Supplementary Material

Relationship of noise estimates between MLDS and MLCM

The statistical models in MLDS and MLCM take the stochasticity of observers' judgments into account. Their parametrizations place an additive and Gaussian noise source at the decision stage.

In MLDS the stimulus triad x_1, x_2 and x_3 evokes a deterministic perceptual response $\Psi(x_1)$, $\Psi(x_2)$, $\Psi(x_3)$ and the observer compares the two perceptual intervals with a differencing rule expressed in the decision variable Δ:

$$\begin{aligned}\Delta_{\text{MLDS}} = &\left[\Psi^i(x_3) - \Psi^i(x_2)\right] \\ &- \left[\Psi^i(x_2) - \Psi^i(x_1)\right] + \epsilon\end{aligned} \tag{S1}$$

where $\Psi^i(x)$ is the perceptual scale in the i-th context, $\epsilon \sim N(0, \sigma^2)$, and σ^2 is the decision noise variance. The observer responds that the pair (x_2, x_3) is perceived as having the biggest perceptual difference if $\Delta_{\text{MLDS}} > 0$, (x_1, x_2) otherwise.

Similarly, for MLCM we consider all possible paired comparisons of context and luminance. Thus we have that the decision variable is

$$\Delta_{\text{MLCM}} = \left[\Psi^j(x_2) - \Psi^i(x_1)\right] + \epsilon \tag{S2}$$

where the observer compares the luminances x_1 and x_2 in contexts i and j, respectively, and $\epsilon \sim N(0, \sigma^2)$. The observer responds that x_2 is perceived as lighter if $\Delta_{\text{MLCM}} > 0$, x_1 otherwise.

Using a binomial generalized linear model (GLM), MLDS and MLCM find the values of $\Psi(x)$ that maximize the likelihood given the observers' responses. Importantly, they also provide an estimate of the decision noise parameter, $\hat{\sigma}$.

We can parametrize the model differently by shifting the noise from the decision stage to the perceptual stage. This parametrization is motivated by the potential comparison of this class of scaling methods with performance-based methods, as they instantiate the same assumptions taken in the simplest version of signal detection theory (equal-variance, independent, Gaussian distributed variables), and thus perceptual scales can be expressed in "d' units" (see applications in e.g. Devinck & Knoblauch, 2012; Aguilar, Wichmann, & Maertens, 2017).

We reparametrize the models by assuming that the noise comes only from the internal dimension $\Psi(x)$ and that there is no decision noise. Each perceptual response is an equal variance, Gaussian distributed random variable with mean equal to the deterministic $\Psi(x)$,

$$\psi^i(x) \sim N(\Psi^i(x), \sigma_P^2) \tag{S3}$$

where σ_P^2 is the noise variance at the perceptual level. Then we can rewrite the decision variables for MLDS and MLCM as follows

$$\Delta_{\text{MLDS}} = \left[\psi^i(x_3) - \psi^i(x_2)\right] - \left[\psi^i(x_2) - \psi^i(x_1)\right] \tag{S4}$$

$$\Delta_{\text{MLCM}} = \left[\psi^j(x_2) - \psi^i(x_1)\right] \tag{S5}$$

MLDS and MLCM provide a noise estimate $\hat{\sigma}$ that reflects the noise of the entire decision model. The next step is to express $\hat{\sigma}$ as a function of the perceptual noise σ_P^2. The decision variables are a linear combination of Gaussian random variables. The variance always adds when linearly combining Gaussian random variables. Thus, the variance of the decision variables is

$$\hat{\sigma}^2_{\text{MLDS}} = 4\sigma_P^2 \tag{S6}$$

$$\hat{\sigma}^2_{\text{MLCM}} = 2\sigma_P^2 \tag{S7}$$

as in the decision variable for MLDS there are four terms, and in the decision variable for MLCM only two terms. Assuming that MLCM and MLDS probe the same internal dimension $\Psi(x)$, we can now put these last two equations together, and we obtained the noise relationship between MLCM and MLDS

$$\hat{\sigma}_{\text{MLCM}} = \frac{\sqrt{2}}{2}\,\hat{\sigma}_{\text{MLDS}} \tag{S8}$$

Goodness of fit

Evaluating the goodness-of-fit (GoF) of general linear models (GLMs) with binomial responses (as in MLDS and MLCM) is not as straightforward as with Gaussian data. We followed the approach suggested by Knoblauch and Maloney (2012) and Wood (2006) to analyze the appropriateness of our scale estimates. The procedures are implemented in the *MLDS* and *MLCM* R packages (Knoblauch & Maloney, 2012). Using bootstrapping we generated responses from the empirically estimated scales. We then fit scales to these simulated data and calculated the deviance residuals for each of the bootstrapped datasets. Finally we compared the actual deviance residuals to the distribution of the simulated residuals. There are two ways to check goodness of fit.

