

Appendix A: Supplementary Methodological Details

A.1 Individual white point adjustment procedure and coverage

(Supplementary to Subsection 2.1.3: Matching software and procedure)

During the individual white point adjustment mode, the central black patch was fully illuminated by the display foreground. Observers adjusted chromaticity along the a^* (red-green) and b^* (yellow-blue) dimensions in CIELAB space—a Cartesian counterpart of the polar CIELCH space that can be directly transformed conversely—until the patch appeared neutrally white, while the lightness component (L^*) remained fixed. While the primary color-matching tasks were conducted in CIELCH, adjusting a^* and b^* provided an intuitive orthogonal control to neutralize hue shifts before returning to CIELCH coordinates.

Observers could re-enter this adjustment mode at any point during their session.

The individual white point adjustment mode was introduced midstream during the study. The primary motivation was that the initial reference white point was noticeably yellowish, causing color trajectories during chroma adjustments to deviate from subjective neutral white and increasing task difficulty for observers. Consequently, this adjustment feature was active during later-executed experimental sessions, whereas some earlier sessions operated without it (as noted in Section 2.4). Appendix Table B1 provides the complete matrix mapping observer participation across experiments and indicating session-specific adjustment availability.

A.2 Observer chromatic adaptation and reference white points

(Supplementary to Section 2.5: Data preparation)

Prior to model fitting, descriptive colorimetric analysis of the observer-adjusted white points was conducted to characterize the chromatic adaptation states of the observers and establish the reference white points for subsequent CIECAM16-UCS transformations (as specified in Table I). During the white point adjustment routine, stimulus luminance (Y) was fixed and non-adjustable; nominal luminance levels (31.8 cd/m² for high luminance and 7.06 cd/m² for low luminance) were retained across all calculations.

Individual white point coordinates were first converted into the CIE 1976 (u' , v') chromaticity diagram, where arithmetic averaging across observers was performed separately for the two ambient luminance conditions. Overall CCT, D_{uv} and CIE 1931 (x , y) chromaticity coordinates were subsequently derived from these condition-wise mean (u' , v') coordinates.

Under the high-luminance condition, the mean (u' , v') coordinates corresponded to a CCT of 5390K (individual observer CCT range: 4733–6092K), a D_{uv} of 0.0060 (individual D_{uv} range: 0.0019–0.0096), and CIE 1931 chromaticity coordinates of (x , y) = (0.3353, 0.3555).

Under the low-luminance condition, the mean (u' , v') coordinates corresponded to a CCT of 5417K (individual observer CCT range: 4781–7615K), a D_{uv} of 0.0038 (individual D_{uv} range: 0.0004–0.0109), and CIE 1931 chromaticity coordinates of (0.3343, 0.3504).

Consistent with prior literature¹, these observer-adjusted CCT values were systematically

¹ H. P. Huang, M. Wei, and S. Chen, “White appearance of virtual stimuli produced by augmented reality,” Color

higher than the nominal light source CCT of 4334K.

A.3 Outlier detection and model sensitivity analysis

(Supplementary to Section 2.5: Data preparation)

Following the computation of CIECAM16-UCS coordinates (J' , a' , b' , dimensionality $p = 3$), the dataset was categorized into distinct subgroups based on the target combinations under identical experimental conditions within each experiment. For each subgroup (8 observers $\times$ 2 repetitions = 16 observations), multivariate outlier detection was performed using squared Mahalanobis distance (D^2) with leave-one-out jackknife resampling to prevent single extreme values from distorting the covariance matrix.

Critical upper-tail thresholds were computed using the F -distribution ($F(p, N - p - 1)$) at $\alpha = 0.05$ with a Bonferroni-correction for subgroup size (N , yielding an initial threshold of $D^2 = 30.5$). Observations exceeding the critical threshold were sequentially removed until no further outliers were detected. This iterative procedure removed a maximum of two observations per subgroup, excluding 12 extreme outliers out of 1088 observations (1.1%).

A model sensitivity analysis was conducted by fitting the GLMMs on both the complete and trimmed dataset. The fixed-effect estimates and core statistical conclusions remained virtually identical between the two fits, confirming the robustness of our results. Outlier removal was

nevertheless retained to ensure high data hygiene by eliminating obvious procedural matching errors.

A.4 Experiment 3 data supplementation during individual modeling

(Supplementary to Section 2.5: Data preparation)

In Experiment 3, which investigated the effects of ambient luminance and presentation medium, display-only conditions were evaluated under both ambient luminance levels because display rendering is virtually unaffected by peak luminance adjustments. However, due to procedural constraints associated with the preheating and chromatic stabilization of the background light sources, the AR medium condition was evaluated only under the low-luminance condition.

When fitting individual models for Experiment 3 (i.e., during preliminary factor screening), this created a structural gap that induced high multicollinearity among predictors. To stabilize individual model estimation, corresponding observations for common target combinations (Targets 3 and 7) were supplemented from Experiments 1 and 2.

A potential concern is whether pooling data across different experiments introduces bias due to inter-experiment heterogeneity. However, as is detailed in Section A.5, including *experiment* as a random effect in the pooled model had negligible effects, indicating high consistency across experiments.

To further verify that this data supplementation did not distort model predictions, a

consistency diagnostic was performed between the model fitted on the supplemented subset and the model fitted on the full Experiment 3 dataset across the shared reference grid. The predictions exhibited near-perfect agreement (Lin's concordance correlation coefficient, CCC = 0.995; Agreement R^2 = 0.995; mean bias = -0.167), confirming that factor screening on the supplemented dataset was highly robust.

When data from all experiments were pooled into a unified GLMM, the multi-experiment dataset naturally provided full cross-condition coverage. Therefore, no data supplementation was performed or required for the consolidated pooled model analysis.

A.5 Evaluation of model architecture and secondary factors

(Supplementary to Subsection 2.6.3: Analytical strategy and model selection)

During model consolidation, candidate random-effects structures and secondary factors were systematically evaluated using top-down AICc comparisons. Including an inter-experiment random intercept “(1 | experiment)” yielded a higher AICc (Δ AICc = +1.6) and negligible variance explained, confirming structural consistency across experiments and supporting the simpler random-effects structure “(0 + axis | observer)” (Table B3.5).

To verify whether procedural or cognitive variations biased matching performance, two secondary factors were evaluated: whether the observer adjusted the white point (*wp_adjusted*, evaluated in Experiments 1–2 where both adjusted and non-adjusted conditions were present), and the observer's explicit cognitive awareness of the display medium

(*disp_aware*, evaluated in Experiments 3–4). Incorporating these terms into their respective experimental datasets did not significantly improve model fit (*wp_adjusted*: $\chi^2 = 5.79$, $p = 0.122$; *disp_aware*: $\chi^2 = 0.49$, $p = 0.920$) and increased AICc values (Tables B3.6 and B3.7). Consequently, these factors were excluded from the final pooled model, confirming that perceptual color shifts were robust against white point calibration status and observer awareness of display media.

A.6 Model diagnostics and collinearity assessment

(Supplementary to Section 3.2: Model selection outcomes and validation)

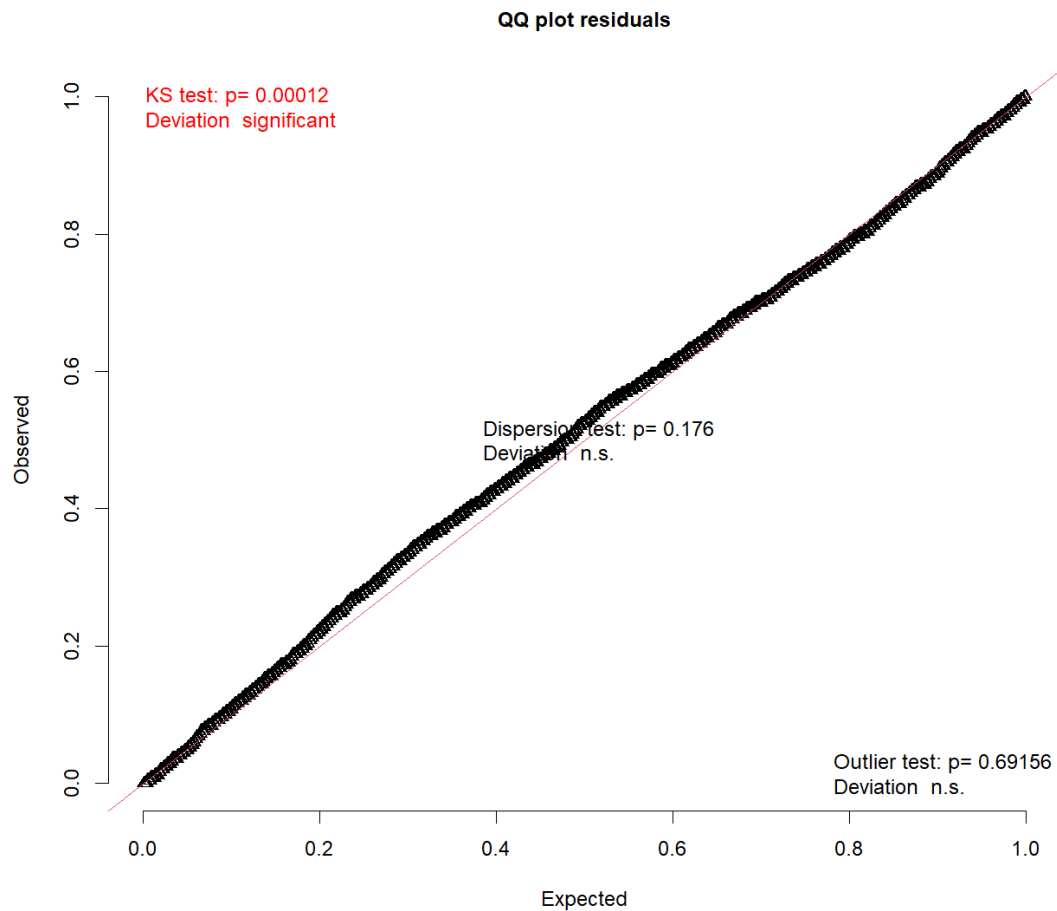


Figure A1. DHARMA residual diagnostic plot for the consolidated pooled GLMM. The quantile-quantile (Q–Q) plot shows observed versus expected uniform scaled residuals, annotated with results for the Kolmogorov-Smirnov (KS) test ($p < 0.001$), dispersion test ($p = 0.176$), and outlier test ($p = 0.692$). The tight visual alignment along the reference line confirms distributional adequacy, with the significant KS test result driven by high statistical power from the large sample size ($N = 3228$).

