Supplementary Material

for

Judgments During Perceptual Comparisons Predict Distinct Forms of Memory Updating

Joseph M. Saito[1], Gi-Yeul Bae[2], & Keisuke Fukuda[1,3]

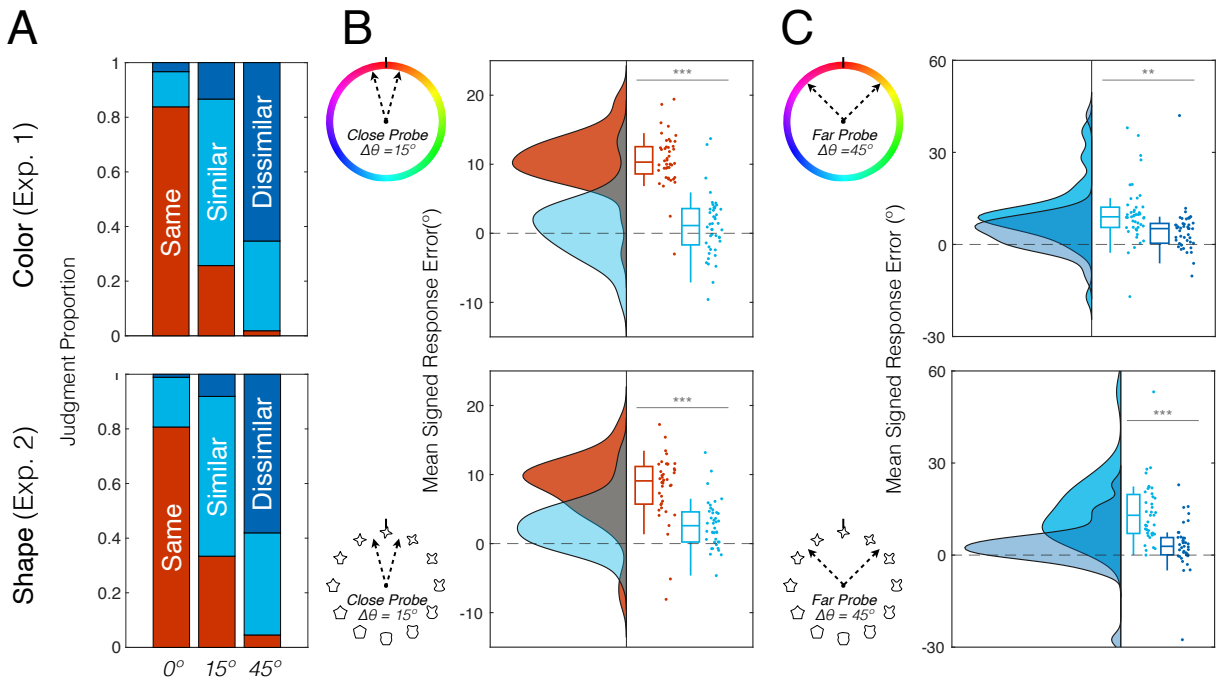
[1]University of Toronto, [2]Arizona State University, [3]University of Toronto Mississauga

Contents:

1 Behavioral results with exclusions included

Figure S1

Systematic VWM Performance Persists with Exclusions Included



Note. (**A**) Stacked bar chart indicating the proportion of 'same', 'similar', and 'dissimilar' judgments at each physical distance in the Compare condition. (**B**) Boxplots with corresponding density distributions depicting the mean signed response error following 'same' and 'similar' judgments in the 15° probe condition and (**C**) 'similar' and 'dissimilar' responses in the 45° probe condition. Colored dots to the right of each boxplot indicate the mean error for a given participant. ** $p < 0.01$, *** $p < 0.001$

We re-conducted our analyses while including the individuals that reported a low number of confident trials or had poor overall task performance. All effects persisted.

As can be seen in **Figure S1B**, we found evidence of systematic memory distortions in the 15° probe condition following 'same' judgments for both stimulus types (color: $M = 10.65°$, 95% CI [9.50°, 11.81°], $t(43) = 18.62$, $p < 0.001$, *Cohen's d* = 2.81, $BF_{10} = 1.19 \times 10^{19}$; shape: $M = 8.13°$, 95% CI [6.60°, 9.66°], $t(39) = 10.73$, $p < 0.001$, *Cohen's d* = 1.70, $BF_{10} = 2.39 \times 10^{10}$) and following 'similar' judgments for shape stimuli (color: $M = 1.21°$, 95% CI [-0.14°, 2.56°], $t(45) = 1.80$, $p = 0.078$, *Cohen's d* = 0.27, $BF_{01} = 1.41$; shape: $M = 2.76°$, 95% CI [1.67°, 3.86°], $t(39) = 5.09$, $p < 0.001$, *Cohen's d* = 0.81, $BF_{10} = 2.13 \times 10^{3}$). These distortions were reliably larger following 'same' judgments than 'similar' judgments (color: $M = 9.58°$, 95% CI [7.98°, 11.19°], $t(43) = 12.03$, $p < 0.001$, *Cohen's d* = 1.81, $BF_{10} = 2.78 \times 10^{12}$; shape: $M = 5.37°$, 95% CI [3.45°, 7.29°], $t(39) = 5.66$, $p < 0.001$, *Cohen's d* = 0.90, $BF_{10} = 1.15 \times 10^{4}$). We also found reliable memory biases following 'similar' judgments (**Figure S1C**; color: $M = 9.96°$, 95% CI [7.14°, 12.78°], $t(44) = 7.12$, $p < 0.001$, *Cohen's d* = 1.06, $BF_{10} = 1.63 \times 10^{6}$; shape: $M = 13.86°$, 95% CI