In the first comparison, the cumulative distribution of empirically observed residuals is plotted together with a 95 % confidence envelope which is obtained from the simulated residuals (Fig. S1A). If our distributional assumption is correct, then the envelope represents the residual distribution. For a satisfactory goodness of fit the empirical scale should fall inside the envelope. We quantified this GoF test by calculating the percentage of trials (P_{in}) in which the empirical residuals were located inside the envelopes. Additionally, the plotting routine detects and visualizes individual trials with high residuals which may point to outliers (see below).

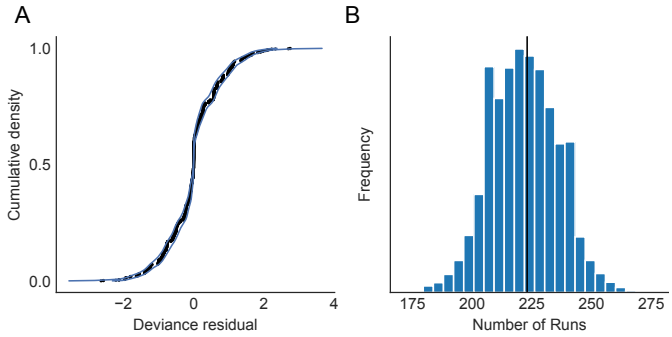

Figure S1: Example of goodness of fit evaluation for one perceptual scale. (A) Cumulative distribution of residuals for the actual data (black dots) and its 95 % confidence envelope (blue lines) calculated from bootstrap simulation. (B) Number of zero-crossings ('runs') for the observed data (black line) and distribution of zero-crossings from bootstrap simulation (blue histogram).

The second comparison tests for systematic patterns in the residuals. In a simple linear model ($y = ax+b+\epsilon \quad \epsilon, \sim N(0, \sigma^2)$) the residuals ($y-\hat{y}$) are plotted against the predicted values ($\hat{y}$) to check that the residuals scatter randomly around zero. Analogously, for binomial GLMs we sort the residuals according to the fitted values and check for sequences of only positive or only negative deviance residuals. Practically we calculate the number of times that adjacent residuals cross zero. To put the empirically observed number of zero-crossings into perspective we compare it with the distribution of zero- crossings in the simulated data (Fig. S1B). The observed value should be within the probability mass of the simulated distribution, i.e. their p-value is not less than 0.01, indicating a random distribution of residuals and hence a satisfactory goodness-of-fit.

On a first evaluation we found that 91 % (29 out of 32) of the cases scales in variegated checkerboards had an satisfactory goodness of fit, against 41 % (13 out of 32) of the cases for the center-surround stimulus (see Suppl. Tables S1 and S2 for details). As many scales did not have a satisfactory goodness of fit, we employed an outlier removal procedure. The procedure is described in Knoblauch and Maloney (2012, pp. 219-222) and it consists of removing the trials which are flagged as having a deviance residual higher than an arbitrary threshold (set to ± 2 in our case). The trials flagged as 'outliers' can be seen in the first comparison plot (Fig. S1A) as datapoints with x-values to the extreme left or right, corresponding to a deviance residual outside the range $[-2, 2]$.

We identified outlier trials in all scaling data using the same criterion as described above, we removed them from their respective datasets, and refitted the models (MLDS or MLCM) to find new scale estimates. These newly estimated scales passed the goodness of fit test in 100 % of the cases for variegated checkerboards, and in 94 % (30 out of 32) of the cases for center-surround stimuli (scales for observers O4 and O5 in 'plain view' did not pass). These updated scales are the ones reported in all results in the main text.

