Supplemental Material

Reprioritization of features of multi-dimensional objects stored in visual working memory

Park, Sy, Hong & Tong

This supplemental material contains:

(1) Response error histograms and best-fitting mixture distributions for Experiments 1–3 (Figure S1–S6)

NOTE: For illustrative purposes, all histograms have 40 bins, with a bin width of 9˚ for color, 4.5˚ for orientation. The mixture model was fitted to raw error data, not the histogram.

(2) Procedure for conducting an a priori power analysis for Experiment 3 (Supplemental Text and Figure S7–S9)

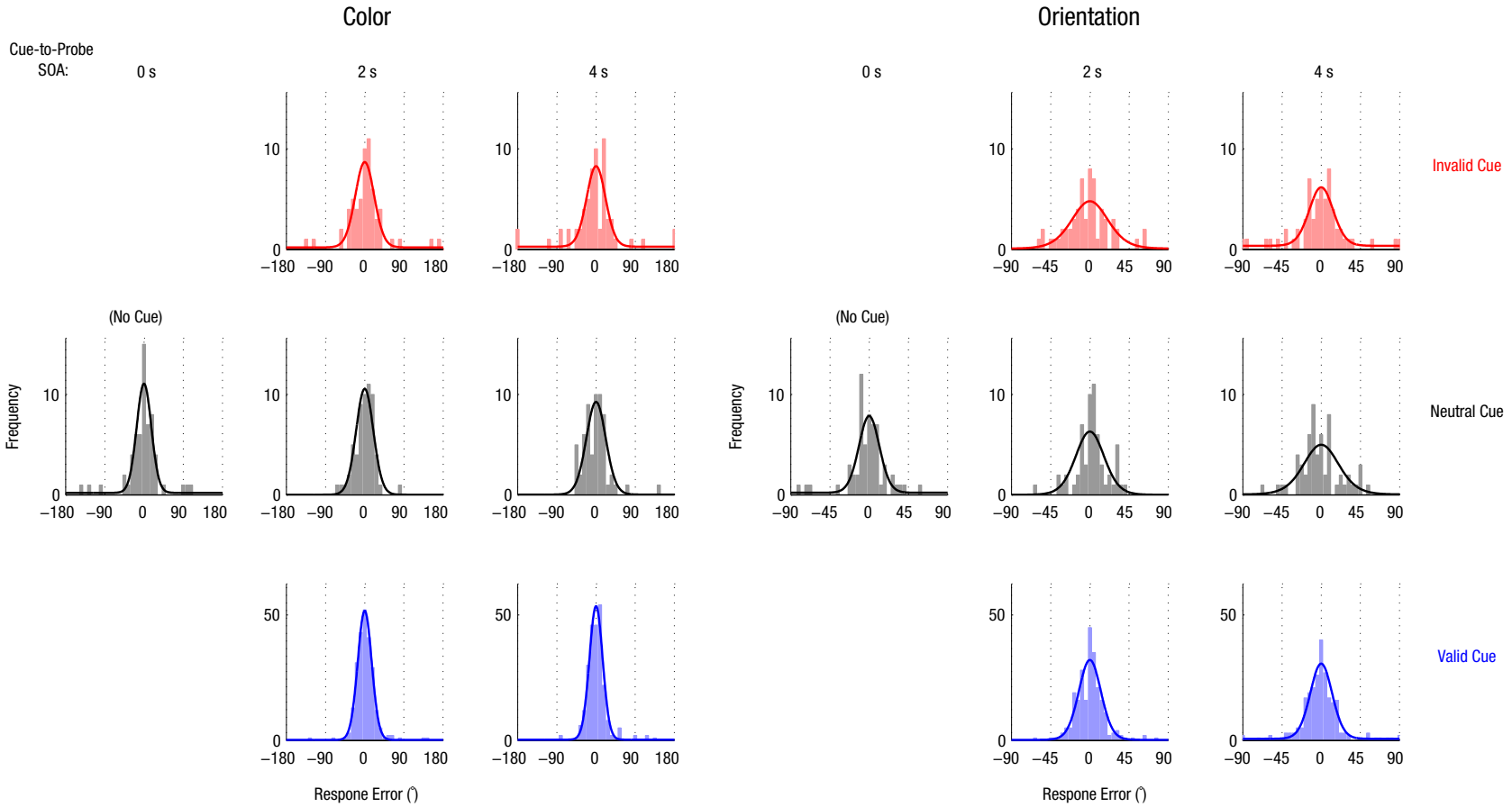

Figure S1. Histogram of response errors for a representative participant in Experiment 1. Response error distribution is shown for each feature dimension, cue type, and cue-to-probe SOA combination. The solid line overlaid on each histogram shows the best-fitting mixture distribution for the respective condition for this participant.

