nature research

https://doi.org/10.1038/s41562-020-00938-0

Supplementary information

Psychophysical scaling reveals a unified theory of visual memory strength

In the format provided by the authors and unedited

Supplementary Information

Supplementary Discussion

Measuring psychophysical similarity

The psychophysical similarity function we measure naturally captures two key aspects of how stimuli are perceived: The relationship between the physical stimulus and the psychological representation of that stimulus is rarely linear (e.g., CIELab is a complex transform of light wavelengths), and the similarity between stimuli as a function of distance is additionally non-linear[17,19]. In spaces that are already scaled to be approximately psychophysically uniform (e.g., CIELab), then, only the approximately-exponential fall-off in similarity remains to be modeled; whereas in spaces that are not equalized in advance (e.g., face space), both factors will be measured together, and inhomogeneities may need to be taken into account when modeling memory (e.g., Extended Data Figure 5, fitting each color separately).

In the current manuscript, we present several examples of tasks that naturally capture both of these insights and can be translated to a psychophysical similarity function, including the triad task, the quad task, and a subjective Likert similarity judgment (see Methods). It is important to note that depending on the number of trials, a large number of data points (and many subjects) may be necessary in order to obtain reliable estimates of a given stimulus space in the triad and quad tasks (in the current methods we collected n = 100 participants and pooled across them completely to obtain reliable group estimates). A Likert similarity task may be sufficient to capture this function under some circumstances, like for color in the current study. In such tasks, participants are simply asked to rate the similarity of two items (varying in distance from one another) on a Likert scale from 1 to 7, and these ratings can then be normalized. In color space, we observed this similarity rating task provided a measure of psychophysical similarity that is in close agreement with the results of the quad and triad tasks and requires considerably less data to estimate (Figure 1).

However, it is important to note that depending on the stimulus space, observers may utilize different strategies in such subjective similarity tasks, and that ultimately objective tasks like the quad task may be best to understand the psychophysical similarity function. In particular, to ensure the similarity function is properly measured, is important to ensure that participants provide judgments of the absolute interval between stimuli and not rely on categories or verbal labels, or, in the triad task, that participants not rely on a relational or relative encoding of the two choice items rather than their absolute distance to the target item (that is, the modeling assumes they compare each choice entirely separately to the target item — not relying on comparing the two choices, say, considering which choice is more clockwise in an orientation task). How best to ensure that participants rely on absolute intervals is represented in a large literature dating to Thurstone[62] and Torgerson[15].

Multidimensional stimuli, like color or faces, seem to have general agreement across many methods of measuring psychophysical similarity. However, we expect that collecting the psychophysical similarity measurements will be particularly challenging in single-dimensional stimulus spaces whose true objective distance function is transparent to participants. For example, when asking to judge orientation similarity or location similarity along a circle, participants are likely to be aware that the stimuli are physically manipulated on only a single dimension (angle), and will thus be inclined to report linear similarity judgments along this dimension. Less transparent similarity tasks, like the quad task, may help with this, but it may ultimately be difficult to prevent participants from using this knowledge. How best to deal with this remains a question for future work. For example, it may be possible to instead “back out”

the similarity function from memory data, or from alternative tasks (like speeded same-different tasks), or to use speeded similarity tasks to reduce such cognitive strategies. Alternatively, performing multidimensional scaling on the stimuli to create a psychophysically uniform space (as in CIELab for color; for example, in orientation this would "stretch" the space near the cardinals and shrink it near the obliques), could allow relatively simple similarity models. After such scaling, it would be likely that the similarity function beyond the perceptual discrimination limit would be an exponential function, which could allow the parameterization of the similarity function in relatively straightforward terms without the need for complex measurements.

It is important to note that while we emphasize the stability of the similarity function across conditions in the current work, the psychophysical similarity we measure could not possibly be a fixed property of the colors per se, but must be at least partially contextual. For example, if the background color of the display was blue rather than light gray, this would certainly alter the perception of — and discriminability of — colors from each other, as would adaptation and many other factors[49,50], which would necessarily have consequences for memory.

In addition, extremely brief presentations or extremely long presentations that allow verbal coding would be expected to alter this similarity function. It is expected this would result in changes in memory performance as well, in the same way that observed memory biases are altered when discriminability is affected by adaptation or contextual effects[48]. Thus, while we find the similarity function is fixed across a wide range of encoding times, delays and set sizes, there are likely to be conditions which change the underlying perception of the memoranda (e.g., very very short encoding times; different backgrounds) which will necessarily have an effect on memory.