Independent, additive and saturated models in MLCM

MLCM includes three different statistical models: the *independent*, the *additive*, and the *saturated* model. The independent model considers that perceptual judgments can be explained by only one of the stimulus dimensions. The additive model considers that judgments can be explained by a sum of the effects from each stimulus dimension. And finally the saturated model is the most general and it considers the contribution of each stimulus combination presented to the observer, thus allowing interactions that are more complex than a sum among stimulus dimensions. The model is 'saturated' as it has the maximum number of possible parameters. These three models are nested, being the saturated the most general, followed by the additive and finally the independent model. Model selection among these three options is done using the nested likelihood ratio test. More details can be found in Knoblauch and Maloney (2012) and in Gerardin, Devinck, Dojat, and Knoblauch (2014).

For our current use of MLCM we considered all three models. As a reminder, the two stimulus dimensions in our study were luminance and a categorical dimension for viewing context (plain, dark transparency, light transparency). And we evaluated three observer models: lightness constant, luminance-based and contrast-based models.

MLCM with the independent model is appropriate only for the luminance-based observer, as judgments would only depend on one stimulus dimension,

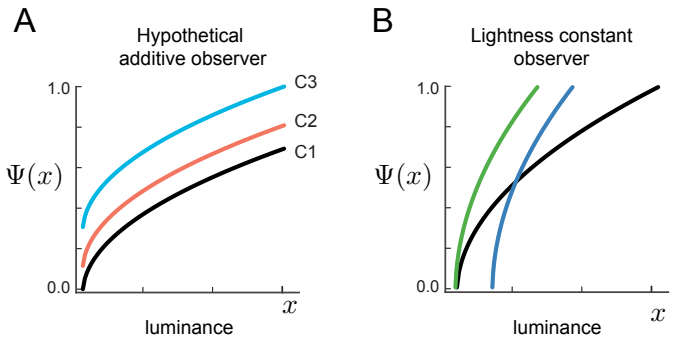

Figure S2: (A) An hypothetical observer model for which additive conjoint measurement would be appropriate. The context dimension (C1-C3) modulates the effect of the other dimension (luminance, x) by an additive factor. (B) Contrarily, luminance range is reduced by the introduction of a transparency (Fig. 2B main text) and the lightness constant observer model expands this range using a multiplicative factor which cannot be captured by the additive model.

i.e. luminance. It is however not appropriate for any other observer model or any other case that we can consider. Thus, we evaluated the next more general model, the additive model.

The additive model would be appropriate if the judgments would depend on the luminance dimension plus an additive offset due to context. An additive offset is however not enough in capturing the interaction of the luminance dimension with the context dimension. In fact, for the lightness constant observer, luminance is transformed to perceived lightness by a multiplicative factor. This multiplicative factor cannot be modeled with additive conjoint measurement (see Fig. S2 for an illustration). Consequently we used MLCM with the saturated model. In this way we cover all scenarios using the most general model.

Estimation issues in MLCM

During our use of MLCM we encountered technical issues worth of mentioning. The general linear model (GLM) - the model underlying the scales' estimation in MLCM and MLDS - often outputs a 'complete separation' warning, which indicate that the model is ill-constrained, having parameters that completely determine the outcome variable. This is not an uncommon occurance in binomial regression (Kosmidis & Firth, 2009). In both MLCM and MLDS the problem can arise when measured proportions of observers judgments are either all 0% or 100%. That could occur when the stimulus spacing is too coarse relative to the observer's internal noise, and thus all judgments are too easy for the observer.

In the present study the 'complete separation' problem did not affect the point estimates of the scales, but it became problematic when calculating confidence intervals using bootstrap. It led to unstable bootstrap estimates and a multi-modal bootstrap distribution, which in turn led to highly skewed confidence intervals. We overcame the problem by setting a coarser stopping rule in the optimization algorithm underlying the GLM fitting routine, in this way avoiding the multi-modality. Alternatively other GLM fitting methods that tackle specifically the problem of 'complete separation' could be used, for example the bias reduction methods implemented in the package *brglm2* (Kosmidis, Pagui, & Sartori, 2018).

References

Aguilar, G., Wichmann, F. A., & Maertens, M. (2017). Comparing sensitivity estimates from MLDS and forced-choice methods in a slant-from-texture experiment. *Journal of Vision*, *17*(1), 37. doi: 10.1167/17.1.37

Devinck, F., & Knoblauch, K. (2012). A common signal detection model accounts for both perception and discrimination of the watercolor effect. *Journal of vision*, *12*(3), 1–14.