[10.62°, 17.10°], $t(38) = 8.66$, $p < 0.001$, *Cohen's d* = 1.39, $BF_{10} = 6.62 \times 10^7$) and 'dissimilar' judgments in the 45° probe condition (**Figure S1C**; color: $M = 4.39°$, 95% CI [2.24°, 6.55°], $t(45) = 4.11$, $p < 0.001$, *Cohen's d* = 0.61, $BF_{10} = 1.46 \times 10^2$; shape: $M = 3.26°$, 95% CI [0.84°, 5.68°], $t(39) = 2.72$, $p = 0.010$, *Cohen's d* = 0.43, $BF_{10} = 4.19$). Biases following 'similar' judgments were larger than those following 'dissimilar' judgments in both stimulus types (color: $M = 5.46°$, 95% CI [1.52°, 9.41°], $t(44) = 2.79$, $p = 0.008$, *Cohen's d* = 0.42, $BF_{10} = 4.88$; shape: $M = 10.62°$, 95% CI [7.47°, 13.77°], $t(38) = 6.83$, $p < 0.001$, *Cohen's d* = 1.09, $BF_{10} = 3.30 \times 10^5$). Thus, all of the behavioral findings reported in our main analyses persist even when we include the noisy estimates of memory distortion collected from participants that were infrequently confident or showed poor overall task proficiency.

2 Behavioral results in high-confidence trials

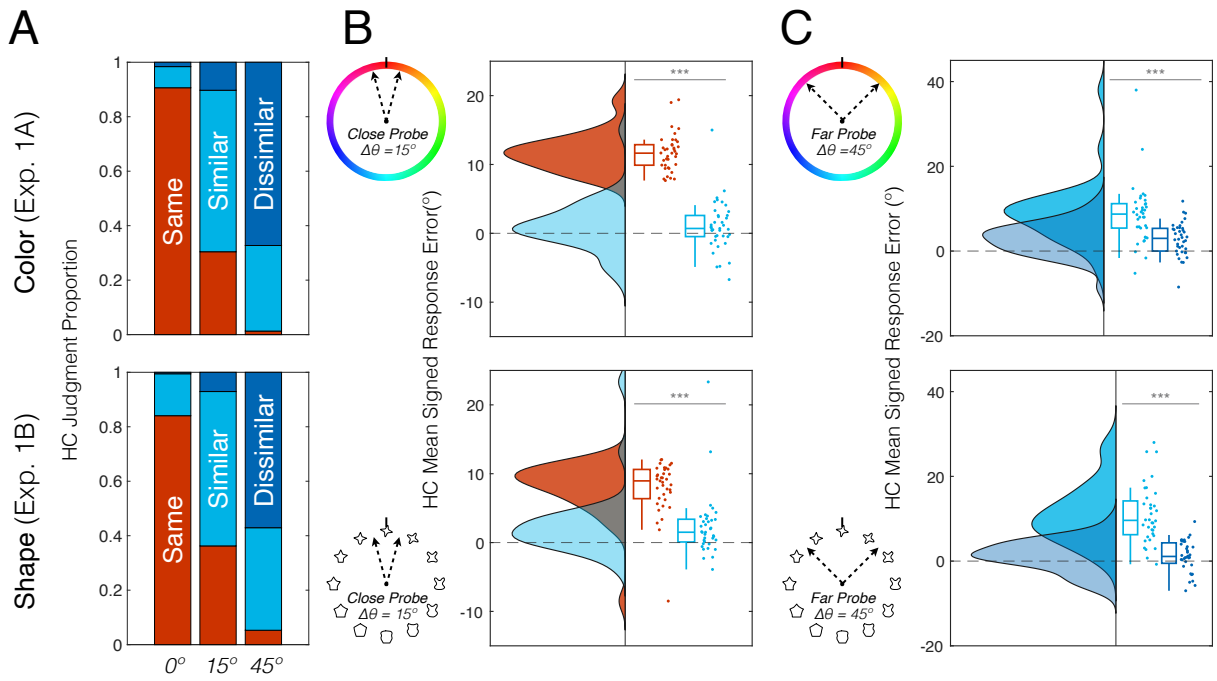
Table S1

Number and Proportion of Confident Memory Reports

	Variable	Short Baseline	Long Baseline	0°-Offset	15°-Offset	45°-Offset
Exp 1A (Color)	*Trial Prop (SD)* *Trial Count (SD)*	0.78 (0.19) 46.70 (11.49)	0.72 (0.22) 43.23 (13.00)	0.84 (0.14) 49.73 (8.49)	0.76 (0.21) 44.73 (12.37)	0.73 (0.23) 42.68 (13.41)
Exp 1B (Shape)	*Trial Prop (SD)* *Trial Count (SD)*	0.79 (0.20) 47.11 (12.06)	0.73 (0.23) 43.81 (13.52)	0.84 (0.16) 49.24 (9.34)	0.79 (0.18) 46.14 (10.79)	0.73 (0.21) 42.27 (12.92)

Figure S2

Systematic VWM Performance Persists in Confident Trials



Note. (**A**) Stacked bar chart indicating the proportion of 'same', 'similar', and 'dissimilar' judgments at each physical distance in the Compare condition. (**B**) Boxplots with corresponding density distributions depicting the mean signed response error following 'same' and 'similar' judgments in the 15° probe condition and (**C**) 'similar' and 'dissimilar' responses in the 45° probe condition. Colored dots to the right of each boxplot indicate the mean error for a given participant. ** $p < 0.01$, *** $p < 0.001$

We re-conducted our analyses while only including trials with memory reports that were made with high confidence. The number and proportion of high-confidence trials in each experiment are reported in **Table S1**.