Model adequacy and distributional assumptions for the pooled GLMM were evaluated using DHARMA-based residual diagnostics (Figure A1). Conditional dispersion ($p = 0.176$) and

residual outlier tests ($p = 0.692$) showed no significant violation. The Kolmogorov-Smirnov (KS) test for residual uniformity yielded $p < 0.001$ on the pooled dataset ($N = 3228$). As shown in Figure A1, despite the KS test result, which was interpreted as a Type I error resulting from excessive statistical power in large sample sizes, the scaled residuals display tight visual alignment along the 1:1 reference diagonal with no systematic deviations from normality. Identical KS tests passed across all four individual experiment models ($p = 0.833, 0.428, 0.345, 0.325$, respectively, with Experiment 3 evaluated on its full, unsupplemented dataset; see Section A.4). Multicollinearity was assessed using variance inflation factors (VIF). Elevated VIF values from terms involving *target_id* (target combination) are a mathematical consequence of multi-level categorical interaction terms within a non-orthogonal repeated-measures design, rather than model misspecification.

Appendix B: Supplementary Tables

Table B1. Matrix of observer participation and white point adjustment mode coverage across Experiments 1–4. Each entry indicates whether an observer completed a session with active individual white point adjustment (solid circle, ●), completed a session with the background white patch as the reference white point (open circle, ○), or did not participate (em dash, —).

Bottom rows summarize the total observer count per experiment.

Observer	Total Sessions	Experiment 1	Experiment 2	Experiment 3	Experiment 4
1	4	○	●	●	●
2	4	○	●	●	●
3	4	○	●	●	●
4	2	—	—	●	●
5	1	—	○	—	—
6	1	—	○	—	—
7	4	○	●	●	●
8	2	●	●	—	—
9	3	●	●	—	●
10	2	—	—	●	●
11	3	○	—	●	●
12	2	●	—	●	—
Total Observers	32	8	8	8	8

Table B2. Experimental conditions and their mapping to conditions. Rows are ordered by experiment and phase. The same condition may appear in multiple rows, indicating identical factor level combinations across experiments. See Table IV for the complete definition of each condition.

Experiment	Phase	Targets	viewing mode	stimulus size	aperture presence	foreground offset	start point	ambient luminance	presentation medium	Condition	N (observations)
Exp 1	P1	3, 7, 13, 19	monocular	small	TRUE	aligned	mid	high	AR	1	63
		3, 7, 13, 19		large						2	63
	P2	3, 7, 13, 19	binocular	small	TRUE	aligned	mid	high	AR	4	63
		3, 7, 13, 19		large						3	63

Exp 2	P1	3, 7	binocular	small	TRUE	aligned	mid	high	AR	4	32
		3, 7					low			5	32
		3, 7					high			6	32
	P2	3, 7			FALSE	offset	mid			4	31
	P3	3, 7					mid			4	32
Exp 3	P1	1, 3, 4, 6, 7, 9	binocular	small	TRUE	aligned	mid	low	AR	7	92
	P2	1, 3, 4, 6, 7, 9							display	9	96
	P3	1, 3, 4, 6, 7, 9						high		11	96
Exp 4	P1	1, 3, 4,	binocular	small	TRUE	aligned	mid	low	AR	7	96

		6, 7, 9									
		1, 3, 4, 6, 7, 9					low			8	95
	P2	1, 3, 4, 6, 7, 9					mid			9	95
		1, 3, 4, 6, 7, 9					low			10	95

Table B3. Model selection results for each experiment. The table comprises seven panels (B3.1-B3.7), one per experiment or pooled analysis. Each panel reports the fixed-effect structure, number of observations, number of observers, and model comparison statistics. The selected model for each comparison (lowest AICc) is indicated in bold. See Section 2.6 for model specification details.

Table B3.1. Experiment 1 (*view_mode* and *stim_size*). 756 observations, eight observers.

Fixed effects:

hmlmm_exp1_4: Diff ~ axis * (view_mode + target_id)

hmlmm_exp1_3: Diff ~ axis * (view_mode + stim_size + target_id)

hmlmm_exp1_2: Diff ~ axis * (view_mode + stim_size * target_id)

hmlmm_exp1_1: Diff ~ axis * (view_mode * stim_size * target_id)

Random effects:

(0 + axis | observer)

Zero-inflation formula:									
~0									
Dispersion formula:									
~axis * target_id									
	Df	AIC	AICc (weights)	BIC	logLik	deviance	Chisq	Chi df	Pr (> Chisq)
hmlmm_exp1_4	34	2864.6	2867.9 (< 0.001)	3021.9	-1398.3	2796.6			
hmlmm_exp1_3	37	2751.9	2755.9 (< 0.001)	2923.2	-1339.0	2677.9	118.635	3	< 2e-16 ***
hmlmm_exp1_2	46	2609.7	2615.8 (0.999)	2822.6	-1258.9	2517.7	160.200	9	< 2e-16 ***
hmlmm_exp1_1	67	2615.6	2628.9 (0.001)	2925.7	-1240.8	2481.6	36.114	21	0.02123 *

Table B3.2. Experiment 2 (*aperture*, *fg_offset*, and *start_point*). 477 observations, eight observers.

Fixed effects:

hmlmm_exp2_6: $\text{Diff} \sim \text{axis} * \text{target_id}$

hmlmm_exp2_5: $\text{Diff} \sim \text{axis} * (\text{start_point} + \text{target_id})$

hmlmm_exp2_4: $\text{Diff} \sim \text{axis} * (\text{start_point} * \text{target_id})$

hmlmm_exp2_2: $\text{Diff} \sim \text{axis} * ((\text{aperture} + \text{start_point}) * \text{target_id})$

hmlmm_exp2_3: $\text{Diff} \sim \text{axis} * ((\text{fg_offset} + \text{start_point}) * \text{target_id})$

hmlmm_exp2_1: $\text{Diff} \sim \text{axis} * ((\text{aperture} + \text{fg_offset} + \text{start_point}) * \text{target_id})$

Random effects:

$(0 + \text{axis} \mid \text{observer})$

Zero-inflation formula:

~ 0

Dispersion formula:									
~axis * target_id									
	Df	AIC	AICc (weights)	BIC	logLik	deviance	Chisq	Chi df	Pr (>Chisq)
hmlmm_exp2_6	19	2078.9	2080.5 (0.045)	2158.1	-1020.4	2040.9			
hmlmm_exp2_5	25	2072.1	2075.0 (0.735)	2176.3	-1011.0	2022.1	18.7974	6	0.00452 **
hmlmm_exp2_4	31	2072.9	2077.4 (0.216)	2202.1	-1005.5	2010.9	11.1298	6	0.08445 .
hmlmm_exp2_2	37	2080.3	2086.7 (0.002)	2234.5	-1003.1	2006.3	4.6403	6	0.59070
hmlmm_exp2_3	37	2082.3	2088.8 (< 0.001)	2236.5	-1004.2	2008.3	0.0000	0	1.00000
hmlmm_exp2_1	43	2090.4	2099.2 (< 0.001)	2269.7	-1002.2	2004.5	3.9006	6	0.69013

Table B3.3. Experiment 3 (*ambient_lum* and *medium*). 477 observations, twelve observers. Dataset supplemented with select data from Experiment 1 and 2 to reduce collinearity.

Fixed effects:

hmlmm_exp3s_7: $\text{Diff} \sim \text{axis} * \text{target_id}$

hmlmm_exp3s_6: $\text{Diff} \sim \text{axis} * (\text{medium} + \text{target_id})$

hmlmm_exp3s_4: $\text{Diff} \sim \text{axis} * (\text{ambient_lum} + \text{medium} + \text{target_id})$

hmlmm_exp3s_5: $\text{Diff} \sim \text{axis} * (\text{medium} * \text{target_id})$

hmlmm_exp3s_2: $\text{Diff} \sim \text{axis} * (\text{ambient_lum} * \text{target_id} + \text{medium})$

hmlmm_exp3s_3: $\text{Diff} \sim \text{axis} * (\text{ambient_lum} + \text{medium} * \text{target_id})$

hmlmm_exp3s_1: $\text{Diff} \sim \text{axis} * ((\text{ambient_lum} + \text{medium}) * \text{target_id})$

Random effects:

(0 + axis | observer)

Zero-inflation formula:									
~0									
Dispersion formula:									
~axis * target_id									
	Df	AIC	AICc (weights)	BIC	logLik	deviance	Chisq	Chi df	Pr (>Chisq)
hmlmm_exp3s_7	19	2191.8	2193.5 (0.016)	2271.0	-1076.9	2153.8			
hmlmm_exp3s_6	22	2191.7	2194.0 (0.012)	2283.4	-1073.9	2147.7	6.0689	3	0.10831
hmlmm_exp3s_5	25	2194.4	2197.3 (0.002)	2298.6	-1072.2	2144.4	3.2851	3	0.34972
hmlmm_exp3s_4	25	2196.2	2199.1 (< 0.001)	2300.4	-1073.1	2146.2	0.0000	0	1.00000
hmlmm_exp3s_2	28	2181.7	2185.3 (0.905)	2298.4	-1062.9	2125.7	20.4759	3	0.00014 ***
hmlmm_exp3s_3	28	2198.9	2202.5 (< 0.001)	2315.6	-1071.5	2142.9	0.0000	0	1.00000
hmlmm_exp3s_1	31	2186.2	2190.7 (0.063)	2315.4	-1062.1	2124.2	18.6866	3	0.00032 ***

Table B3.4. Experiment 4 (*start_point* and *medium*). 1143 observations, eight observers.