Gerardin, P., Devinck, F., Dojat, M., & Knoblauch, K. (2014). Contributions of contour frequency, amplitude, and luminance to the watercolor effect estimated by conjoint measurement. *Journal of Vision*, *14*(4), 9–9.

Knoblauch, K., & Maloney, L. T. (2012). *Modeling Psychophysical Data in R*. Springer New York.

Kosmidis, I., & Firth, D. (2009). Bias reduction in exponential family nonlinear models. *Biometrika*, *96*(4), 793–804. doi: 10.1093/biomet/asp055

Kosmidis, I., Pagui, E. C. K., & Sartori, N. (2018). *Mean and median bias reduction in generalized linear models.* Retrieved from `https://arxiv.org/abs/1804.04085`

Wood, S. (2006). *Generalized additive models : an introduction with r*. Boca Raton, FL: Chapman & Hall/CRC.

Variegated checkerboards

observer	dataset experiment	context	before outlier removal GoF measure P_{in}	before outlier removal GoF measure p	after outlier removal GoF measure P_{in}	after outlier removal GoF measure p
O1	MLDS	plain	96.5	0.54	99.4	0.94
		dark	93.9	0.27	97.5	0.56
		light	89.8	0.08	99.2	0.79
	MLCM	all	97.1	0.06	95.5	0.71
O2/MM	MLDS	plain	99.3	0.57	99.6	0.69
		dark	96.1	0.09	99.1	0.51
		light	99.4	0.95	98.1	0.99
	MLCM	all	99.7	0.71	99.2	0.89
O3/GA	MLDS	plain	45.17	0.27	63.74	0.41
		dark	24.75	0.20	58.35	0.54
		light	22.58	0.02	46.53	0.11
	MLCM	all	94.33	0.04	94.25	0.72
O4/MK	MLDS	plain	99.08	0.05	99.75	0.36
		dark	99.33	0.71	99.92	0.94
		light	99.42	0.54	97.63	0.78
	MLCM	all	96.54	< 0.01 *	98.92	0.80
O5	MLDS	plain	98.83	0.14	99.58	0.58
		dark	99.08	0.10	98.04	0.40
		light	99.17	0.04	99.58	0.12
	MLCM	all	99.64	0.44	99.67	0.77
O6	MLDS	plain	52.67	0.05	68.22	0.27
		dark	97.25	0.03	96.27	0.65
		light	69.58	0.01	76.63	0.20
	MLCM	all	96.72	< 0.01 *	99.43	0.85
O7	MLDS	plain	94.33	< 0.01 *	99.25	0.51
		dark	98.83	0.22	99.16	0.725
		light	99.08	0.69	98.90	0.95
	MLCM	all	96.93	0.85	61.30	0.95
O8	MLDS	plain	98.67	0.15	99.75	0.63
		dark	98.92	0.32	97.02	0.58
		light	99.00	0.29	97.02	0.76
	MLCM	all	99.46	0.10	99.85	0.93

Table S1: Goodness of fit measures for scales in variegated checkerboards before and after outlier removal. P_{in}: percentage of residuals inside envelope (see Fig. S1A), p: p-value statistic of zero-crossings distribution (see Fig. S1B). Asterisks mark the cases with inappropriate goodness of fit. Detailed description can be found in the text.