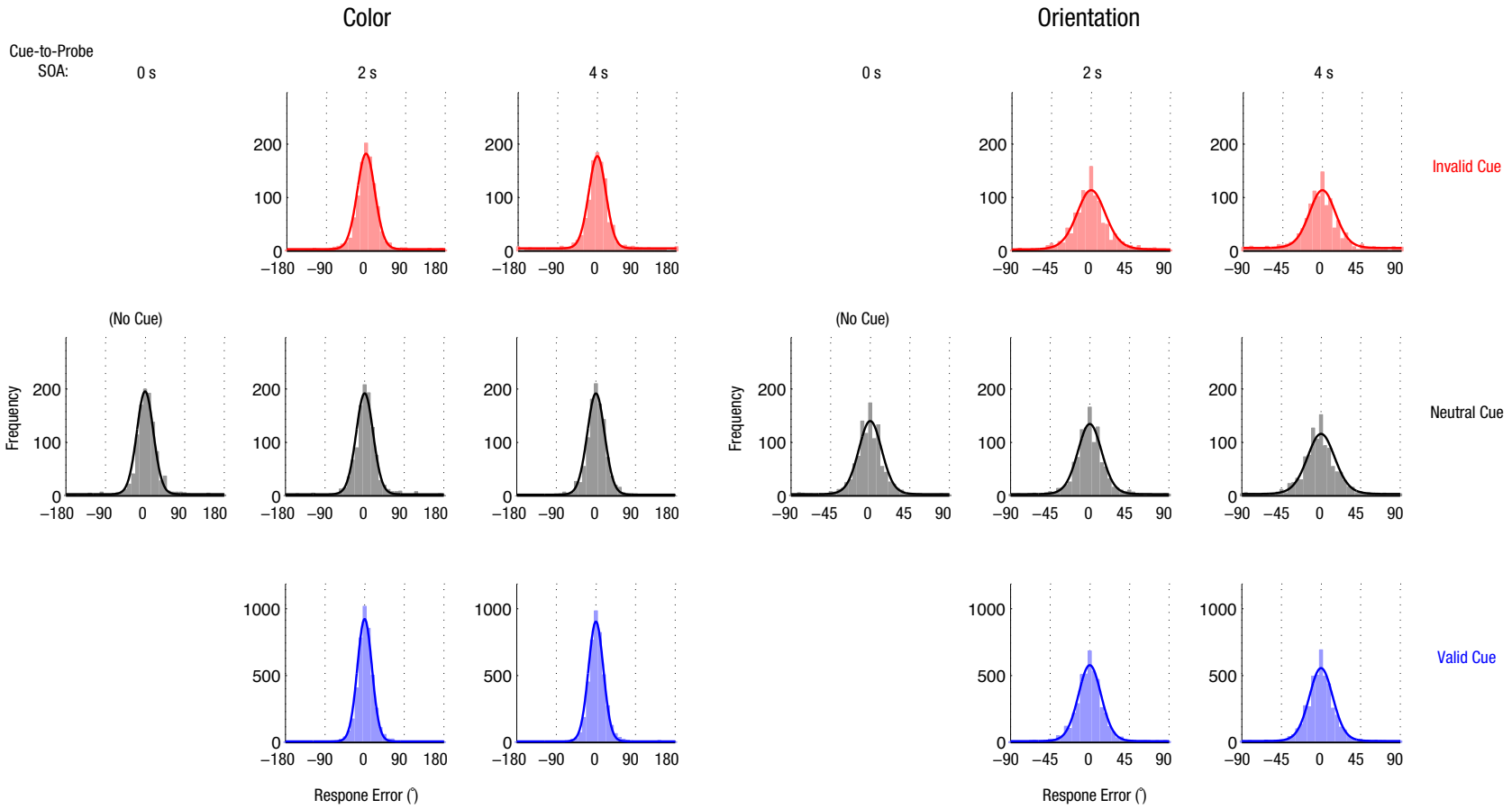

Figure S2. Histogram of response errors pooled across all participants (N = 19) in Experiment 1. Response error distribution is shown for each feature dimension, cue type, and cue-to-probe SOA combination. The solid line overlaid on each histogram shows the mixture distribution based on the average of individual participants' best-fitting parameters.