Fixed effects:

hmlmm_exp4_7: $\text{Diff} \sim \text{axis} * \text{target_id}$

hmlmm_exp4_6: $\text{Diff} \sim \text{axis} * (\text{start_point} + \text{target_id})$

hmlmm_exp4_5: $\text{Diff} \sim \text{axis} * (\text{start_point} + \text{target_id}) + \text{medium}$

hmlmm_exp4_4: $\text{Diff} \sim \text{axis} * (\text{start_point} + \text{target_id} + \text{medium})$

hmlmm_exp4_3: $\text{Diff} \sim \text{axis} * (\text{start_point} * \text{target_id})$

hmlmm_exp4_2: $\text{Diff} \sim \text{axis} * (\text{start_point} * \text{target_id} + \text{medium})$

hmlmm_exp4_1: $\text{Diff} \sim \text{axis} * (\text{start_point} * \text{medium} * \text{target_id})$

Random effects:

$(0 + \text{axis} \mid \text{observer})$

Zero-inflation formula:

~0 Dispersion formula: ~axis * target_id									
	Df	AIC	AICc (weights)	BIC	logLik	deviance	Chisq	Chi df	Pr (>Chisq)
hmlmm_exp4_7	43	5038.8	5042.3 (< 0.001)	5255.6	-2476.4	4952.8			
hmlmm_exp4_6	46	5029.5	5033.4 (0.037)	5261.4	-2468.8	4937.5	15.3488	3	0.00154 **
hmlmm_exp4_5	47	5023.2	5027.3 (0.811)	5260.1	-2464.6	4929.2	8.3282	1	0.00390 **
hmlmm_exp4_4	49	5026.2	5030.7 (0.147)	5273.3	-2464.1	4928.2	0.9488	2	0.62225
hmlmm_exp4_2	61	5034.3	5041.3 (< 0.001)	5341.8	-2456.1	4912.3	15.9538	12	0.19336
hmlmm_exp4_3	64	5030.9	5038.6 (0.003)	5353.6	-2451.4	4902.9	9.3653	3	0.02481 *
hmlmm_exp4_1	97	5065.5	5083.7 (< 0.001)	5554.5	-2435.7	4871.5	31.4388	33	0.54490

Table B3.5. Pooled model of all experiments (*experiment* as a random effect). 3228 observations, twelve observers, four experiments.

Fixed effects: Diff ~ axis * (view_mode + (stim_size + ambient_lum) * target_id + start_point + medium) Random effects: hmlmm_pooled_woExp: (0 + axis observer) hmlmm_pooled_wExp: (0 + axis observer) + (1 experiment) Zero-inflation formula: ~0 Dispersion formula: ~axis * target_id									
	Df	AIC	AICc (weights)	BIC	logLik	deviance	Chisq	Chi df	Pr (>Chisq)
hmlmm_pooled_woExp	97	13566	13571.9 (0.694)	14156	-6685.9	13372			

hmlmm_pooled_wExp	98	13567	13573.5 (0.306)	14163	-6695.6	13371	0.4895	1	0.4842
-------------------	----	-------	-----------------	-------	---------	-------	--------	---	--------

Table B3.6. Partially pooled (*wp_adjusted* as a secondary fixed effect). 1233 observations, ten observers.

Fixed effects: hmlmm_exp1n2_woWP: Diff ~ axis * (view_mode + stim_size * target_id + start_point) hmlmm_exp1n2_wWP: Diff ~ axis * (view_mode + stim_size * target_id + start_point + wp_adjusted) Random effects: (0 + axis observer) Zero-inflation formula: ~0 Dispersion formula: ~axis * target_id									
	Df	AIC	AICc (weights)	BIC	logLik	deviance	Chisq	Chi df	Pr (>Chisq)
hmlmm_exp1n2_woWP	52	4754.6	4759.3 (0.595)	5020.7	-2325.3	4650.6			

hmlmm_exp1n2_wWP	55	4754.8	4760.0 (0.405)	5036.3	-2322.4	4644.8	5.7918	3	0.1222
------------------	----	--------	----------------	--------	---------	--------	--------	---	--------

Table B3.7. Partially pooled (*disp_aware* as a secondary fixed effect). 1146 observations, nine observers.

Fixed effects: hmlmm_expOnDisplay_woDA: Diff ~ axis * (ambient_lum * target_id + start_point) hmlmm_expOnDisplay_wDA: Diff ~ axis * (ambient_lum * target_id + start_point + disp_aware) Random effects: (0 + axis observer) Zero-inflation formula: ~0 Dispersion formula: ~axis * target_id									
	Df	AIC	AICc (weights)	BIC	logLik	deviance	Chisq	Chi df	Pr (>Chisq)
hmlmm_expOnDisplay_woDA	64	5024.2	5031.9 (0.958)	5347.1	-2448.1	4896.2			

hmlmm_expOnDisplay_wDA	67	5029.8	5038.2 (0.042)	55367.7	-2447.9	4895.8	0.4948	3	0.92
------------------------	----	--------	----------------	---------	---------	--------	--------	---	------

Table B4. Full Type III Wald χ^2 test results, random effects, and model diagnostics for the final pooled linear mixed-effects model.

Model metrics included AIC, AICc, BIC, log-likelihood, McFadden’s Pseudo-R² (marginal and conditional), and likelihood-based ICC. Residual diagnostics are based on DHARMa. All diagnostics were satisfactory except for DHARMa Kolmogorov-Smirnov (KS) test, for which the significant p-value ($p < 0.001$) is attributable to the large sample size ($N = 3228$; see Section A.6). Residual prediction errors are reported as mean absolute error (MAE) and root mean square error (RMSE) separately for each axis.

For fixed effects, χ^2 values (Type-III Wald tests) and p-values are reported. Statistically significant p-values ($p < 0.05$) are highlighted in bold. VIF values indicate collinearity among predictors and were largely influenced by the non-orthogonal experimental design. Specifically, large VIF values for terms involving *target_id* (target combination) are expected due to the multi-level categorical nature of this factor and do not indicate model misspecification. Random effects are reported as standard deviations (Std. Dev) for each axis across observers.

Category	Term	Value / χ^2	Df	P	VIF
Model Metric	AIC	13744.1			
Model Metric	AICc	13750.2			
Model Metric	BIC	14333.8			
Model Metric	LogLik	-6775.1			
Model Metric	Pseudo R2 (marginal)	0.131			
Model Metric	Pseudo R2 (conditional)	0.172			
Model Metric	ICC (likelihood-based)	0.019			
Model Metric	Residual MAE (J')	2.761			
Model Metric	Residual MAE (a')	1.050			
Model Metric	Residual MAE (b')	1.345			
Model Metric	Residual RMSE (J')	3.939			
Model Metric	Residual RMSE (a')	1.424			

Model Metric	Residual RMSE (b')	2.028			
Diagnostic	DHARMa KS Test			0.0001	
Diagnostic	DHARMa Dispersion Test			0.1760	
Diagnostic	DHARMa Outliers Test			0.6916	
Fixed Effect	(Intercept)	3.447	1	0.0634	
Fixed Effect	axis	26.450	2	< 0.0001	3.72
Fixed Effect	view_mode	5.639	1	0.0176	2.86
Fixed Effect	stim_size	14.586	1	0.0001	9.30
Fixed Effect	ambient_lum	2.252	1	0.1335	23.5
Fixed Effect	target_id	340.474	7	< 0.0001	4.80e6
Fixed Effect	start_point	1.184	2	0.5533	3.50
Fixed Effect	medium	0.884	1	0.3472	2.86
Fixed Effect	stim_size*target_id	25.831	3	< 0.0001	511
Fixed Effect	ambient_lum*target_id	17.642	5	0.0034	3.40e4
Fixed Effect	axis*view_mode	14.173	2	0.0008	4.25
Fixed Effect	axis*stim_size	20.804	2	< 0.0001	43.2
Fixed Effect	axis*ambient_lum	3.237	2	0.1982	311
Fixed Effect	axis*target_id	1121.714	14	< 0.0001	1.26e11
Fixed Effect	axis*start_point	51.016	4	< 0.0001	5.56
Fixed Effect	axis*medium	8.989	2	0.0112	4.49
Fixed Effect	axis*stim_size*target_id	227.055	6	< 0.0001	2.89e4
Fixed Effect	axis*ambient_lum* target_id	121.646	10	< 0.0001	2.85e7
Random Effect	Std. Dev (J')	1.692			
Random Effect	Std. Dev (a')	0.635			
Random Effect	Std. Dev (b')	0.677			

Table B5. Estimated marginal means (EMMs) and pairwise contrasts from the final pooled model.

Table B5.1. Estimated marginal means (EMMs) for all factor levels under standardized reference conditions. Values are predicted differences (match – target) in CIECAM16-UCS coordinates (J' , a' , and b') with standard errors. The standardized reference condition (binocular viewing, small stimulus size, high ambient luminance, mid starting point, AR medium) is listed first, followed by the alternative levels of each significant predictor. Each row corresponds to a specific factor level and target.

Factor and level	Target	EMM ± SE		
		J'	a'	b'
Standardized reference condition	1	-3.52 ± 0.92	0.34 ± 0.37	0.39 ± 0.31
	3	-1.18 ± 0.54	2.27 ± 0.25	-4.52 ± 0.24
	4	-7.34 ± 1.46	2.87 ± 0.47	-1.57 ± 0.81
	6	1.01 ± 0.97	-0.13 ± 0.32	0.76 ± 0.58
	7	-7.76 ± 0.67	1.77 ± 0.22	-0.93 ± 0.29
	9	1.38 ± 0.81	0.19 ± 0.38	-0.42 ± 0.29
	13	0.49 ± 0.62	1.32 ± 0.25	-1.21 ± 0.25
	19	-0.26 ± 0.64	0.40 ± 0.24	-1.24 ± 0.24
view_mode: monocular	3	-0.64 ± 0.59	2.10 ± 0.27	-4.12 ± 0.25
	7	-7.22 ± 0.71	1.61 ± 0.24	-0.52 ± 0.31
	13	1.03 ± 0.62	1.16 ± 0.25	-0.80 ± 0.25
	19	0.28 ± 0.64	0.23 ± 0.24	-0.84 ± 0.24
stim_size: large	3	1.94 ± 0.65	0.23 ± 0.34	-1.09 ± 0.30
	7	-1.35 ± 0.95	1.16 ± 0.27	1.26 ± 0.43
	13	2.69 ± 0.63	1.08 ± 0.24	-1.01 ± 0.24
	19	2.99 ± 0.65	1.56 ± 0.23	-1.42 ± 0.24
ambient_lum: low	1	-2.41 ± 0.62	0.50 ± 0.24	0.53 ± 0.23