"Dissociating" guess rate and precision

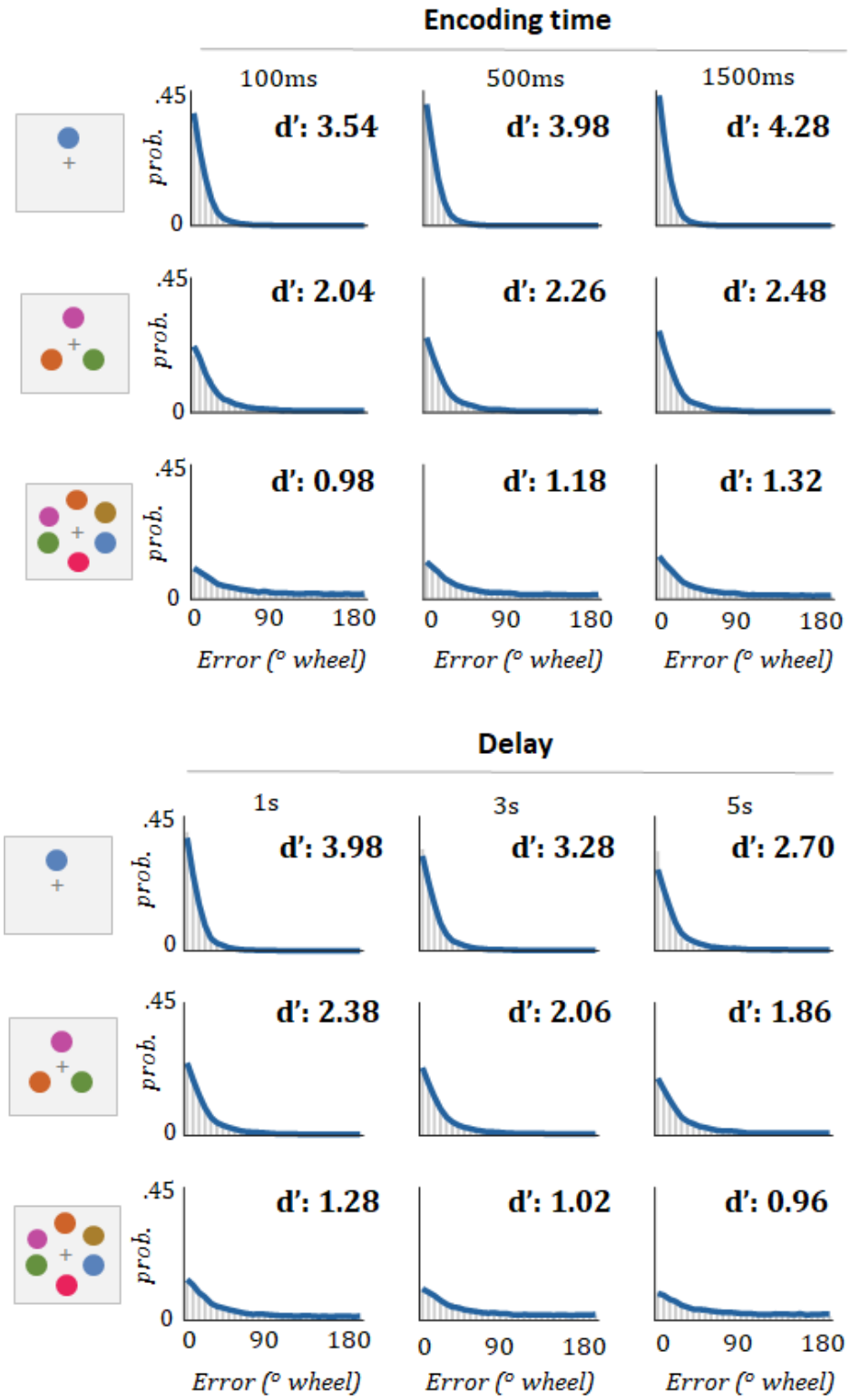
In addition to fitting a two parameter model, some previous research has claimed to dissociate these parameters. If a one-parameter model can account for the data, how has previous research so often found dissociations between these parameters?