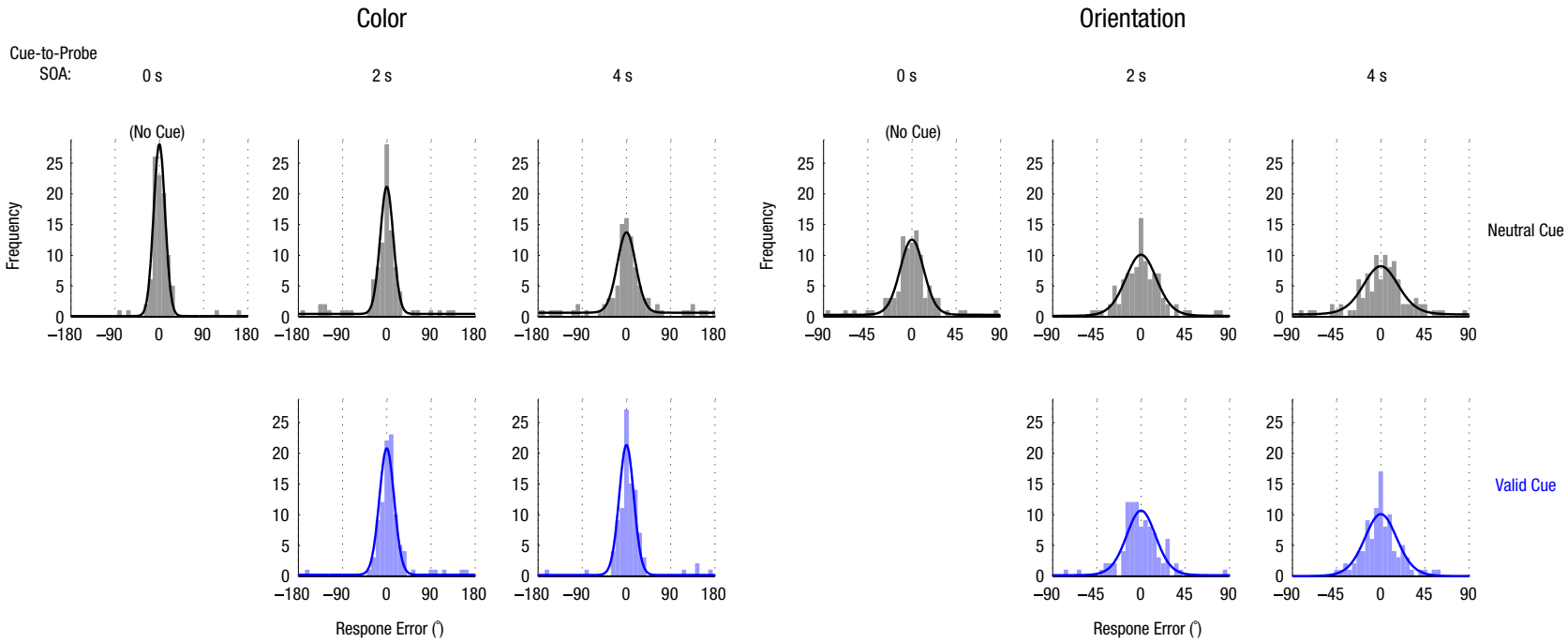

Figure S3. Histogram of response errors for a representative participant in Experiment 2. Response error distribution is shown for each feature dimension, cue type, and cue-to-probe SOA combination. The solid line overlaid on each histogram shows the best-fitting mixture distribution for the respective condition for this participant.

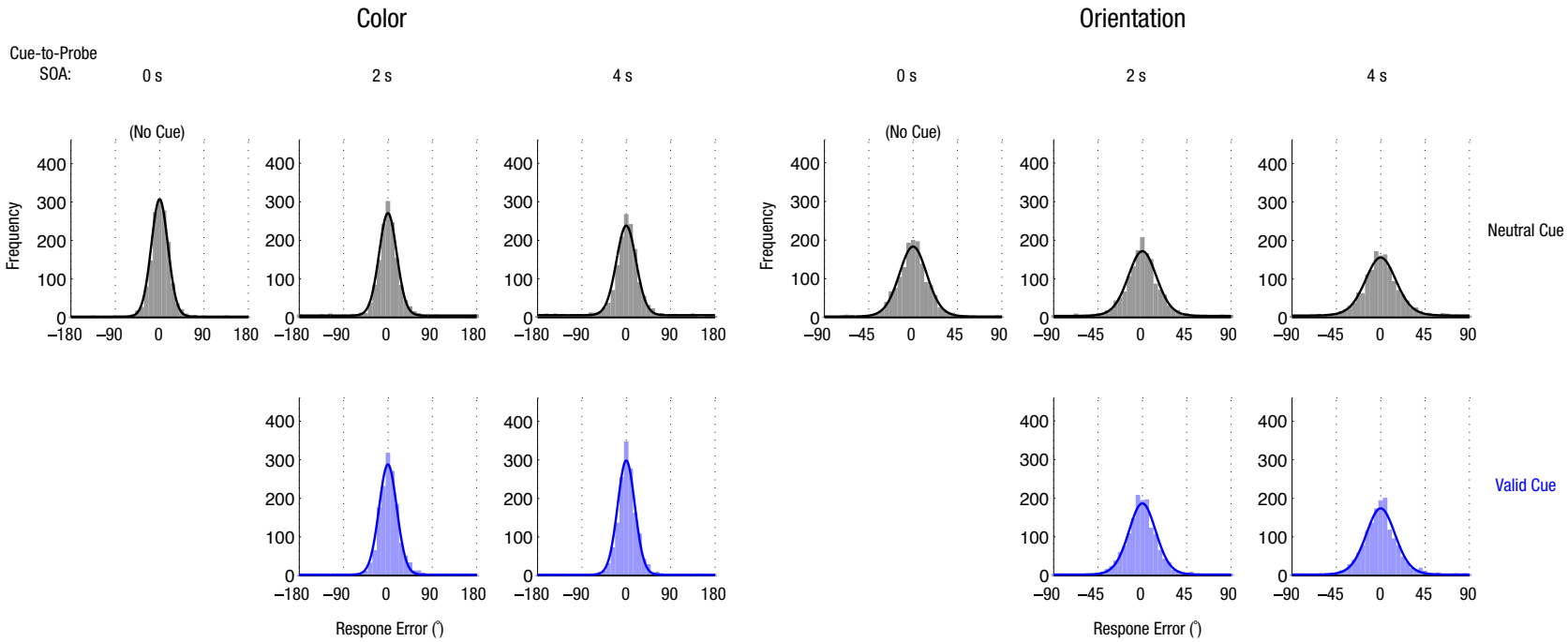

Figure S4. Histogram of response errors pooled across all participants (N = 16) in Experiment 2. Response error distribution is shown for each feature dimension, cue type, and cue-to-probe SOA combination. The solid line overlaid on each histogram shows the mixture distribution based on the average of individual participants' best-fitting parameters.