Center-surround stimuli

observer	dataset experiment	context	before outlier removal GoF measure P_{in}	before outlier removal GoF measure p	after outlier removal GoF measure P_{in}	after outlier removal GoF measure p
O1	MLDS	plain	98.83	0.32	98.73	0.73
		dark	85.17	0.06	70.55	0.42
		light	96.25	0.09	99.40	0.43
	MLCM	all	88.48	< 0.01 *	98.70	0.91
O2/MM	MLDS	plain	99.25	0.01	99.07	0.25
		dark	99.25	0.21	98.24	0.54
		light	98.42	0.14	99.16	0.67
	MLCM	all	99.22	0.43	99.49	0.65
O3/GA	MLDS	plain	92.17	< 0.01 *	97.05	0.04
		dark	95.33	0.02	97.72	0.22
		light	99.50	0.06	99.41	0.39
	MLCM	all	96.51	< 0.01 *	99.67	0.58
O4/MK	MLDS	plain	97.58	< 0.01 *	98.82	< 0.01 *
		dark	93.17	< 0.01 *	98.57	0.22
		light	96.00	< 0.01 *	97.81	0.08
	MLCM	all	93.16	< 0.01 *	99.40	0.34
O5	MLDS	plain	87.08	< 0.01 *	98.30	< 0.01 *
		dark	96.50	< 0.01 *	94.59	0.05
		light	93.67	< 0.01 *	97.80	0.03
	MLCM	all	99.76	0.28	99.46	0.95
O6	MLDS	plain	83.08	< 0.01 *	95.47	0.04
		dark	96.75	0.01	98.82	0.51
		light	71.67	0.02	90.14	0.30
	MLCM	all	96.63	< 0.01 *	99.61	0.76
O7	MLDS	plain	99.25	0.03	99.16	0.16
		dark	97.83	< 0.01 *	98.82	0.15
		light	99.25	0.01	98.31	0.01
	MLCM	all	94.27	< 0.01 *	95.46	0.61
O8	MLDS	plain	96.33	< 0.01 *	96.93	0.04
		dark	53.25	< 0.01 *	67.52	0.22
		light	84.42	< 0.01 *	93.32	0.06
	MLCM	all	97.34	< 0.01 *	96.83	0.45

Table S2: Goodness of fit measures for scales in center-surround stimili, same presentation format as Table S1.

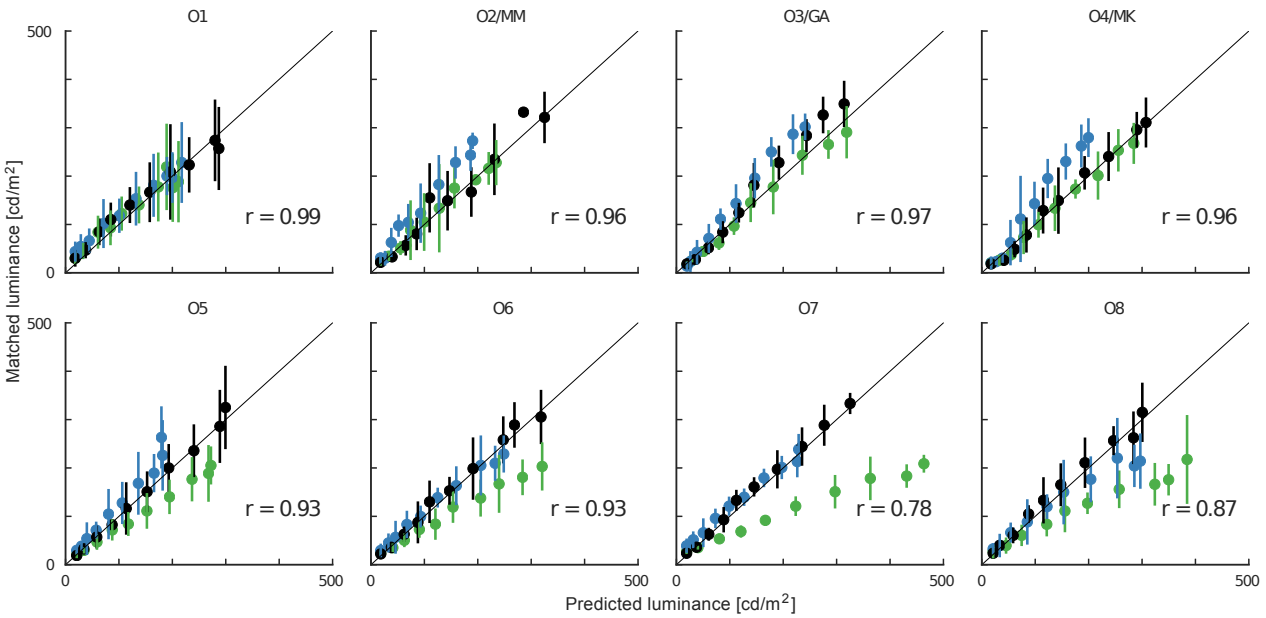

Figure S3: Consistency between matching data and prediction from MLDS perceptual scales in variegated checkerboards, for each observer individually. Datapoints represent mean ± 95 % C.I.