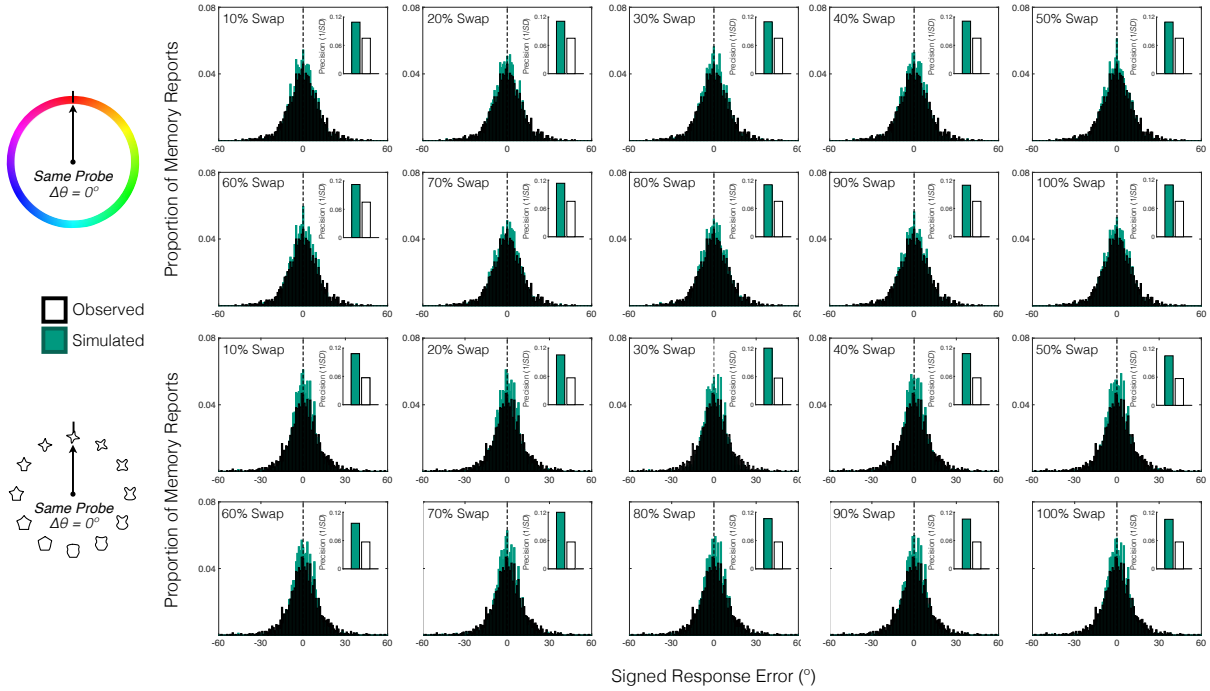
As can be seen in **Figure S2B**, we found evidence of systematic memory distortions in the 15° probe condition following 'same' judgments for both stimulus types (color: $M = 11.69°$, 95% CI

[10.82°, 12.56°], $t(38) = 27.21$, $p < 0.001$, *Cohen's d* = 4.36, $BF_{10} = 1.78 \times 10^{23}$; shape: M = 7.94°, 95% CI [6.67°, 9.21°], $t(36) = 12.64$, $p < 0.001$, *Cohen's d* = 2.08, $BF_{10} = 7.94 \times 10^{11}$) and following 'similar' judgments for shape stimuli (color: M = 0.98°, 95% CI [-0.19°, 2.15°], $t(39) = 1.70$, $p = 0.097$, *Cohen's d* = 0.27, $BF_{01} = 1.57$; shape: M = 2.32°, 95% CI [0.80°, 3.84°], $t(36) = 3.10$, $p = 0.004$, *Cohen's d* = 0.51, $BF_{10} = 9.67$). These distortions were reliably larger following 'same' judgments than 'similar' judgments (color: M = 10.67°, 95% CI [9.18°, 12.17°], $t(38) = 14.45$, $p < 0.001$, *Cohen's d* = 2.31, $BF_{10} = 1.12 \times 10^{14}$; shape: M = 5.62°, 95% CI [3.49°, 7.75°], $t(36) = 5.35$, $p < 0.001$, *Cohen's d* = 0.88, $BF_{10} = 3.72 \times 10^{3}$). We also found reliable memory biases following 'similar' judgments (**Figure S2C**; color: M = 8.67°, 95% CI [6.32°, 11.02°], $t(37) = 7.47$, $p < 0.001$, *Cohen's d* = 1.21, $BF_{10} = 1.86 \times 10^{6}$; shape: M = 10.97°, 95% CI [8.64°, 13.30°], $t(35) = 9.54$, $p < 0.001$, *Cohen's d* = 1.59, $BF_{10} = 3.35 \times 10^{8}$) and 'dissimilar' judgments in the 45° probe condition (color: M = 2.78°, 95% CI [1.55°, 4.01°], $t(39) = 4.58$, $p < 0.001$, *Cohen's d* = 0.72, $BF_{10} = 4.85 \times 10^{2}$; shape: M = 1.30°, 95% CI [0.08°, 2.52°], $t(36) = 2.17$, $p = 0.037$, *Cohen's d* = 0.36, $BF_{10} = 1.41$). Biases following 'similar' judgments were larger than those following 'dissimilar' judgments in both stimulus types (color: M = 5.82°, 95% CI [3.18°, 8.46°], $t(37) = 4.47$, $p < 0.001$, *Cohen's d* = 0.73, $BF_{10} = 3.34 \times 10^{2}$; shape: M = 9.75°, 95% CI [7.17°, 12.33°], $t(35) = 7.66$, $p < 0.001$, *Cohen's d* = 1.28, $BF_{10} = 2.27 \times 10^{6}$). Therefore, we conclude that all behavioral patterns persist when we focus on trials with high-confidence memory reports.