	3	-0.54 ± 0.57	1.96 ± 0.28	-4.30 ± 0.26
	4	-9.51 ± 0.79	3.86 ± 0.26	-4.27 ± 0.41
	6	1.12 ± 0.63	0.09 ± 0.22	-0.15 ± 0.30
	7	-7.55 ± 0.74	2.63 ± 0.23	-4.02 ± 0.33
	9	1.07 ± 0.57	-0.60 ± 0.24	-0.48 ± 0.23
start_point: high	1	-3.38 ± 1.06	0.63 ± 0.45	-0.24 ± 0.43
	3	-1.03 ± 0.71	2.55 ± 0.34	-5.16 ± 0.35
	4	-7.19 ± 1.55	3.15 ± 0.53	-2.21 ± 0.86
	6	1.15 ± 1.10	0.15 ± 0.41	0.12 ± 0.65
	7	-7.61 ± 0.82	2.05 ± 0.31	-1.57 ± 0.39
	9	1.52 ± 0.96	0.48 ± 0.46	-1.06 ± 0.42
start_point: low	1	-4.22 ± 0.95	1.00 ± 0.38	0.11 ± 0.33
	3	-1.87 ± 0.59	2.92 ± 0.27	-4.80 ± 0.25
	4	-8.03 ± 1.48	3.52 ± 0.48	-1.85 ± 0.81
	6	0.31 ± 1.01	0.52 ± 0.34	0.48 ± 0.59
	7	-8.45 ± 0.70	2.42 ± 0.24	-1.21 ± 0.31
	9	0.68 ± 0.85	0.85 ± 0.40	-0.70 ± 0.31
medium: display	1	-3.19 ± 0.89	0.16 ± 0.36	-0.02 ± 0.29
	3	-0.84 ± 0.58	2.08 ± 0.27	-4.94 ± 0.25
	4	-7.00 ± 1.44	2.68 ± 0.46	-1.99 ± 0.80
	6	1.35 ± 0.94	-0.32 ± 0.30	0.35 ± 0.57
	7	-7.42 ± 0.70	1.58 ± 0.23	-1.34 ± 0.30
	9	1.72 ± 0.77	0.01 ± 0.37	0.84 ± 0.28

Table B5.2. Pairwise contrasts between factor levels. Contrasts are reported in CIECAM16-UCS coordinates (J' , a' , and b') with standard errors (SE) and p-values. Statistically significant p-values ($p < 0.05$) are highlighted in bold. The Euclidean distance ($\Delta E'$), computed as the Euclidean norm of the three contrast components, is also provided. Dashes (—) indicate that the corresponding factor did not interact with target combination (*target_id*), and in these cases, a single contrast value applies to all targets.

Factor	Target	contrast $\pm$ SE (p-value)			$\Delta E'$
		$\Delta J'$	$\Delta a'$	$\Delta b'$	
view_mode: binocular – monocular	—	-0.54 $\pm$ 0.29 (0.0596)	0.16 $\pm$ 0.12 (0.1693)	-0.41 $\pm$ 0.11 (0.0002)	0.694
stim_size: small – large	3	-3.11 $\pm$ 0.47 (< 0.0001)	2.04 $\pm$ 0.33 (< 0.0001)	-3.43 $\pm$ 0.26 (< 0.0001)	5.064
	7	-6.40 $\pm$ 0.92 (< 0.0001)	0.61 $\pm$ 0.22 (0.0055)	-2.19 $\pm$ 0.44 (< 0.0001)	6.795
	13	-2.20 $\pm$ 0.49 (< 0.0001)	0.25 $\pm$ 0.21 (0.2437)	-0.20 $\pm$ 0.17 (0.2472)	2.221
	19	-3.25 $\pm$ 0.54 (< 0.0001)	-1.17 $\pm$ 0.19 (< 0.0001)	0.18 $\pm$ 0.16 (0.2686)	3.458
ambient_lum: high – low	1	-1.11 $\pm$ 0.82 (0.1761)	-0.16 $\pm$ 0.33 (0.6371)	-0.13 $\pm$ 0.24 (0.5782)	1.131
	3	-0.63 $\pm$ 0.36	0.30 $\pm$ 0.26	-0.22 $\pm$ 0.20	0.738

		(0.0751)	(0.2473)	(0.2708)	
	4	2.18 ± 1.48 (0.1416)	-0.99 ± 0.45 (0.0283)	2.70 ± 0.85 (0.0015)	3.608
	6	-0.11 ± 0.88 (0.8979)	-0.22 ± 0.27 (0.4061)	0.91 ± 0.57 (0.1131)	0.942
	7	-0.20 ± 0.70 (0.7735)	-0.85 ± 0.17 (< 0.0001)	3.09 ± 0.34 (< 0.0001)	3.209
	9	0.31 ± 0.65 (0.6360)	0.79 ± 0.35 (0.0234)	0.05 ± 0.21 (0.7998)	0.851
start_point: mid – high	—	-0.14 ± 0.55 (0.9628)	-0.28 ± 0.26 (0.5140)	0.64 ± 0.31 (0.0969)	0.714
start_point: mid – low	—	0.70 ± 0.26 (0.0200)	-0.65 ± 0.11 (< 0.0001)	0.28 ± 0.11 (0.0304)	0.996
start_point: high – low	—	0.84 ± 0.59 (0.3312)	-0.37 ± 0.28 (0.3682)	-0.36 ± 0.32 (0.5075)	0.986
medium: AR – display	—	-0.34 ± 0.24 (0.1620)	0.18 ± 0.10 (0.0693)	0.41 ± 0.10 (< 0.0001)	0.564

Appendix C: Supplementary Figures

Figure C1. Trajectory plots for Conditions 1–11. The left panels show the $a'-b'$ plane and the right panels show the $J'-b'$ plane for all targets measured in the respective condition. Symbols are defined in the figure legends. Condition definitions are provided in Table IV.

Trajectory - Condition 1

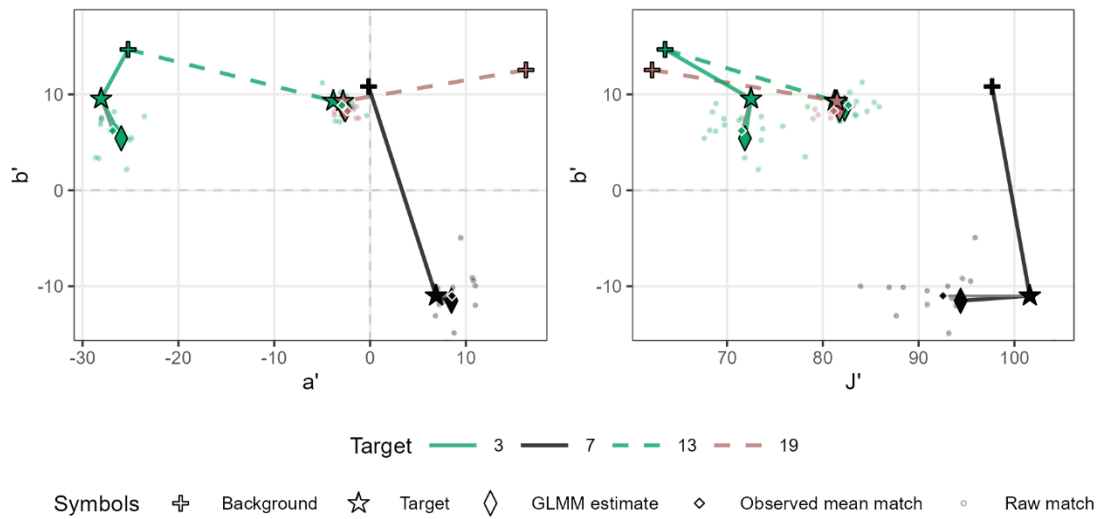


Figure C1.1. Condition 1 (monocular, small stimulus, mid starting point, high luminance, AR).

Trajectory - Condition 2

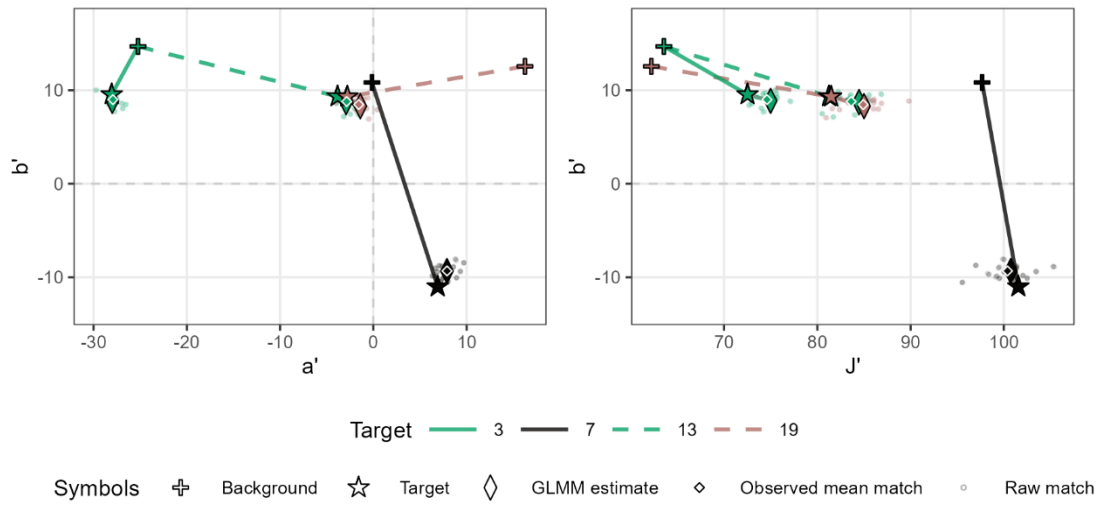


Figure C1.2. Condition 2 (monocular, large stimulus, mid starting point, high luminance, AR).

Trajectory - Condition 3

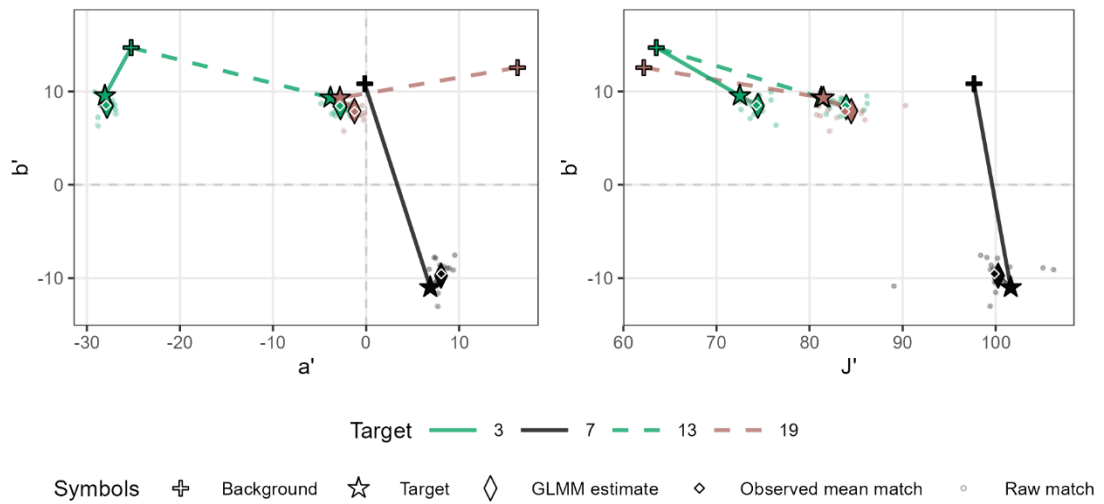


Figure C1.3. Condition 3 (binocular, large stimulus, mid starting point, high luminance, AR).