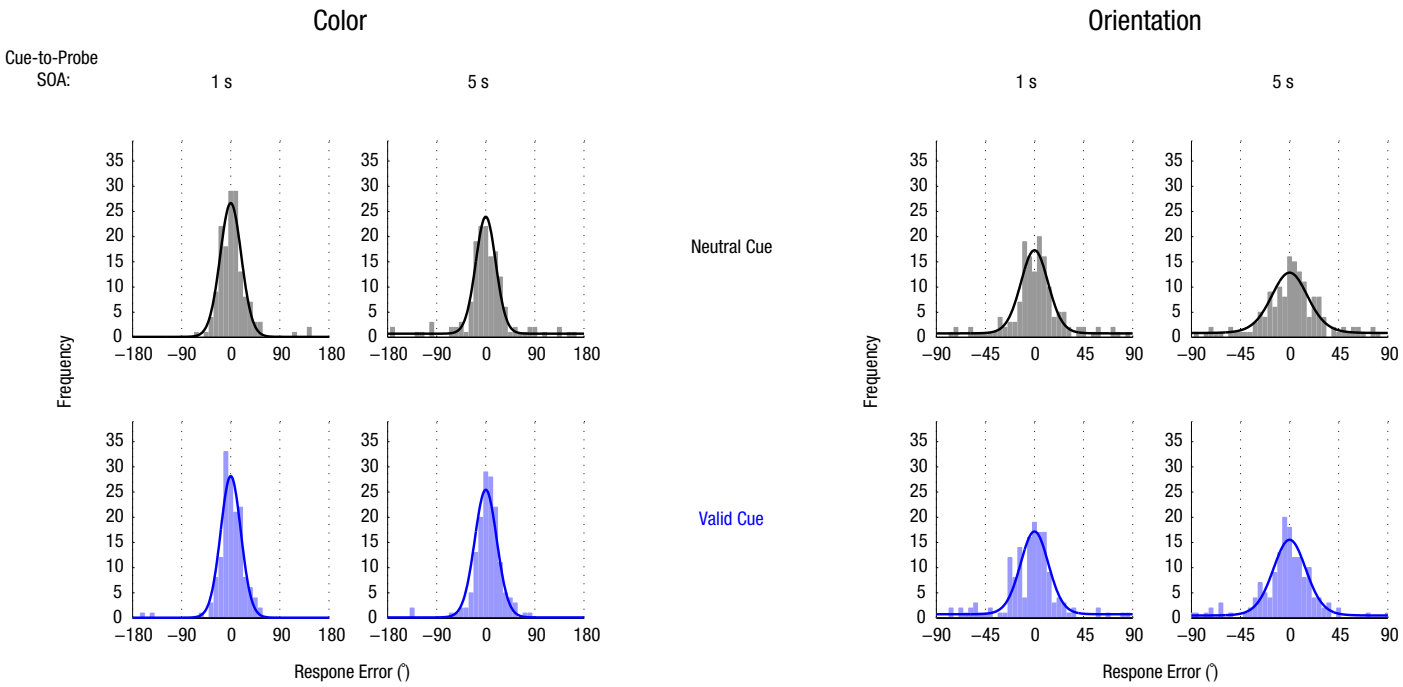

Figure S5. Histogram of response errors for a representative participant in Experiment 3. Response error distribution is shown for each feature dimension, cue type, and cue-to-probe SOA combination. The solid line overlaid on each histogram shows the best-fitting mixture distribution for the respective condition for this participant.

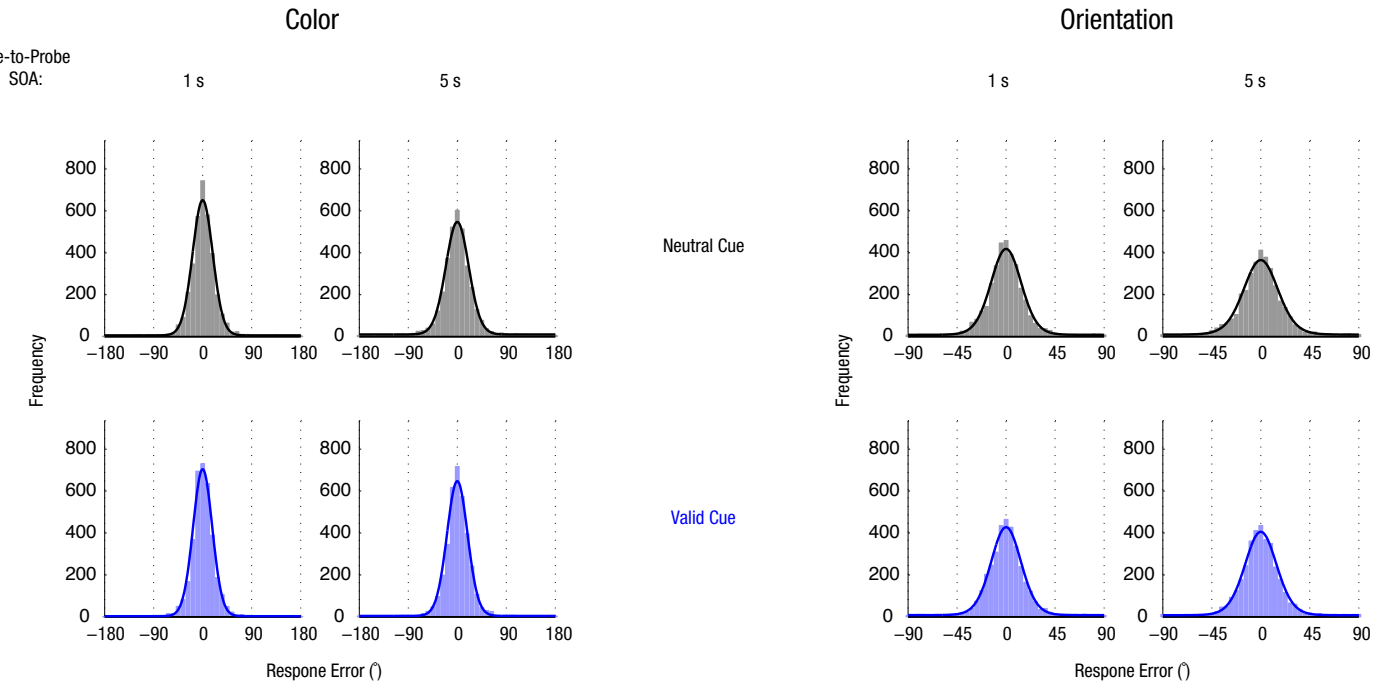

Figure S6. Histogram of response errors pooled across all participants (N = 24) in Experiment 3. Response error distribution is shown for each feature dimension, cue type, and cue-to-probe SOA combination. The solid line overlaid on each histogram shows the mixture distribution based on the average of individual participants' best-fitting parameters.

