to each. The resulting 100% confidence intervals are shown as colored regions surrounding each ellipse. (A, middle panel) Threshold comparisons at 25 reference locations of the MOCS conditions between estimates from AEPsych and MOCS trials. Red error bars indicate 95% confidence intervals of thresholds estimated from MOCS trials using Weibull psychometric function fits. (A, right panel) Correlation between threshold estimates derived from AEPsych (Wishart model) and MOCS (Weibull fit). A linear fit (solid black line) was constrained to pass through the origin. (B) Same as (A) with a different selected decay rate. (C) Same as (B) except that the data were simulated based on the Wishart fits to CIE1994-derived thresholds.

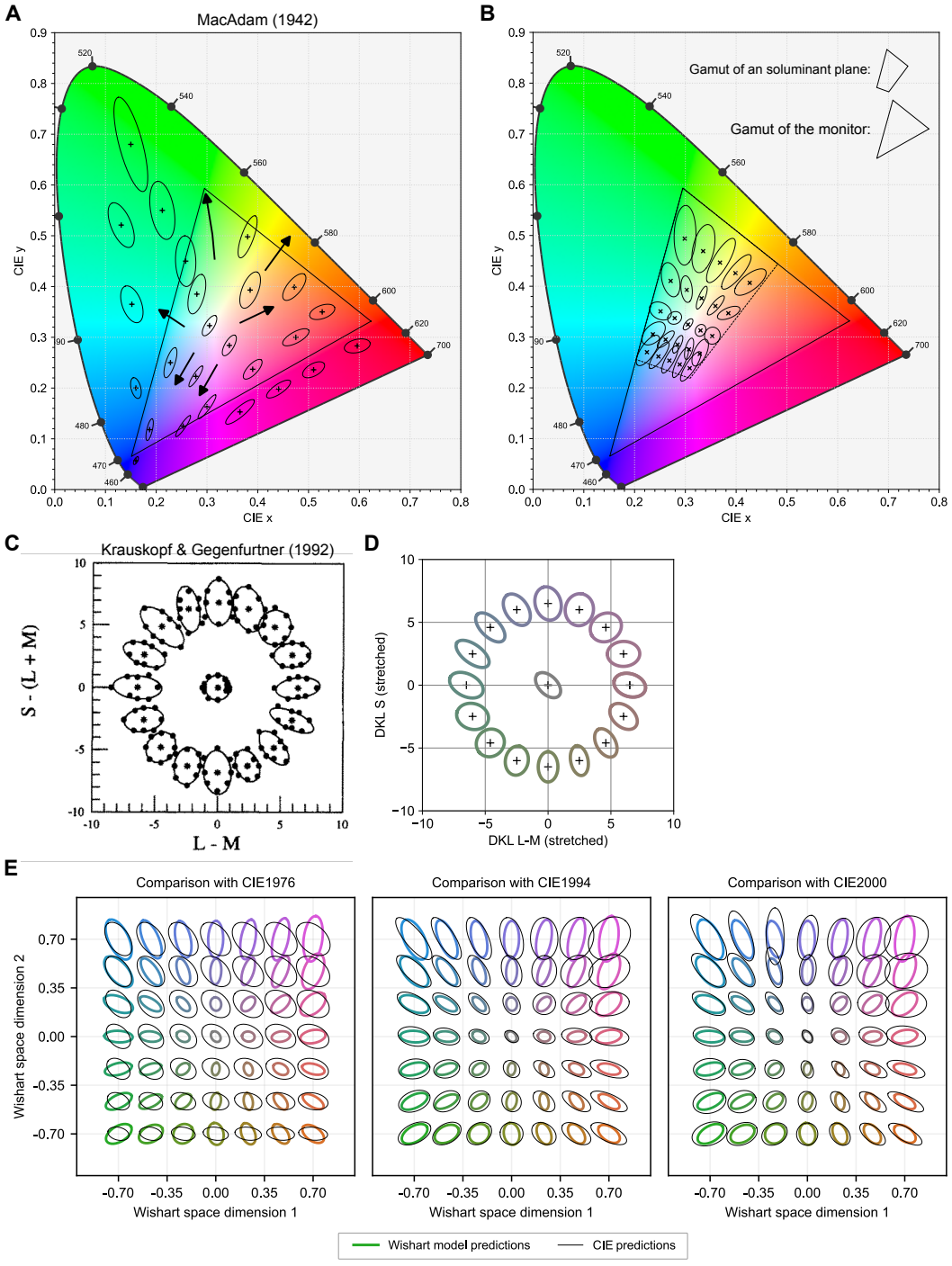


**Figure 3. Comparison of chromatic discrimination thresholds between our experiment and previous studies.** (A) MacAdam ellipses. Black arrows: the guesstimated directions of the major axes at untested reference color, inferred from neighboring ellipses. Ellipses are enlarged by 10 times for illustrative purposes. (B) Threshold ellipses from Subject #1 in our study, transformed from the Wishart model space to CIE 1931 chromaticity space. Reference colors were sampled on a 5 × 5 grid, evenly spaced between –0.75 and 0.75 along each dimension in the Wishart space. Ellipses are enlarged by 2.5 times to roughly match the scale of (A), which itself reports a sacled version of the measured jnds. (C) Threshold measurements from Krauskopf & Gegenfurtner (1992), shown for comparison. (D) Threshold ellipses from Subject #1, transformed from the Wishart model space to DKL color space (L–M and S axes). Axes are scaled so that the threshold at the achromatic reference in each DKL direction has unit length corresponds. This was the scaling convention used by Krauskopf & Gegenfurtner. (E) Comparison of Subject #1's threshold ellipses with those derived from CIE1976 (scaled x5), CIE1994 (scaled x2.5), and CIE2000 (scaled x2.5). Note that the Wishart model predictions shown were computed using a decay rate of 0.5.