Trajectory - Condition 4

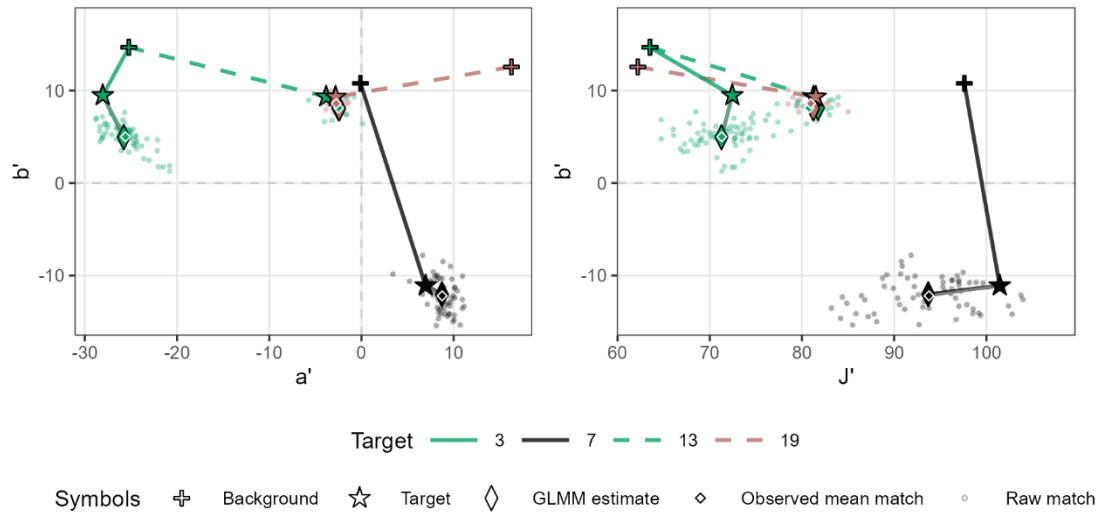


Figure C1.4. Condition 4 (binocular, small stimulus, mid starting point, high luminance, AR).

Trajectory - Condition 5

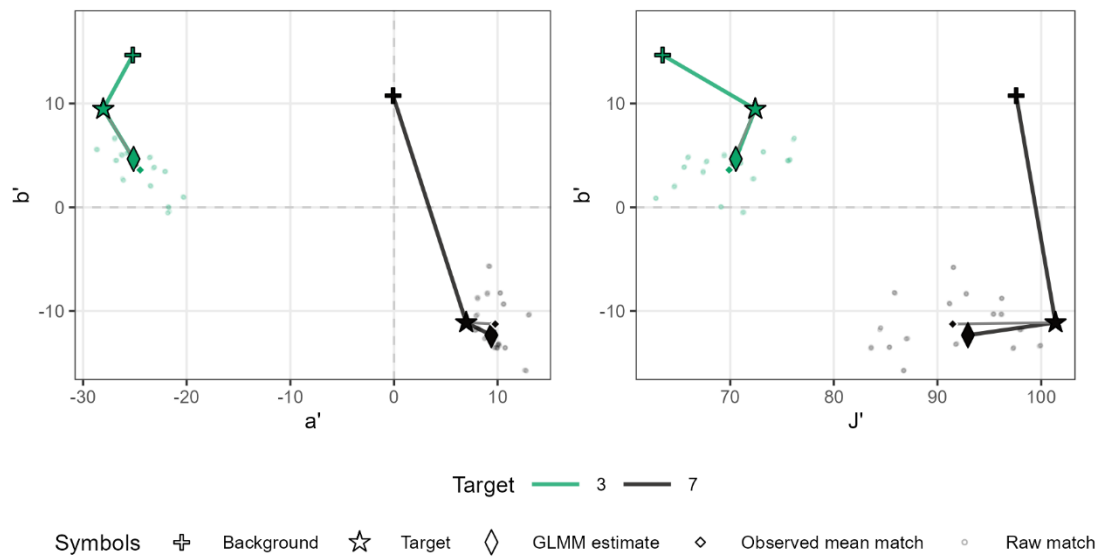


Figure C1.5. Condition 5 (binocular, small stimulus, low starting point, high luminance, AR).

Trajectory - Condition 6

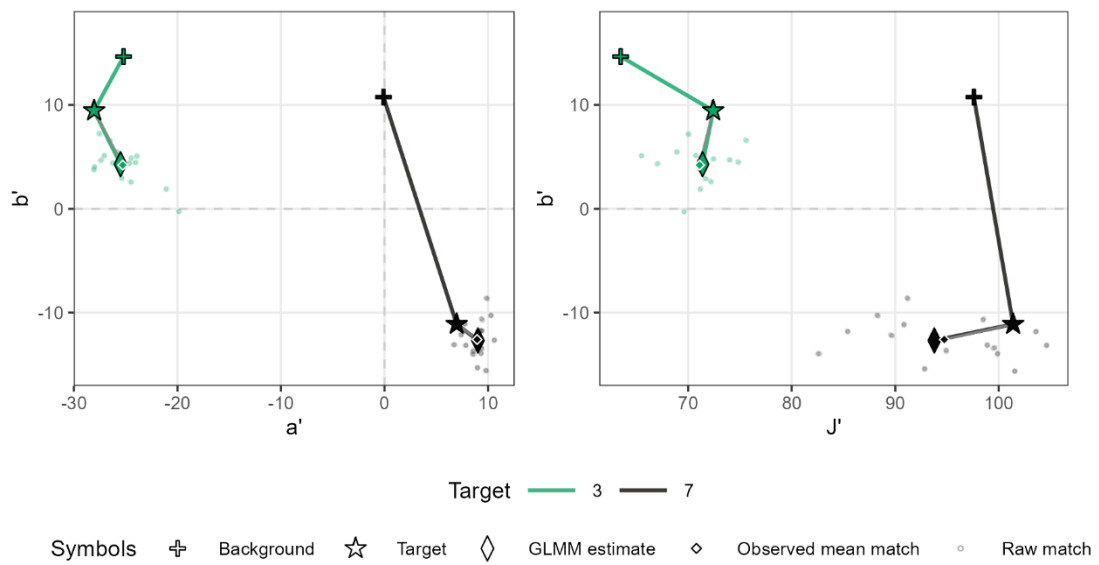


Figure C1.6. Condition 6 (binocular, small stimulus, high starting point, high luminance, AR).

Trajectory - Condition 7

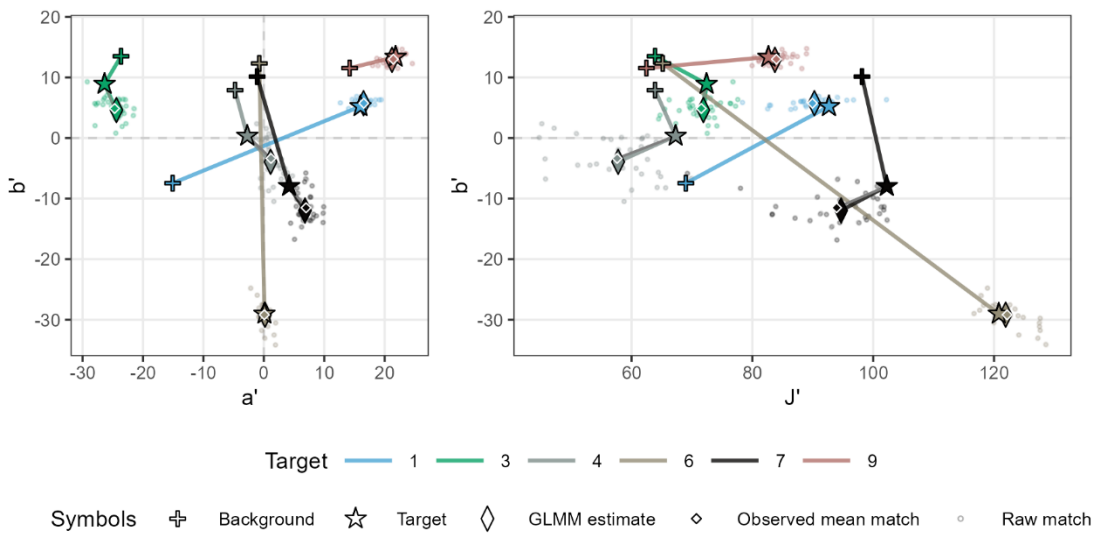


Figure C1.7. Condition 7 (binocular, small stimulus, mid starting point, low luminance, AR).

Trajectory - Condition 8

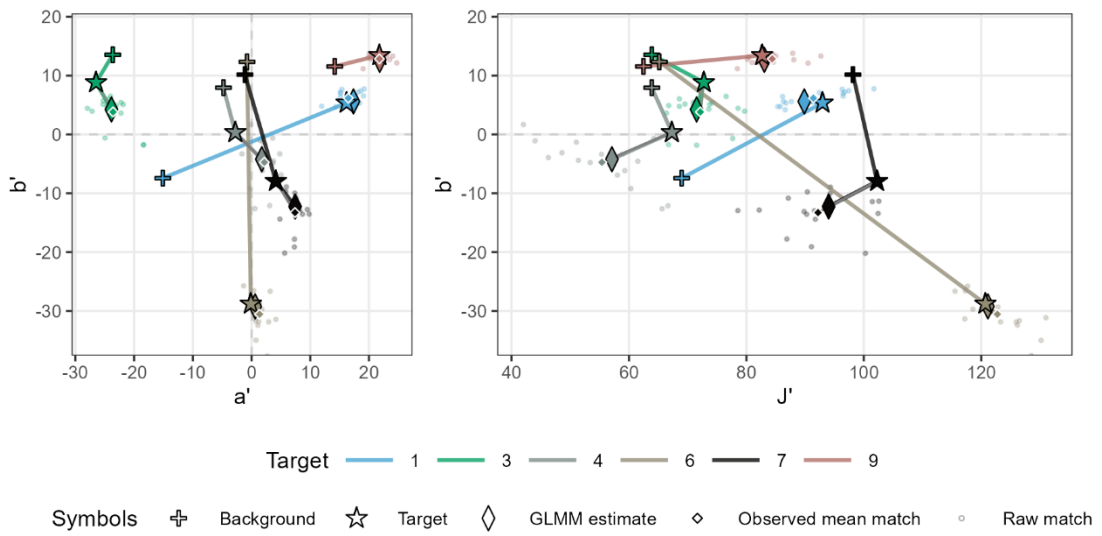


Figure C1.8. Condition 8 (binocular, small stimulus, low starting point, low luminance, AR).