A priori power analysis for Experiment 3

The effect of interest

The key statistical test for Experiment 3 is the interaction between cue-to-probe SOA (short and long) and cue type (valid and neutral), on the SD and $P_{failure}$ estimates averaged across color and orientation. The F-test for the interaction effect in a 2 x 2 repeated measures design is essentially equivalent to a one-sample T-test comparing the mean interaction score to zero. The interaction scores can be calculated by applying a contrast $V_{short} – V_{long} – N_{short} + N_{long}$ (V = valid; N = neutral; short = short SOA; long = long SOA) to each participant's SD and $P_{failure}$ data. The mean and standard deviation of these scores will be used to calculate the standardized effect size (Cohen's *d*) for the interaction effect.

When we determined the sample sizes (*N*) for Experiments 1 and 2, however, the power calculation was based on detecting the main effect of cue type (valid and neutral) rather than the interaction effect. To assess our chance of obtaining significant results in the previous experiments if the observed effects were true, we conducted *post hoc* power analyses with regard to the interaction effect as defined above. In Experiment 1 (only comparing valid and neutral conditions), Cohen's *d* was 0.46 for SD, and 0.16 for $P_{failure}$. With *N* = 19 and a two-tailed alpha of .05, the achieved power for this experiment was 47% for SD, and 10% for $P_{failure}$. In Experiment 2 (*N* = 16), which had higher cue validity and a larger number of trials, Cohen's *d* was 0.32 for SD, and 0.38 for $P_{failure}$, achieving a power of 23% and 29%, respectively. These results indicate that our previous experiments were severely underpowered to detect the interaction effects in both parameters.

Prediction of the effect

Experiment 3 used a broader range of cue-to-probe SOAs (1 and 5 s) than those used in the previous experiments (2 and 4 s), in an attempt to obtain a larger interaction effect. We used the data from Experiment 2, which was more comparable than Experiment 1 in terms of the cue validity and the use of articulatory suppression, to predict the mean interaction effect with these new delay durations. Based on our previous work showing that changes in SD and $P_{failure}$ over delay durations of 1 to 12 s can be described reasonably well by a linear function over time (Rademaker, Park, Sack, & Tong, under review), we linearly extrapolated the group-mean SD and $P_{failure}$ estimates obtained at 2 and 4-s SOAs, separately for each cue-type and each feature. The original and the extrapolated data points are shown on top panels of Figure S7 and Figure S8, respectively (filled circles and squares). With the linear extrapolation, the unstandardized effect size (i.e., mean interaction score averaged

across features) increased twofold, from 0.0280 to 0.0560 (rad) for SD, and from 0.0278 to 0.0556 for $P_{failure}$.

Prediction of the variability

Calculation of the standardized effect size (Cohen's *d*) requires an estimation of the variability associated with the interaction effect, which stems from the variability of the effect in the population, as well as the variability in the parameter estimation process. We assumed that the total variance of the interaction effect that can be expected in a given experiment is the sum of the variances from the two sources:

$$\sigma^2_{total} = \sigma^2_{effect} + \sigma^2_{estimation}$$

σ^2_{effect} is the variance of the interaction scores across individuals in the population. $\sigma^2_{estimation}$ is an additional source of variance arising from the estimation process, which depends on the specific magnitudes of SD and $P_{failure}$ parameters comprising each individual's interaction scores and the amount of data available. If an infinite amount of data were available, the estimated parameter values would be error-free ($\sigma^2_{estimation} = 0$), and the total variance would directly reflect the variance of the effect in the population ($\sigma^2_{total} = \sigma^2_{effect}$). In reality, the population variance needs to be isolated by subtracting the estimation variance from the total variance ($\sigma^2_{effect} = \sigma^2_{total} - \sigma^2_{estimation}$).

The $\sigma^2_{estimation}$ for Experiment 2 can be reliably estimated by simulation, based on the observed range of parameter values and the number of trials per condition. For simplicity, we used the group-average data, rather than individual participants' parameter estimates. We took the group-average SD and $P_{failure}$ values for each combination of SOA, cue type, and feature. 96 trials were randomly generated for each of the 8 conditions using the appropriate pair of parameters, which were then recovered by fitting the mixture model. This process was repeated 5000 times to assess the accuracy and precision of parameter estimation.

The mean recovered parameters for each of the four conditions (V_{short}, V_{long}, N_{short}, and N_{long}) are shown on the top row of Figure S7, separately for color (left), orientation (middle), and their average (right). The recovered parameters accurately mirrored the true parameters used to generate the data (filled circles and squares). The level of noise varied across conditions, as indicated by the error bars (±1 standard deviation or sd). The sd values are replotted on the second row of Figure S7. The mean interaction scores ($V_{short} - V_{long} - N_{short} + N_{long}$) calculated from these data are shown on the third row of Figure S7 (a black square), with the error bars indicating ±1 sd. The interaction effects are highly variable, as the independent noise from the four conditions would add up. The simulation results indicated that a substantial portion of the total variance (sd^2_{total})

observed in Experiment 2 (i.e., sd^2 of the interaction scores across subjects) can be attributed to the variance from the estimation process ($sd^2_{estimation}$): 66.3% for SD, and 65.4% for $P_{failure}$ (after averaging across two features). This is illustrated on the bottom row of Figure S7, which plots the sd of the interaction scores from the simulation (gray), along with that observed in Experiment 2 (red).