The majority of these dissociations find that precision (SD) does not change when the 'guess rate' (or capacity) does change[7,31]. However, this dissociation is naturally explained by TCC because at low d' values, 'guess rate' can change by a huge amount with SD changing by only a few degrees. For example, over a wide range of guess rates, precision may only vary between SD=21 and SD=24, a difference that is visually indistinguishable and would require extremely high power to detect (e.g., Supplementary Figure 4). As an example, sampling 20 subjects of 100 trials each of data from the TCC at d'=1.0 vs. d'=0.7 and fitting these data with the 2-parameter mixture model reveals that such an experiment would find $p<0.05$ for 'capacity' greater than 60% of the time but $p<0.05$ for 'precision' approximately 11% of the time, despite both parameters being necessarily linked in the data from TCC. In line with this interpretation, many researchers have now found that with high enough power, previous studies claiming only a change in 'guess rate' but not 'SD' actually find changes in both, with very small changes in SD present along with large changes in 'guess rate'[64]. Other dissociations have sometimes been found — for example, Zhang and Luck[7] report a manipulation that causes a change in SD but not 'guess rate' — but these dissociations inevitably rely on comparisons across different sets of stimuli with different psychophysical similarity functions (e.g., the Zhang and Luck manipulation adds color noise to the items, making them less distinct), which is perfectly consistent with TCC.

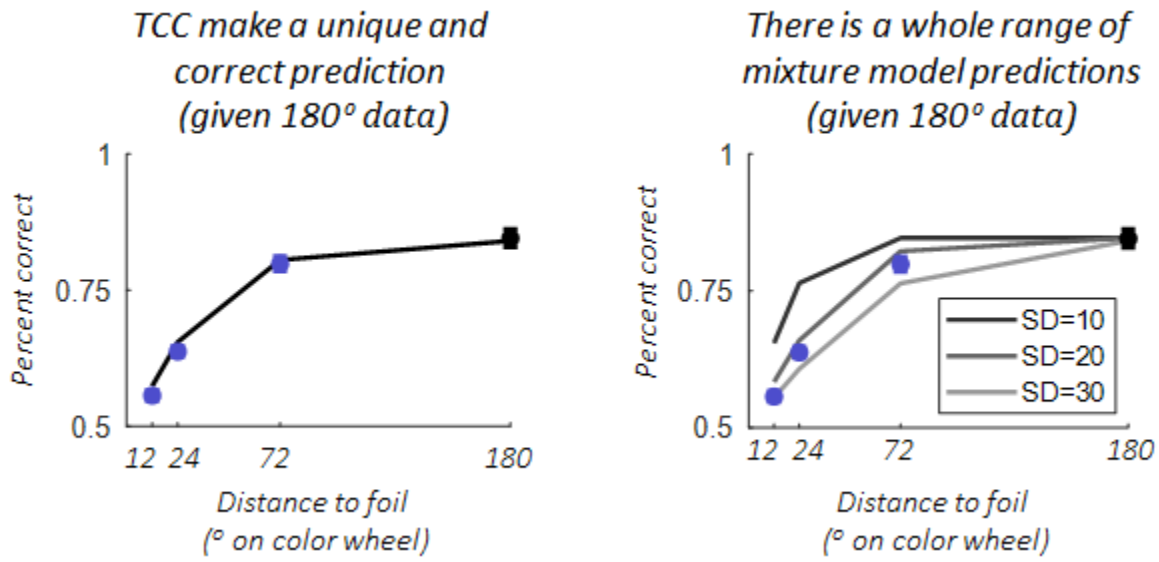
Supplemental References

64. Rademaker, R. L., Park, Y., Sack, A. T. & Tong, F. Evidence of gradual loss of precision for simple features and complex objects in visual working memory. *J. Exp. Psychol. Hum. Percept. Perform.* **44**, 925–940 (2018).