Trajectory - Condition 9

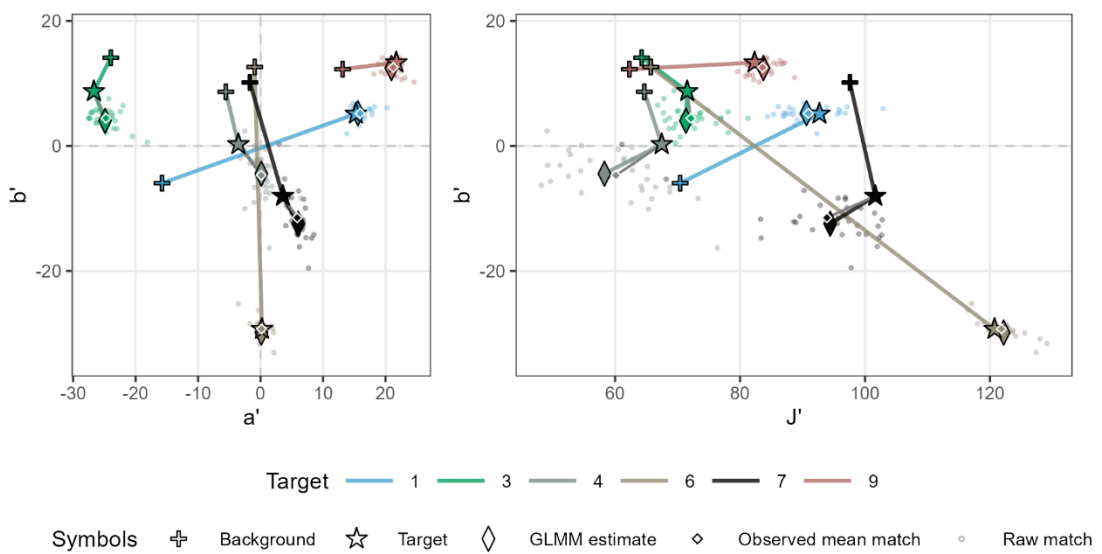


Figure C1.9. Condition 9 (binocular, small stimulus, mid starting point, low luminance, display).

Trajectory - Condition 10

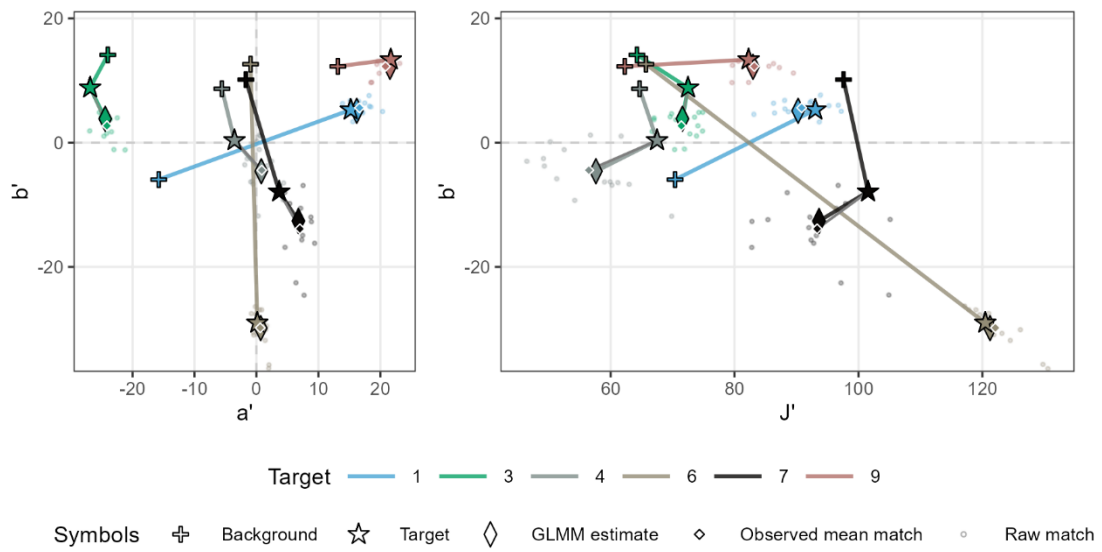


Figure C1.10. Condition 10 (binocular, small stimulus, low starting point, low luminance, display).

Trajectory - Condition 11

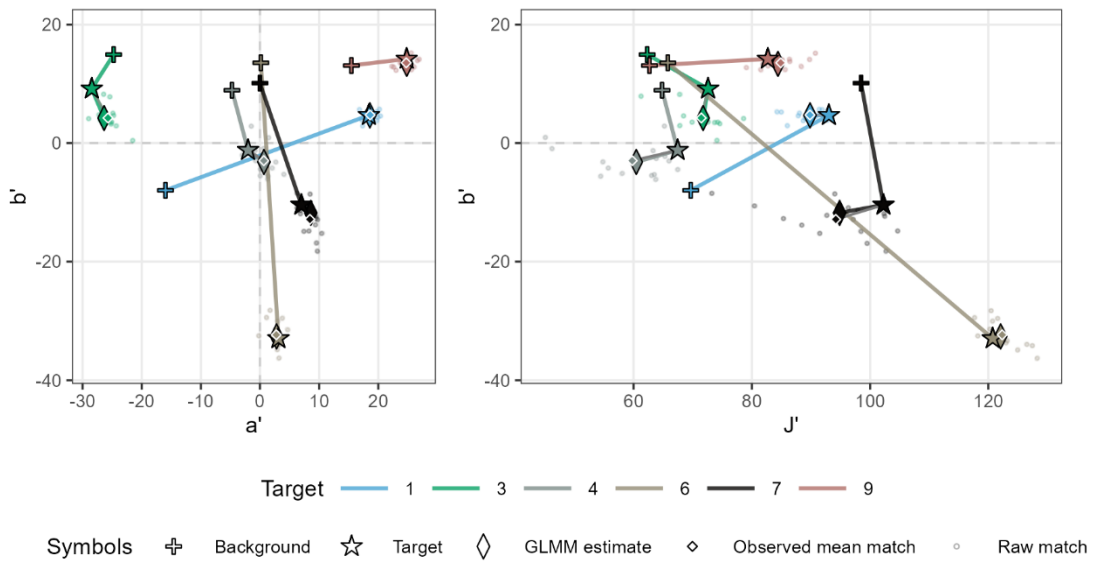


Figure C1.11. Condition 11 (binocular, small stimulus, mid starting point, high luminance, display).

Figure C2. Forest plots of estimated marginal means (EMMs) and raw match distributions across primary predictors. (Note: The results for the ambient luminance factor are presented in Figure 4 in the main text). Horizontal error bars represent 95% confidence intervals derived from model standard errors, while other predictors were held at their baseline levels. The vertical dashed lines indicate the intended target color (zero deviation). For predictors that interact with target combination (*target_id*), results are shown separately for each target color; otherwise, aggregated results across targets are displayed. Symbols are defined in the figure legends.

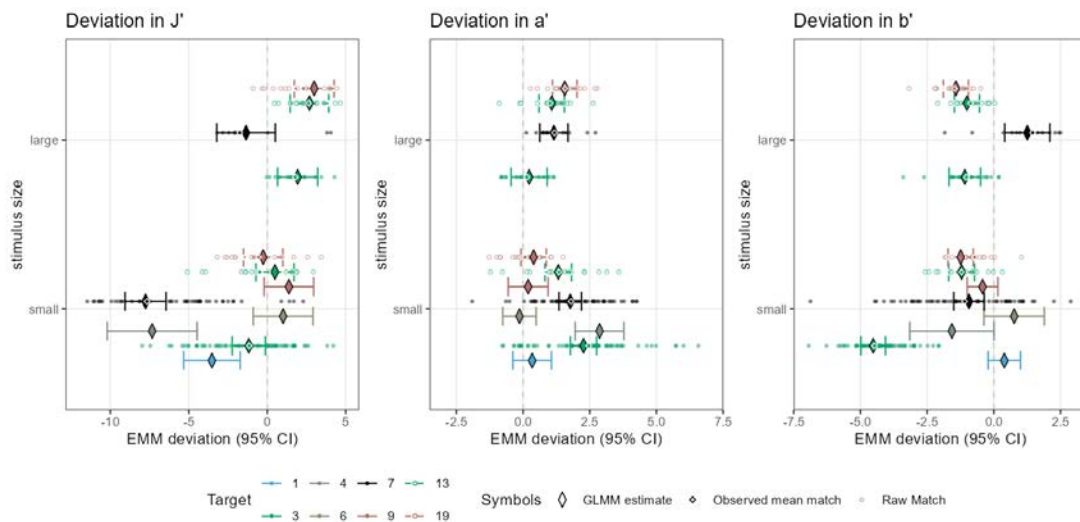


Figure C2.1. Stimulus size (binocular, mid starting point, high luminance, AR).

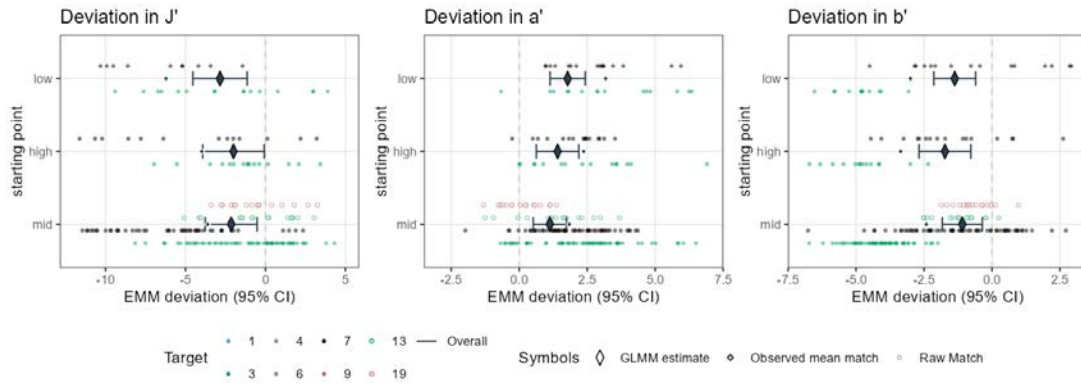


Figure C2.2. Starting point (binocular, small stimulus, high luminance, AR).