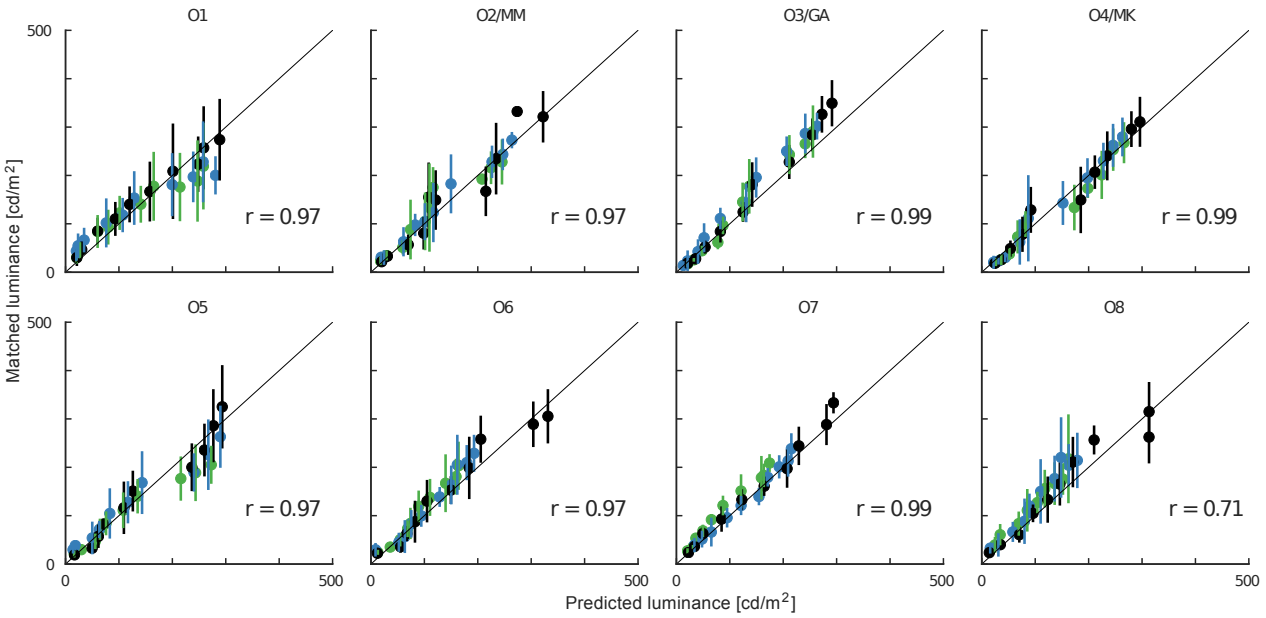

Figure S4: Individual observer data. Consistency between matching data and prediction from MLCM perceptual scales in variegated checkerboards, for each observer individually. Datapoints represent mean ± 95 % C.I.