3 Precision simulation across a full range of replacement rates

Figure S3

Improved memory precision, irrespective of replacement rate

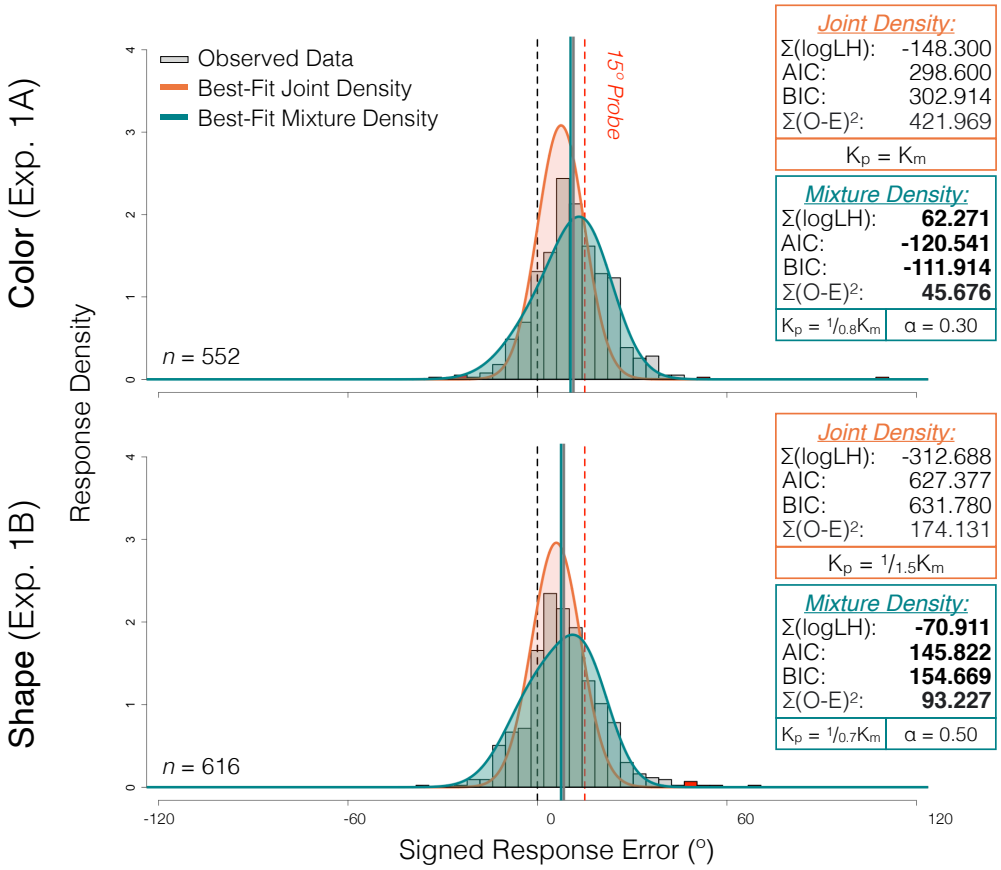


Note. Signed response distributions of simulated responses based on an encoding accuracy account and observed responses in the short baseline condition. In the simulated response distributions, the percentage of probe-based memory reports (i.e., replacement errors) varied from 10% to 100% in 10% increments. The inset bar charts show the precision of the simulated responses compared to the observed baseline responses. In both color and shape stimuli, increased response precision is observed regardless of how frequently or infrequently replacement was set to occur.

We reported in the main text that constraining the dataset to trials with 'same' judgments increases report precision and that this increase occurs even when replacement is assumed to occur on 50% of trials in our simulation. To ensure that this pattern persists regardless of how frequently replacement is assumed to occur, we reconducted our simulation multiple times while changing the replacement rate on each iteration. In **Figure S3**, we report the results of ten different simulations that were performed for each experiment where the replacement rate was set to vary from 10% to 100% in steps of 10%. In both experiments, all ten simulated response distributions showed higher response precision than the observed responses in the short baseline condition. This corroborates our conclusions in the main text that improved report precision occurs due to fluctuations in the quality of initial target encoding that are tracked by subjective judgments and that memory replacement is not mutually-exclusive to this encoding accuracy account.