decays exponentially with the polynomial order $d$ of the basis function: $\eta = \gamma \cdot (\epsilon^d)$, where $\gamma$ is a fixed scalar (3e-4), and $\epsilon$ is the decay rate. The decay rate $\epsilon$ serves as a tunable hyperparameter controlling how quickly the weights diminish with increasing polynomial order—larger values of $\epsilon$ lead to slower decay, allowing for more complex spatial variation in the covariance matrix field. We examined how varying $\epsilon$ affects the model's predictions. While increasing $\epsilon$ introduces slightly greater uncertainty in the estimated thresholds (compare **Fig. 2A** for $\epsilon = 0.4$ with **Fig. 2B** for $\epsilon = 0.5$), it does not alter the qualitative agreement between thresholds estimated from AEPsych trials (using the Wishart model) and those estimated from MOCS trials (using Weibull fits). More importantly, we found comparable agreement between the two trial types using simulated data using a Wishart Model fit to CIELAB $\Delta$E94 isoperformance contours as ground truth (**Fig. 2C**), although the correlation coefficient for this simulated case was a little higher.

Our next steps are as follows: the remaining seven subjects will complete all 12,000 trials without additional sessions. Once data collection is complete, we will conduct the analysis we did for Subject #1 as described above. In the Wishart model fit, we will start with a smoothing parameter of 0.5 and continue with this unless we judge by eye that the confidence intervals are unreasonably large. In this case we will decrease the smoothing parameter (more smoothness) until the confidence intervals stabilize. The choice of smoothing parameter to be used in the reported comparison with the MOCS thresholds will be made before analyzing the MOCS data. Thus it is possible that a different value of the smoothing parameter will be used for each subject in this comparison. That said, our expectation based on what we have learned to date is that the value of 0.5 will lead to reasonable confidence intervals and thus we expect to proceed with that value for most subjects. We will report the comparisons in the format of **Figure 2** above for each subject, and also provide summaries of descriptive statistics resulting from the comparison (correlation values, slope of regression line, etc.).

In the fullness of time we plan to explore how the comparison between Wishart model predictions and the MOCS thresholds varies with smoothing parameters, as well as explore the efficacy of cross-validation approaches to choosing the smoothing parameters. We will explore these procedures with Subject 1 and the pre-register our plan for comparisons with the rest of the subjects.

**Comparison of thresholds between our experiment and past studies**. We have compared our data to two previous studies that have studied color discrimination at discrete reference locations in the isoluminant plane (MacAdam, 1942; **Fig. 3A;** Krauskopf & Gegenfurtner, 1992; **Fig. 3C**). Many details of the experiments differ across these studies and ours, which we do not describe here. None-the-less, we find making these comparisons interesting at a qualitative level. One advantage of the Wishart model is its ability to interpolate threshold contours continuously across the isoluminant plane. Thus we used this capability to generate predictions for reference colors examined in those two studies. In doing so, we did not correct for differences in overall luminance across the studies. We also focussed interest on the shape of the threshold contours, not the overall size, since many experimental factors not matched across the studies will affect that overall size.

To compare with MacAdam's ellipses, we transformed the threshold ellipses from the Wishart space to CIE xyY via a series of color space conversions: Wishart → RGB → XYZ → xyY. Notably, the RGB-to-XYZ transformation involves a matrix computed as the product of the CIE 1931 color matching functions and the spectral power distributions of our monitor primaries. This comparison is somewhat constrained, as MacAdam's stimuli spanned a wider color gamut than ours (**Fig. 3B**). Only a small subset of MacAdam's tested reference colors overlapped with our sampled region, as he had access to monochromatic primaries in his apparatus. As an alternative, we visualized our model-predicted thresholds on a grid of reference colors (sampled in the Wishart space) and compared the overall pattern of threshold variation. In summary, despite differences in stimulus properties and experimental design, our results share similar variation in ellipse orientation and shape.

To compare our measurements with the threshold ellipses reported by Krauskopf & Gegenfurtner (1992), we first computed the discrimination threshold at the achromatic reference point in Wishart space. This threshold was then transformed into DKL space. We stretched the L–M and S axes such that the threshold ellipse at the achromatic point had unit length along both the x- and y-axes. In this stretched DKL space, we sampled a set of reference colors surrounding the achromatic point. For each reference, we computed the corresponding threshold in Wishart space and transformed it into the stretched DKL space (**Fig. 3D**). Overall, the shape and size of the resulting threshold ellipses qualitatively resemble those reported by Krauskopf & Gegenfurtner (1992), despite some differences that may be to individual variability and differences in experimental stimuli and setup.

We have also compared our data to the shape of isoperformance contours predicted by CIELAB $\Delta$E (1976), $\Delta$E94, and $\Delta$E00. We find poor agreement with $\Delta$E (1976), reasonable agreement with $\Delta$E94 but with some deviations, and reasonable agreement with $\Delta$E00 but with this not as good as with $\Delta$E94 (**Fig. 3E**).

Once data collection is complete, we will provide the comparisons described above to the MacAdam and Krauskopf and Gegenfurtner data, as well as to the CIELAB $\Delta$E metrics for each subject. We will also consider comparisons to other data sets and models. Because of differences in experimental parameters, we view these as qualitative comparisons of interest rather than validations of our procedures.