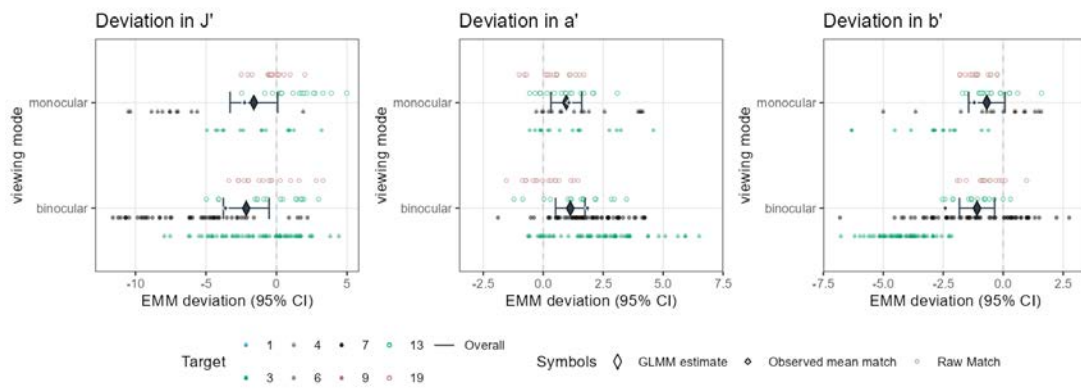


Figure C2.3. Viewing mode (small stimulus, mid starting point, high luminance, AR).

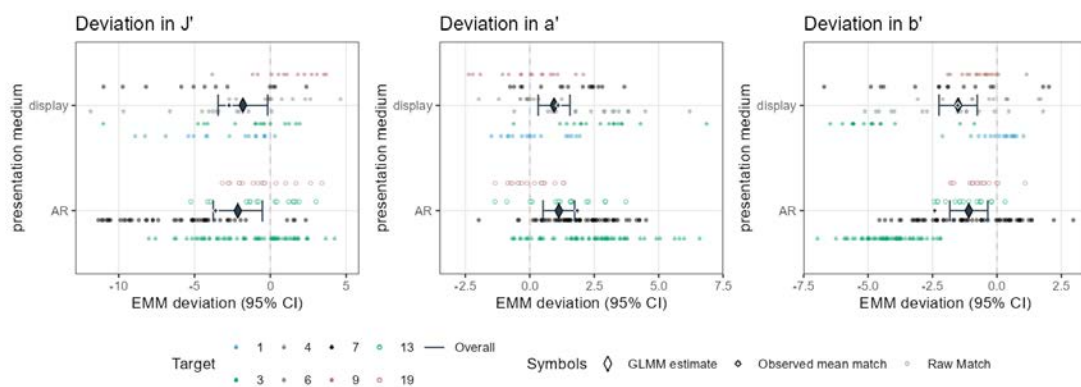


Figure C2.4. Presentation medium (binocular, small stimulus, mid starting point, high luminance).

Figure C3. Forest plots of estimated marginal means (EMMs) for individual targets overall comparison across Conditions 1–11. Figures C3.1–C3.7 show target-wise results for each individual target color (Targets 13 and 19 are presented together in Figure C3.7 for comparison). Figure C3.8 presents the aggregated results across all target colors for overall comparison, without raw data points. Horizontal error bars represent 95% confidence intervals derived from model standard errors. The vertical dashed lines indicate the intended target color (zero deviation). Symbols are defined in the figure legends. Condition definitions are provided in Table IV.

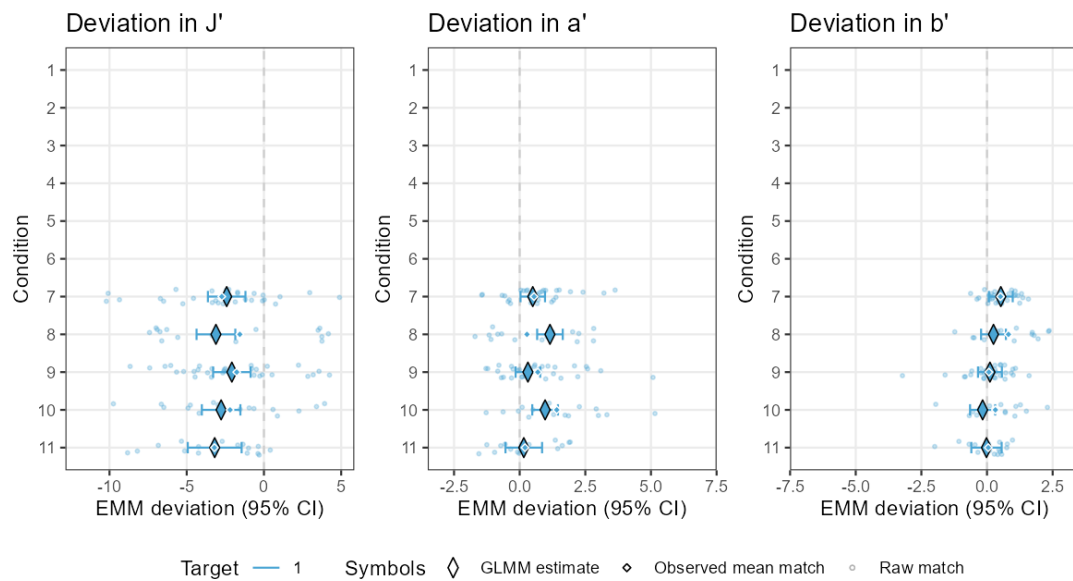


Figure C3.1. Target 1 (BG: blue, FG: orange).

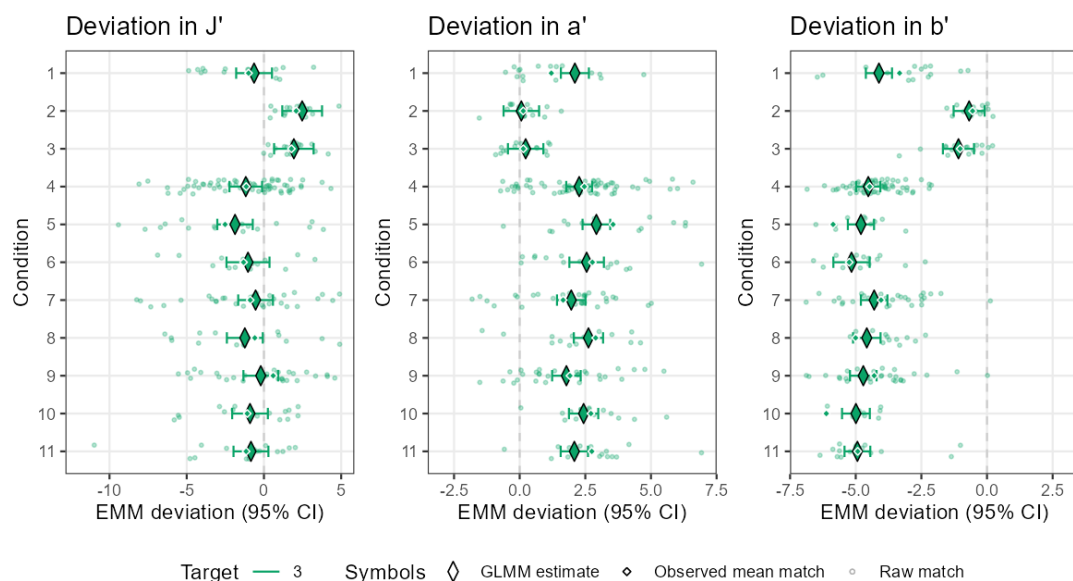


Figure C3.2. Target 3 (BG: green, FG: green).

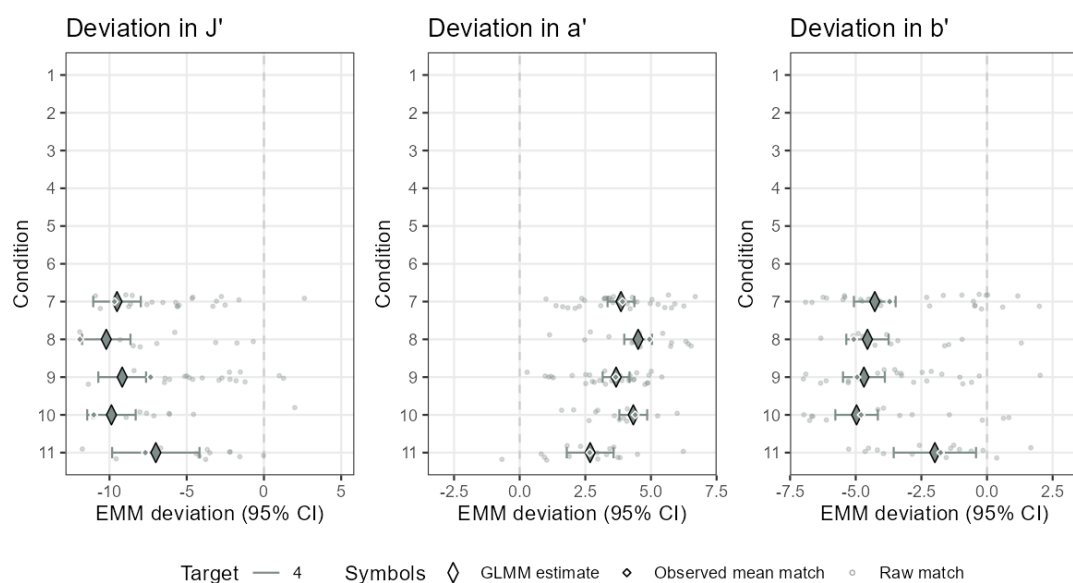


Figure C3.3. Target 4 (BG: gray, FG: purple).