Figure S4

Modeling of VWM Response Errors Following All Confident 'Same' Judgments

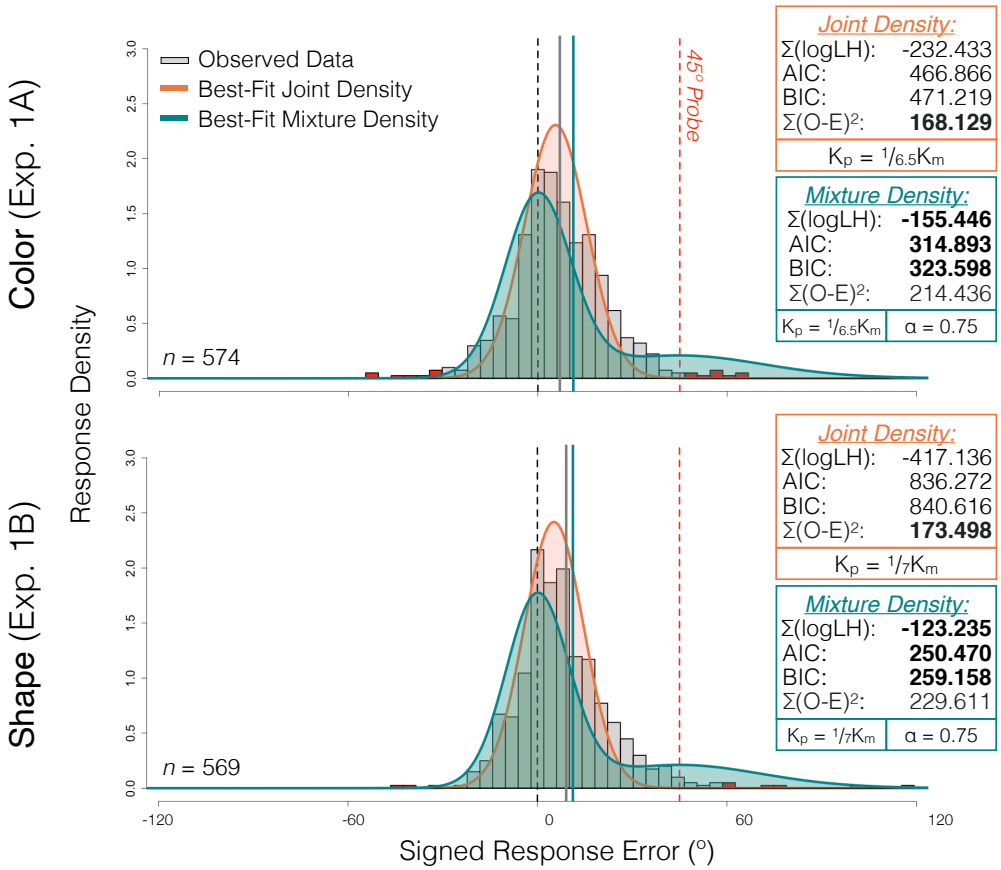


Note. Best-fitting Joint Density (JD) and Mixture Density (MD) distributions for the color and shape experiments overlaid on the Observed Data. For visualization, the Observed Data in each experiment are plotted as a histogram with 90 bins (0.07 radians/bin). The x-axes are abbreviated to -120 to 120 degrees surrounding the zero-centered target to emphasize the shape of the distributions. Red bars in the Observed Data histograms represent outlier response errors (>2.5SD above or below mean response error) that were identified prior to model fitting. Vertical black and red dashed lines indicate the location of the target and probe stimuli across trials, respectively. Vertical teal and gray solid lines indicate the mean response error in the Best-Fitting MD and Observed Data distributions, respectively. Formal model fit statistics and sum of squared differences between the Observed and Best-Fitting Model data are reported with bold face values indicating the preferred model. Free parameters identified within the best-fitting models are reported below the fit statistics.

Despite focusing on trials where participants reported being highly confident in their memory report, there were several trials with memory reports that were far from both the target and the probe. As can be seen in **Figures S4-5**, all of these reports fell outside the central density function of the observed data, suggesting that these trials likely represented rare instances where participants reported being highly confident despite guessing (e.g., due to a pre-potent confidence response or a response strategy). As such, we decided to remove these outlier trials in our main analysis since they can disproportionately skew formal model fit statistics (e.g., log-

Figure S5

Modeling of VWM Response Errors Following All Confident 'Similar' Judgments



Note. Best-fitting Joint Density (JD) and Mixture Density (MD) distributions for the color and shape experiments overlaid on the Observed Data. For visualization, the Observed Data in each experiment are plotted as a histogram with 90 bins (0.07 radians/bin). The x-axes are abbreviated to -120 to 120 degrees surrounding the zero-centered target to emphasize the shape of the distributions. Red bars in the Observed Data histograms represent outlier response errors (>2.5SD above or below mean response error) that were identified prior to model fitting. Vertical black and red dashed lines indicate the location of the target and probe stimuli across trials, respectively. Vertical teal and gray solid lines indicate the mean response error in the Best-Fitting MD and Observed Data distributions, respectively. Formal model fit statistics and sum of squared differences between the Observed and Best-Fitting Model data are reported with bold face values indicating the preferred model. Free parameters identified within the best-fitting models are reported below the fit statistics.

likelihood; Huber, 2004). Here, we report the results of our modeling analysis with these outliers included.

As anticipated, re-introducing outliers weakened the fit of all 8 models. This weakening was especially drastic for the JD model. Unlike the JD model, the MD model was sometimes preserved against severe reductions in log-likelihood by its ability to account for outliers that were beyond the probe on the positive side of the feature space. For example, in the 45° probe condition, the MD model predicted a long, widely-distributed positive tail that was intended to reflect bimodality that was not actually present in the observed distribution, with or without

outliers (see **Figure S5**). However, because some of the outliers fell underneath this tail, they did not significantly reduce log-likelihood in the MD model. Note that our only fit measure that did not incorporate log-likelihood (i.e., sum of squared differences) was only slightly changed in all 8 models by including the outliers.

Model comparisons that followed from these individual fit measures tell an identical story. In the 15° probe condition, where the JD and MD models both struggled to account for outliers, both model fits were worse, but model comparisons were unchanged. However, in the 45° probe condition, re-introducing outliers severely weakened fit measures in the JD model that were dependent on log-likelihood (i.e., sum of log likelihood, AIC, BIC), whereas these fit measures were less severely weakened in the MD model. As a result, measures dependent on log-likelihood now preferred the MD model over the JD model, while sum of squared differences continued to prefer the JD model over the MD model. These findings support our initial decision to remove the outliers and perform a fair model comparison that would not be disproportionately influenced by < 3% of trials.