We estimated the variability of the interaction effect in the population (σ^2_{effect}) by subtracting the variance of simulated interaction scores from the total variance observed in Experiment 2:

$$sd^2_{effect} = sd^2_{total} - sd^2_{estimation}$$

We assumed that this σ^2_{effect} estimate from Experiment 2 can approximate the σ^2_{effect} for Experiment 3. To estimate the variability of parameter estimation ($\sigma^2_{estimation}$) associated with the new experimental design in Experiment 3, we ran the same simulation using the extrapolated SD and $P_{failure}$ values (see Figure S8, top panels) and 150 trials per condition. This simulation was also run using a range of different numbers of trials per condition (n = 96-168) to examine the impact of the amount of data on parameter estimation. As shown in Figure S9, the variability of parameter estimation for each condition systematically decreases as the n increases (top two rows), leading to a more reliable estimation of the interaction effect (bottom two rows).

Now, we can calculate total variance expected for Experiment 3 by summing the σ^2_{effect} estimated from Experiment 2 and the $\sigma^2_{estimation}$ obtained from simulation. The predicted total variability (sd) of the interaction effect averaged across features is shown on the bottom row of Figure S9 (red line) for each parameter. The simulation results indicate that as the number of trials, n, increases from 96 to 150, the predicted variability (sd) is reduced by 13.2% for SD, and 12.9% for $P_{failure}$. The results from n = 150 are re-plotted in Figure S8, for comparison with the Experiment 2 results.

Power analysis

The prediction of mean interaction effects for SD and $P_{failure}$ is shown on the third row of Figure S8, along with the error bar indicating the variability (±1 sd) from the estimation alone (black) or the predicted total variability (red). Using the predicted mean and sd of the interaction effect averaged across two features, we calculated the standardized effect size (Cohen's d) for each parameter. The predicted Cohen's d was 0.63 for SD, and 0.74 for $P_{failure}$, which increased almost twofold compared to those actually observed in Experiment 2 (0.32 for SD, and 0.38 for $P_{failure}$; see Figure S7, third row).

We conducted an *a priori* statistical power analysis using G*Power 3.1 software (Faul, Erdfelder, Buchner, & Lang, 2009). We used a

one-sample T-test comparing the mean interaction effect to zero, with a two-tailed alpha of .05. The power analysis revealed that 22, 25, and 29 subjects would be required to achieve 80%, 85%, and 90% power for SD, and 17, 19, and 22 participants would be required to achieve the corresponding power for $P_{failure}$. We decided that a sample size of 24 would provide adequate power to detect the interaction effects in both parameters.