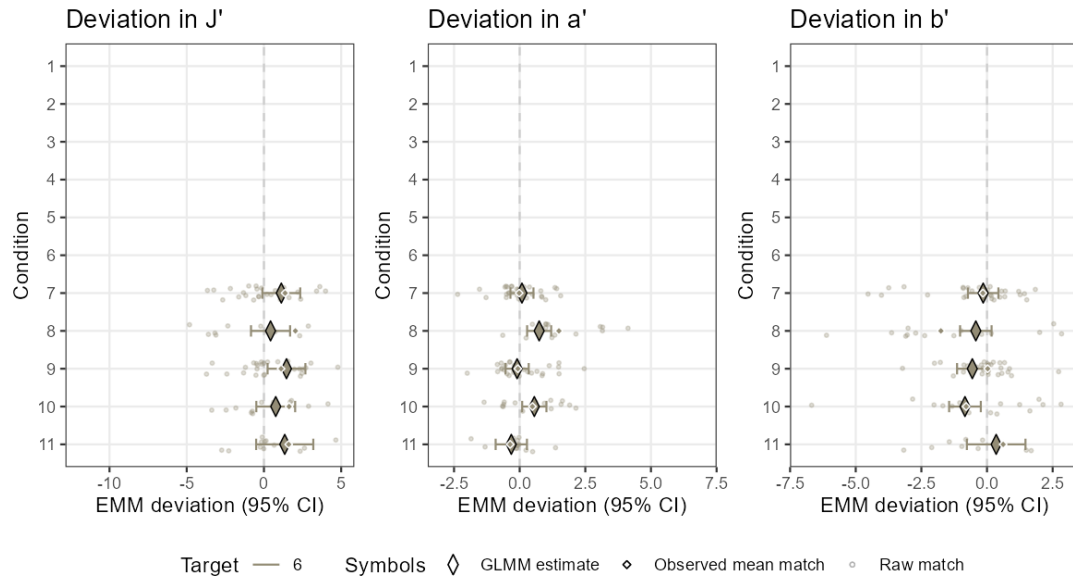


Figure C3.4. Target 6 (BG: brown, FG: white).

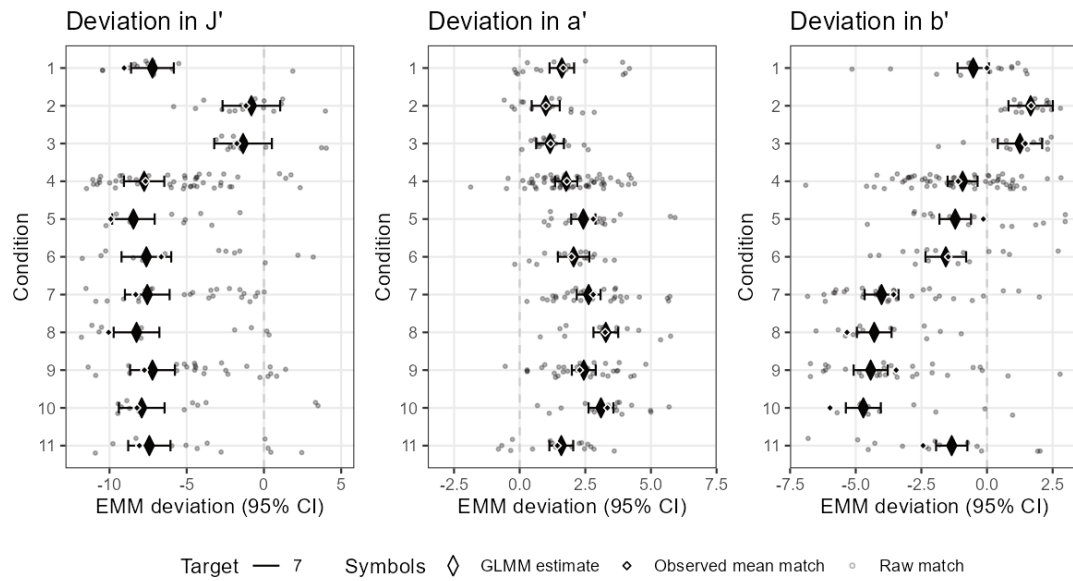


Figure C3.5. Target 7 (BG: white, FG: blue).

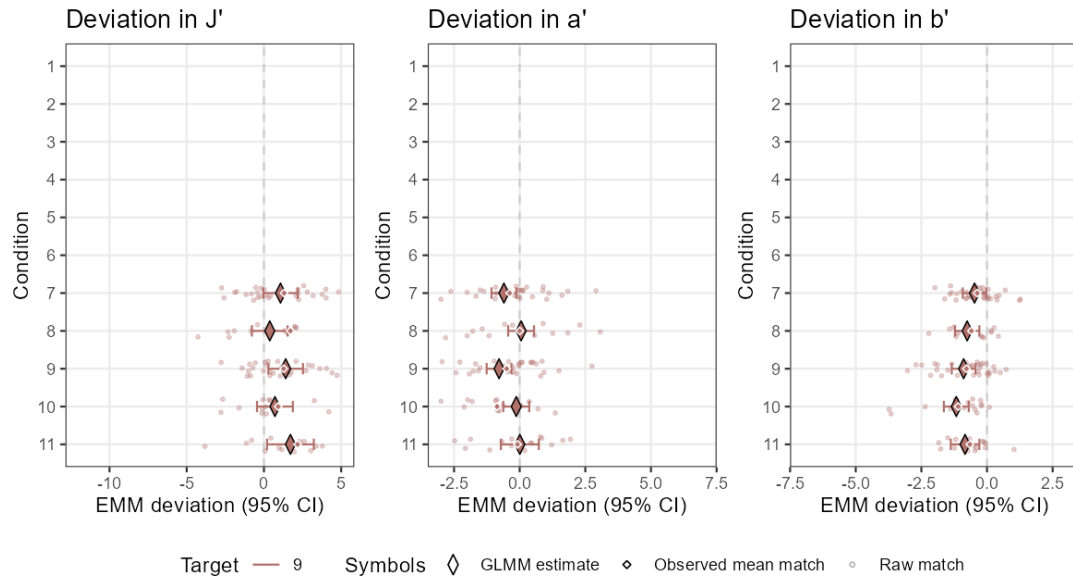


Figure C3.6. Target 9 (BG: red, FG: orange).

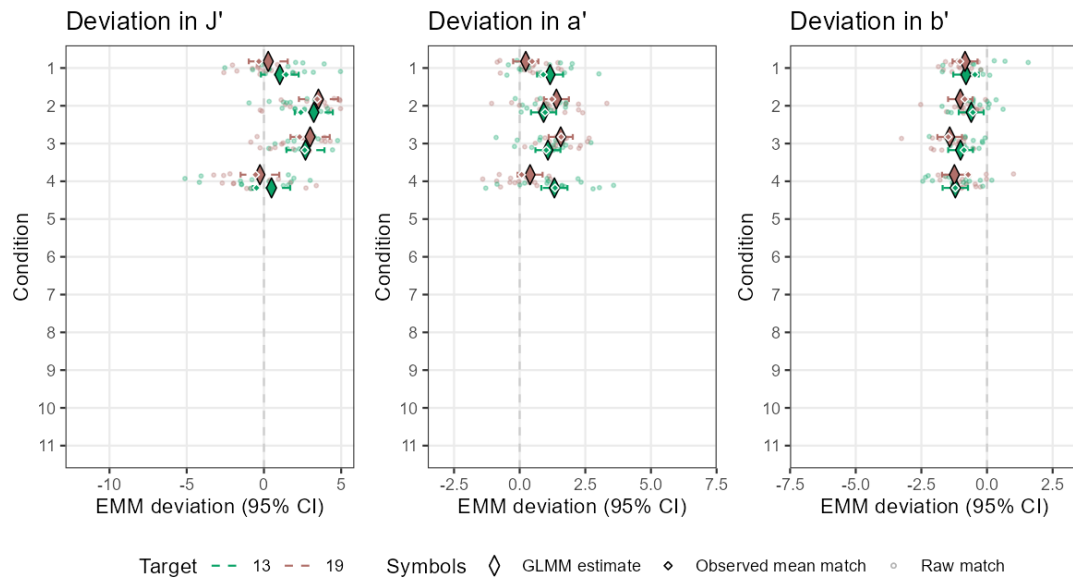


Figure C3.7. Targets 13 (BG: green, FG: red) and 19 (BG: red, FG: green).

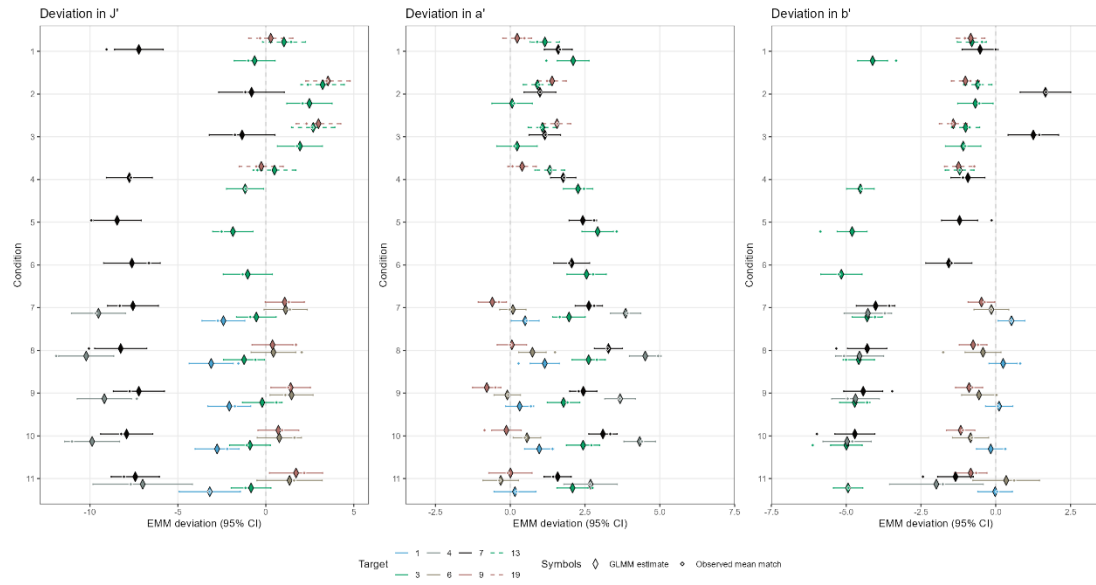


Figure C3.8. Overall comparison across all targets.