We acknowledge that positive outliers captured by the MD model in the 45° probe condition may or may not have represented a handful of trials where memory replacement occurred. However, the purpose of our model comparison was to determine which of two plausible mechanisms provided a better explanation for the systematic trends observed across hundreds of trials. As depicted in **Figures S4-5**, re-introducing a handful of potential memory replacements to the data does not meaningfully change the shape of the resulting distributions and therefore cannot be treated as evidence of a reliable effect. The fact that the MD model was still preferred over the JD model in the 15° probe condition confirms that *reliable* patterns of memory replacement were not hampered by outlier exclusion. Thus, we conclude that removing outliers helped minimize undue contamination of our model comparisons without removing any meaningful systematicity in participants' responses.

5 Computational modeling with non-linear psychophysical scaling

In the main manuscript, Joint and Mixture Density models were constructed following the assumptions of conventional models of VWM which state that the psychophysical similarity between stimuli in a given feature space is linearly related to their physical similarity (e.g., Zhang & Luck, 2008). However, recent work by Schurgin et al. (2020) challenged this conventional modeling assumption by demonstrating that psychophysical similarity is actually *nonlinearly* related to physical similarity. The researchers used a perceptual task to show that observers are better at selecting which of two probe stimuli is more similar to a target stimulus when the probes are close to the target (e.g., offset by 20° and 40°, respectively) than when they are further away (e.g., offset by 120° and 140°, respectively). Using this empirical data, the researchers constructed a psychophysical scaling function that accounts for nonlinearity in the global similarity structure when modeling VWM performance. In doing so, the researchers found that conventional models of VWM were outperformed by a Target Confusability Competition (TCC) model that was formalized by incorporating psychophysical scaling into a classic model of signal detection. For a detailed discussion of the theoretical implications of these findings and the continued debate surrounding the viability of the TCC model, see Schurgin et al. (2020), Oberauer et al. (2023), and Tomić & Bays (2022).

While nonlinearity in psychophysical scaling does not necessarily invalidate the model comparisons that are reported in the main manuscript, it does question whether the results might change if the models were formalized to assume nonlinearity instead. To test this possibility, we re-conducted our model comparisons for Experiment 1 using the same psychophysical scaling function and TCC model that Schurgin et al. (2020) employed to address nonlinearity in color space. In doing so, we replicated the results of our original model comparison by showing that report errors following 'similar' judgments were better explained by representational integration between the target and probe, while report errors following 'same' responses were better explained by probabilistic replacement of the target by the probe. This replication suggests that, despite meaningful theoretical differences between the TCC model and conventional models of VWM, both frameworks converge regarding the memory updating processes that are induced by performing perceptual comparisons between memories and new visual inputs.

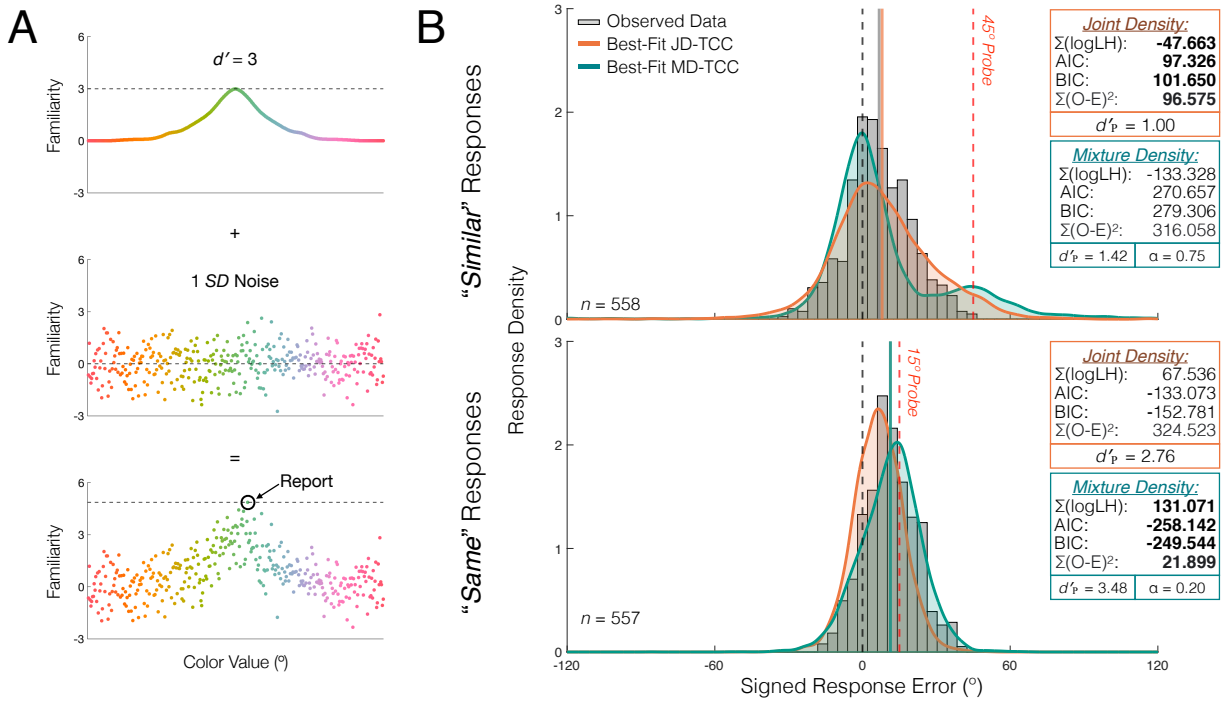
Method

TCC Model

Figure S6A shows the formalization of the TCC model. The model assumes that encoding a target stimulus into memory briefly enhances the familiarity of that target value in memory (d'_T). However, unlike the standard signal detection model, which assumes that only the familiarity of the target is enhanced (Macmillan & Creelman, 2005), the TCC model assumes that the familiarity signal generated by the target stimulus propagates to neighboring non-targets in the feature space as well. The spread in familiarity is assumed to follow a fixed psychophysical similarity function ($f(x)$) that is approximately exponential in shape. When prompted to report (r) the target from memory, the model assumes that the observer selects whichever stimulus value appears maximally familiar within the feature space at that time. In the case of a single target that is remembered, memory reports often cluster tightly around the true target value, but are