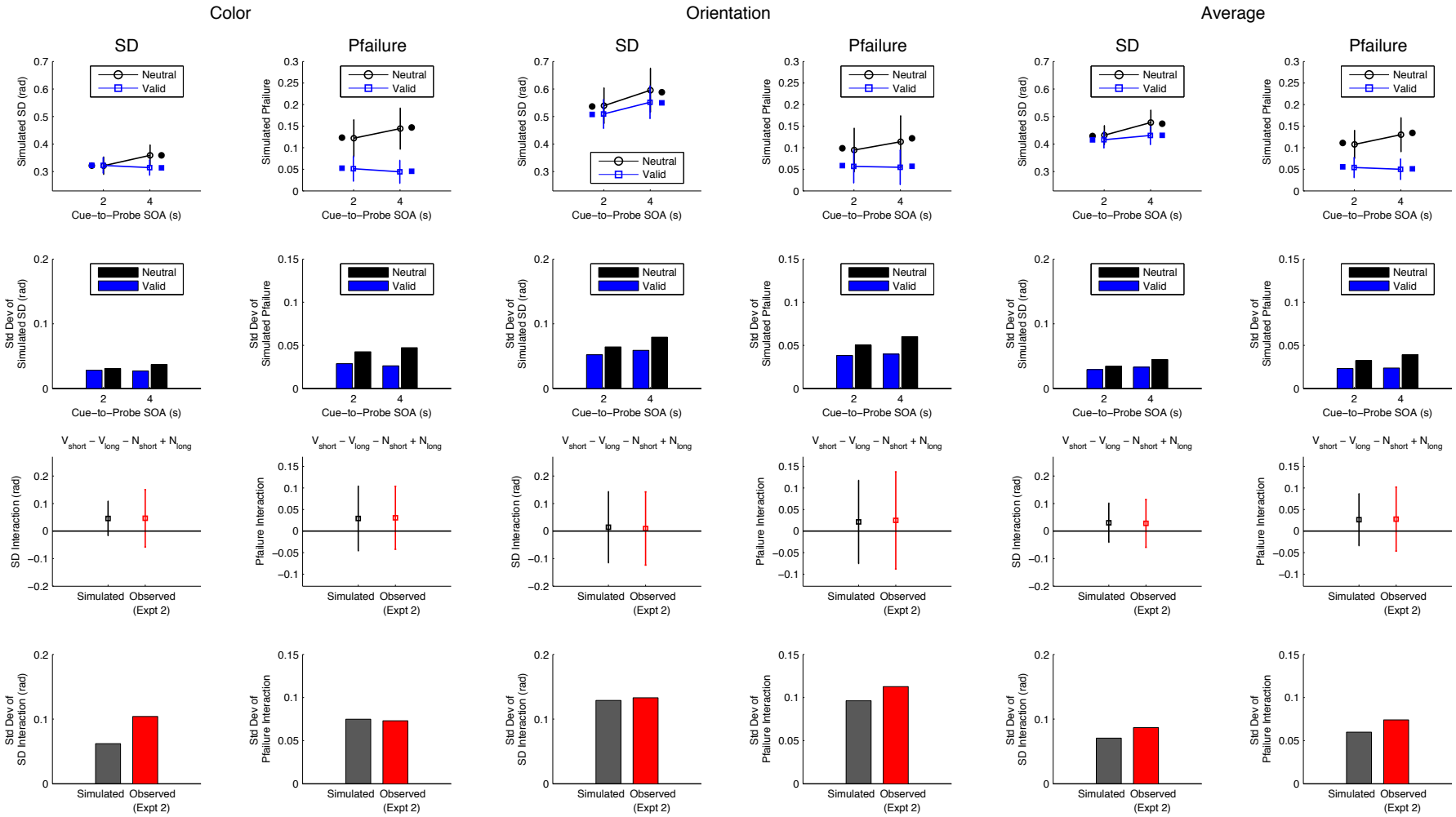

Figure S7. Simulation of the variability in parameter estimation in Experiment 2.
On the top row, the mean recovered SD and $P_{failure}$ parameters for each of the four conditions (valid and neutral cues, at 2- and 4-s SOAs) are shown, separately for color (left), orientation (middle), and their average (right). The filled circles and squares indicate the true parameters that were used to generate the data (i.e. group-averaged parameter values from Experiment 2). The simulation was run with 96 trials per condition, and repeated 5000 times. The error bars represent ±1 sd of the estimated parameters across simulations. The sd values for the four conditions are re-plotted on the second row. On the third row, the mean of the interaction scores ($V_{short} – V_{long} – N_{short} + N_{long}$; V = valid, N = neutral, short = short SOA, long = long SOA) computed from the simulated data is shown in black, with the error bar representing ±1 sd of the interaction scores across simulations. For comparison, the mean interaction score actually observed in Experiment 2 is shown in red, along with the ±1 sd of the interaction scores across 16 participants. The sd values of simulated and observed interaction effects are re-plotted on the bottom row.