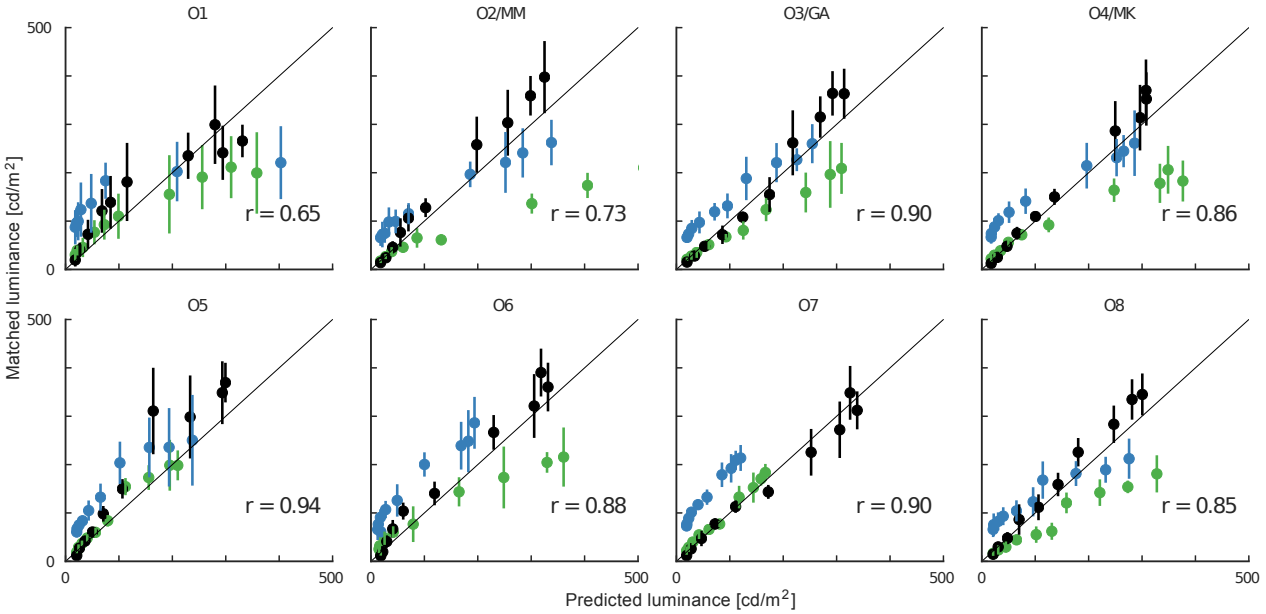

Figure S5: Similar to S3 but for center-surround stimuli.

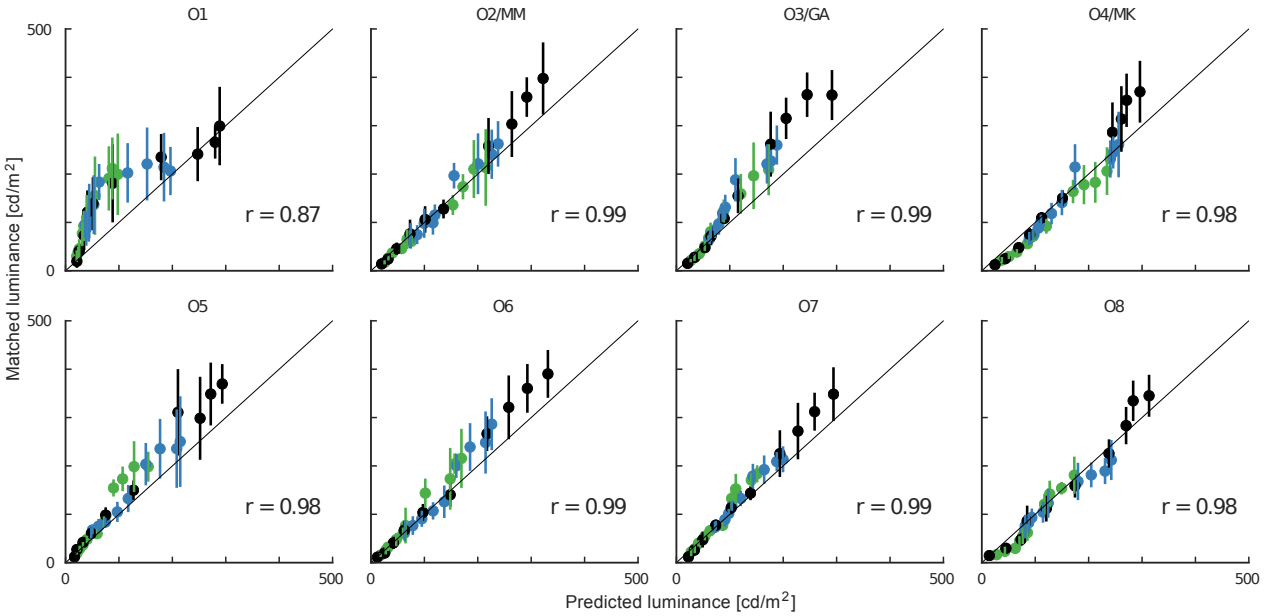

Figure S6: Similar to S4 but for center-surround stimuli.

r	Luminance $[cd/m^2]$		
	plain	dark transparency	light transparency
0.06	15	18	69
0.11	25	22	73
0.19	40	28	79
0.31	60	36	87
0.46	89	48	99
0.63	120	60	111
0.82	155	74	125
1.05	199	92	144
1.29	242	108	159
1.50	281	125	176
1.67	312	137	188
1.95	365	157	209
2.22	415	177	229
mean	174	83	135

Table S3: Target luminance values in different viewing conditions. The first column contains the 13 target reflectance values (r, in *povray* units). The next three columns contain the corresponding luminance values (in cd/m^2) for each viewing condition. Reflectances from 0.11 to 1.67 were used as targets in both experiments.