Figure S6

Computational Modeling of VWM Response Errors with TCC Model



Note. **(A)** TCC model formalization. When a target color is encoded into memory with a given memory strength (e.g., $d' = 3$), the familiarity of the target and its neighboring colors increases according to the psychophysical similarity function mapped to the feature space (*Top*). From here, one standard deviation of noise is added to each color value in the feature space to mimic the noisy processes tied to forming a memory representation (*Middle*). The resultant distribution formed by noise corruption represents the familiarity distribution that the observer uses to make their memory report (*Bottom*). In the case of continuous estimation, the observer reports the color that is maximally familiar within the distribution. **(B)** Best-fitting Joint Density-TCC (JD-TCC) and Mixture Density-TCC (MD-TCC) distributions for the color experiment overlaid on the Observed Data. For visualization, the Observed Data are plotted as histograms with 90 bins (0.07 radians/bin). The x-axes are abbreviated to -120 to 120 degrees surrounding the zero-centered target to emphasize the shape of the distributions. Vertical black and red dashed lines indicate the location of the target and probe stimuli in the feature space across trials, respectively. Vertical orange, turquoise, and gray solid lines indicate the mean response error in the Best-Fitting JD-TCC, Best-Fitting MD-TCC, and Observed Data distributions, respectively. Formal model fit statistics and sum of squared differences between the Observed and Best-Fitting Model data are reported with bold face values indicating the preferred model. Free parameters identified within the best-fitting models are reported below the fit statistics.

imperfect, nonetheless. This imperfection in reporting the exact target is assumed to arise from noise that corrupts the familiarity signal on every trial (N_{-179}, …, N_{180})[1], sometimes causing non-target values to appear more familiar than the target. From these assumptions, the TCC model is

[1] In the present procedure, we fit the model using uncorrelated noise since this is a more parsimonious formalization and is shown to produce fits that are nearly identical to TCC models that assume correlated noise instead (Schurgin et al., 2020).

able to closely recapitulate memory reports made by observers across a wide array of VWM tasks (Schurgin et al., 2020).

$$r = \max(d'_T * f(x) + (N_{-179}, \ldots, N_{180})) \quad (1)$$

Joint Density TCC (JD-TCC) Model

The JD-TCC model is nearly identical to the original JD model. The model assumes that participants' VWM representation of the target is integrated with the probe representation during perceptual comparisons to form a joint representation of the two items. Based on the signal detection principles posited by the TCC model, memory reports following integration are based on the maximally familiar stimulus value drawn from the joint distribution.

As before, the JD-TCC model is constructed in three steps:

First, we construct a representation (X_M) of the target stimulus (S_M) that follows a roughly exponential psychophysical similarity function ($f(x)$) that is centered at the location of the target stimulus in the feature space with a given memory strength (d'_T) and is contaminated by noise ($N_{-179T}, \ldots, N_{180T}$).

$$p(X_M|S_M) = d'_T * f(x) + (N_{-179T}, \ldots, N_{180T}) \quad (2)$$

d'_T is obtained by fitting the TCC model to memory reports in the delay-matched short baseline condition where no probe was presented. This allowed us to estimate the strength of the target at the time of the perceptual comparison using empirical data ($d'_T = 3.41$).

Second, we construct a representation (X_P) of the probe stimulus (S_P) that again follows a roughly exponential psychophysical similarity function ($f(x)$) that is centered at the location of the probe stimulus in the feature space with a given memory strength (d'_P) and is contaminated by noise ($N_{-179P}, \ldots, N_{180P}$).

$$p(X_P|S_P) = d'_P * f(x) + (N_{-179P}, \ldots, N_{180P}) \quad (3)$$

To estimate the strength of the probe representation, d'_P was set to vary in the parameter search process from 0 to 10 in increments of 0.02, yielding 501 possible memory strength estimates.

Third, we construct an integrated representation of the memory and probe that is assumed to follow a joint density of the memory and probe distributions (X_{JD}).

$$p(X_{JD}|\, S_M, S_P) = \frac{p(X_M|S_M)p(X_P|S_P)}{\sum p(X_M|S_M)p(X_P|S_P)} \quad (4)$$

This joint density distribution is constructed by a straightforward multiplication of the target (2) and probe (3) density functions. The two strength parameters in the joint density function are identical to those identified in (2) and (3). Thus, we estimated one free parameter in the JD-TCC model (i.e., d'_P).

Mixture Density TCC (MD-TCC) Model

The MD-TCC model is also nearly identical to the original MD model. The model assumes that participants' VWM representation of the target is sometimes replaced by the probe representation, such that individuals rely on the probe representation during the memory report. Thus, the model assumes that perceptual comparisons do not change the underlying memory representations, but instead change the likelihood that the probe representation will be used to represent the original target stimulus.