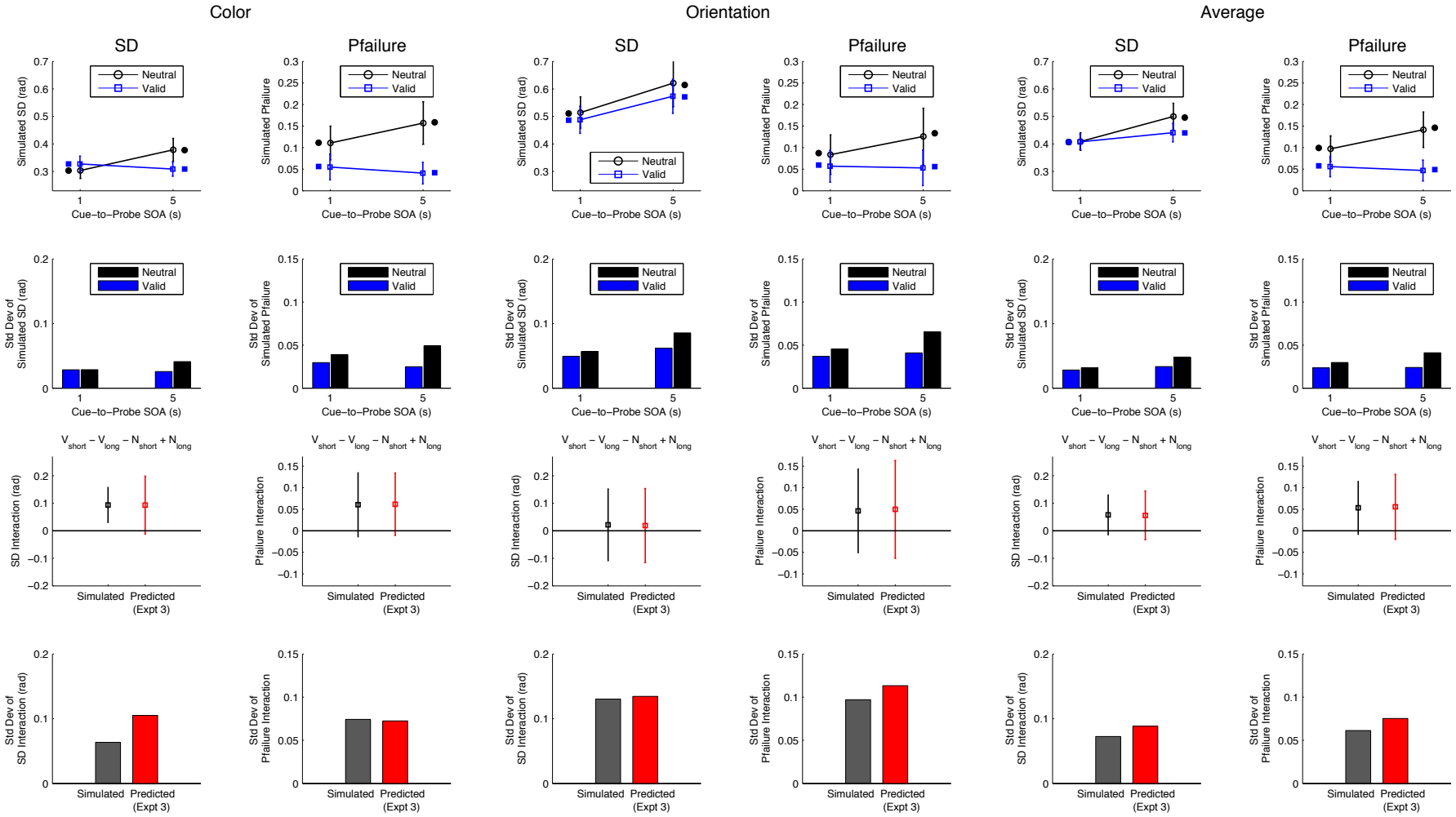

Figure S8. Simulation of the variability in parameter estimation in Experiment 3.
On the top row, the mean recovered SD and $P_{failure}$ parameters for each of the four conditions (valid and neutral cues, at 1- and 5-s SOAs) are shown, separately for color (left), orientation (middle), and their average (right). The filled circles and squares indicate the true parameters that were used to generate the data (i.e. the linear extrapolation of the group-averaged parameter values from Experiment 2). The simulation was run with 150 trials per condition, and repeated 5000 times. The error bars represent ±1 sd of the estimated parameters across simulations. The sd values for the four conditions are re-plotted on the second row. On the third row, the mean of the interaction scores ($V_{short} – V_{long} – N_{short} + N_{long}$) computed from the simulated data is shown in black, with the error bar representing ±1 sd of the interaction scores across simulations. The predicted mean interaction score for Experiment 3, calculated from the extrapolated group-averaged parameters, is shown in red, with the error bar representing the predicted variability (±1 sd) of the interaction scores, estimated by combining the estimation errors and the variability across subjects. The sd values of simulated and predicted interaction effects are re-plotted on the bottom row.

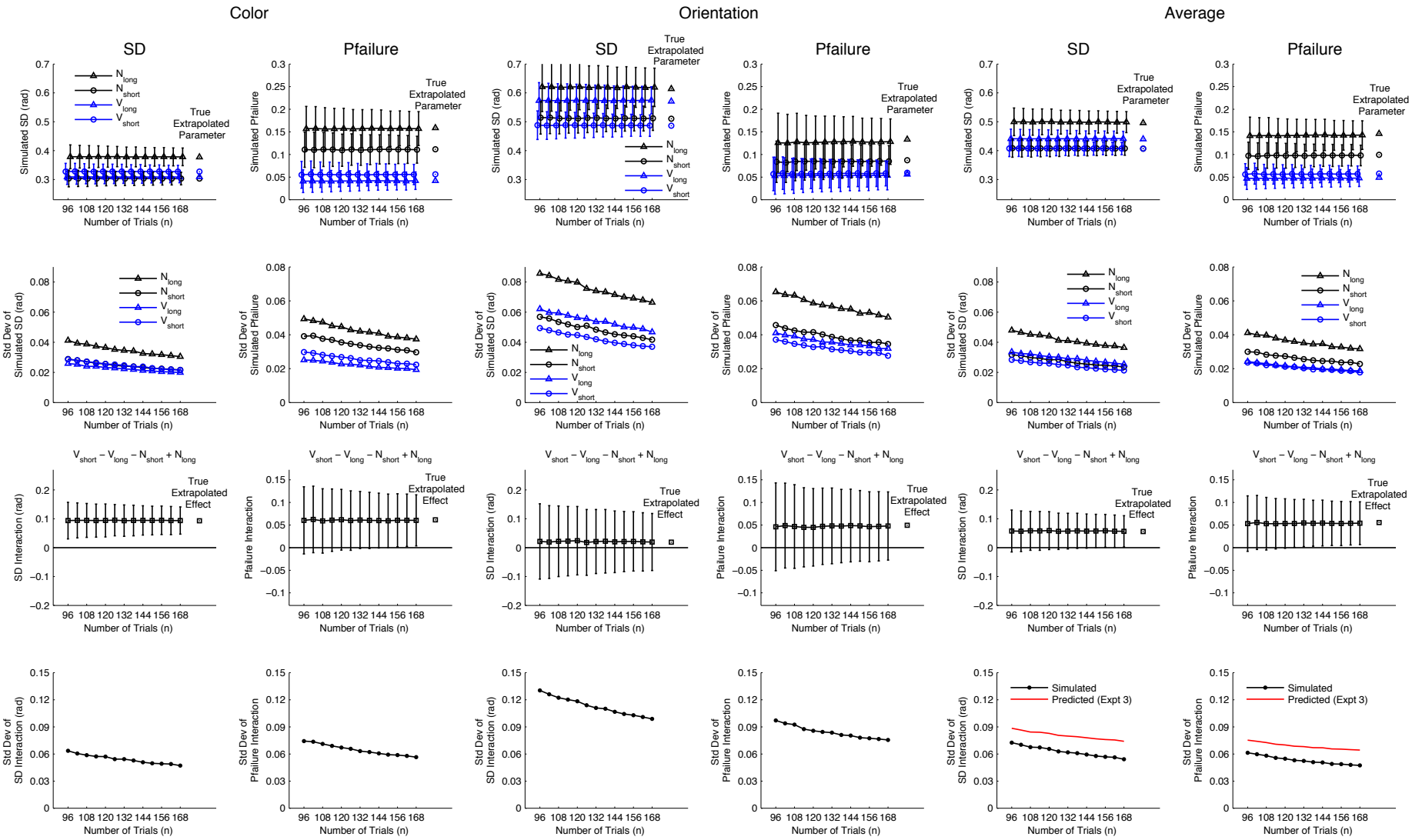

Figure S9. The effect of the number of trials on parameter estimation in Experiment 3.
Simulation data were generated using the linear extrapolation of the group-averaged parameter values from Experiment 2. The number of trials per condition (*n*) varied from 96 to 168, and the simulation was run 5000 times for each *n*.

References.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*, 1149-1160.

Rademaker, R. L., Park, Y. E., Sack, A. T., Tong, F. (Under review). Evidence of gradual loss of precision for simple features and complex objects in visual working memory. Manuscript submitted for publication.