As before, the MD-TCC model is constructed in three steps. Steps 1-2 are identical to the JD-TCC model. The third step is as follows:

$$p(X_{Mix}|\ S_M, S_P) = \alpha p(X_M|S_M) + (1-\alpha)p(X_P|S_P) \quad (5)$$

α is used as a mixture parameter that estimates the proportion of trials where the memory report was based on the target representation (2). The remaining trials are assumed to be based on the probe representation (3) (i.e., 1-α). α was allowed to vary between 0 and 1 in increments of 0.05, yielding 21 possible mixture parameters. That is, the percentage of memory-based reports was allowed to vary from 0% (100% probe-based) to 100% (0% probe-based) in 5% increments. All other aspects of the MD-TCC model are identical to the JD-TCC model. Thus, we estimated two free parameters in the MD model (i.e., α and d'_P).

Procedure

The model fitting procedure was identical to the one reported in the main manuscript.

Results

First, we fit both models to errors observed following 'similar' judgments in the 45° probe condition to test whether representational integration provided a better explanation for these errors than representational replacement (**Figure S6B**). The MD-TCC model produced a best-fitting distribution that was nearly identical to the distribution produced by the original MD model. Because the MD-TCC model assumes a bimodal response pattern comprised of memory and probe-based reports, the best fitting MD-TCC model was incapable of re-producing the shifted central gaussian in the observed data and attempted to compensate for this by assuming that replacement occurred in only 25% of trials (i.e., alpha = 0.75). The MD-TCC model was also forced to assume that responses based on the probe representation were widely-distributed, producing a long, positive-going tail in the distribution that was not present in the observed data. The JD-TCC model, on the other hand, did a superior job at capturing the quality of the observed data, including the positive skewing in the probe-side tail, which the original JD model struggled to recapitulate. This was likely due to the inclusion of uncorrelated noise in the TCC model that allowed it to better approximate trial-wise variability in the magnitude of the bias despite being fit at the aggregate "super subject" level. Lastly, just like the original JD and MD models, the JD-TCC and MD-TCC models both assumed weaker probe representations due to fixing the strength of the target during model fitting.

Sum of squared differences measures corroborated this qualitative assessment (JD = 96.575, MD = 316.058), confirming that the shape of the best-fitting JD-TCC model more closely resembled the shape of the observed data than the best-fitting MD-TCC model. Sum of log-likelihood (JD = -47.663, MD = -133.328), AIC (JD = 97.326, MD = 270.657), and BIC (JD = 101.650, MD =

279.306) measurements all unanimously preferred the JD-TCC model over the MD-TCC model. Thus, even when adopting a different model of VWM that assumes nonlinearity in psychophysical scaling, we find that response errors following 'similar' judgments were better explained by representational integration than probabilistic replacement.

We then moved to complement these initial findings by providing divergent evidence favoring the MD-TCC model in memory reports following 'same' judgments in the 15° probe condition (**Figure S6B**). Much like the original MD model, the close proximity between the target and probe representations resulted in a predicted MD-TCC distribution that was negatively-skewed towards the probe rather than comprised of distinct bimodal peaks. Consistent with the original JD model, the JD-TCC model struggled to account for response errors beyond 15° since integration necessarily produces a representation whose center falls between the target and the probe (i.e., 0° < center < 15°). Both models replicated the estimation of a probe representation that was quite strong, even stronger than the target in the case of the MD-TCC model. Again, given the close physical proximity between the target and probe in the feature space, it is reasonable that both representations could be encoded with high strength, yet still considerably overlapping.

Consistent with our qualitative assessment of the distributions, sum of squared differences measures were lower in the MD-TCC model than the JD-TCC model (JD = 324.523, MD = 21.889). Formal model comparisons using sum of log-likelihood (JD = 67.536, MD = 131.071), AIC (JD = -133.073; MD = -258.142), and BIC (JD = -128.774; MD = -249.544) measurements all unanimously preferred the MD-TCC model over the JD-TCC model. Together, we find clear computational evidence for the recruitment of qualitatively distinct memory updating mechanisms following 'same' and 'similar' judgments even when assuming a non-linear relationship between physical and psychophysical similarity in the underlying feature space.

References

Cousineau, D. (2005). Confidence intervals in within-subject designs: A simpler solution to Loftus and Masson's method. *Tutorials in Quantitative Methods for Psychology*, *1*(1), 42-45. https://doi.org/10.20982/tqmp.01.1.p042

Fukuda, K., Pereira, A. E., Saito, J. M., Tang, T. Y., Tsubomi, H., & Bae, G. Y. (2022). Working memory content is distorted by its use in perceptual comparisons. *Psychological Science*, *33*(5), 816-829. https://doi.org/10.1177/09567976211055375

Huber, P. J. (2004). *Robust Statistics* (Vol. 523). John Wiley & Sons, Inc.

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide* (2nd ed.). Lawrence Erlbaum Associates Publishers.

Oberauer, K. (2023). Measurement models for visual working memory—A factorial model comparison. *Psychological Review*, *130*(3), 841–852. https://doi.org/10.1037/rev0000328

Saito, J. M., Kolisnyk, M., & Fukuda, K. (2022). Perceptual comparisons modulate memory biases induced by new visual inputs. *Psychonomic Bulletin & Review*, 1-12. https://doi.org/10.3758/s13423-022-02133-w

Schurgin, M. W., Wixted, J. T., & Brady, T. F. (2020). Psychophysical scaling reveals a unified theory of visual memory strength. *Nature Human Behaviour*, *4*(11), 1156-1172. https://doi.org/10.1038/s41562-020-00938-0

Tomić, I., & Bays, P. M. (2022). Perceptual similarity judgments do not predict the distribution of errors in working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. Advance online publication. https://doi.org/10.1037/xlm0001172

Zhang, W., & Luck, S. J. (2008). Discrete fixed-resolution representations in visual working memory. *Nature, 453*(7192), 233-235. https://doi.org/10.1038/nature06860