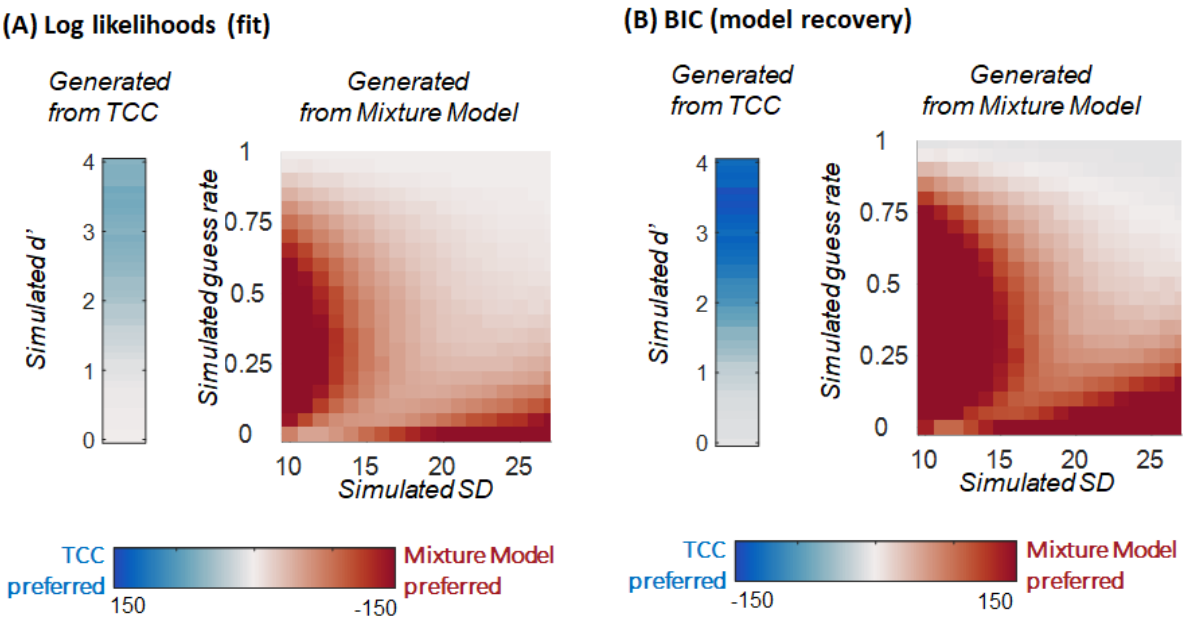
Supplemental Figures



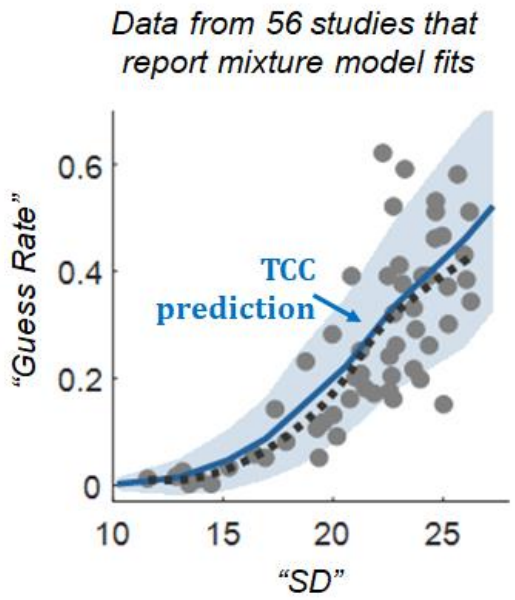
Supplementary Figure 1. Fits of TCC to the all encoding and delay conditions, including those not plotted in Fig. 3. TCC provides a strong fit at all encoding and delays (see correlations and model comparisons in Fig. 3).



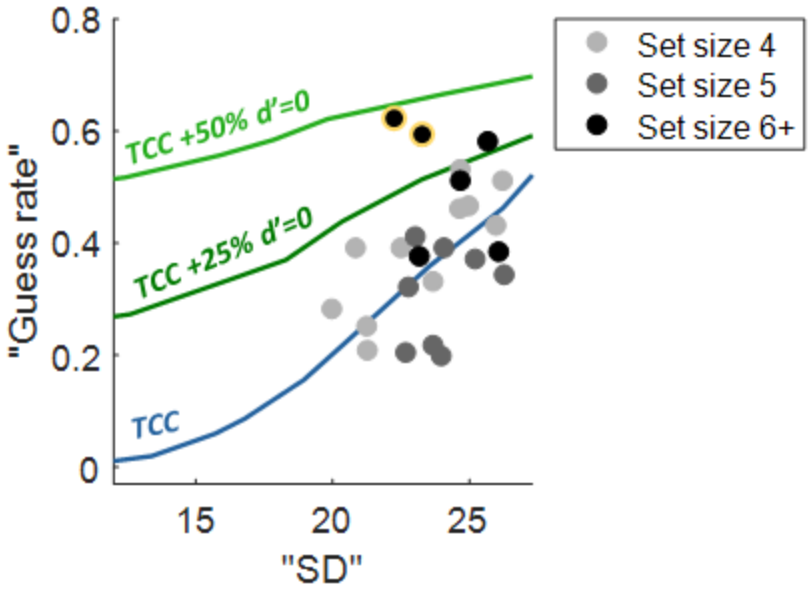
Supplementary Figure 2. Model predictions vs. data for 2-AFC generalization task reported in the main text. Given 2-AFC performance with maximally distinct 180 degree foils (black dot), TCC makes a unique prediction about exactly how well people should perform on other foils — with no free parameters. By contrast, using the 180 degree foils to constrain the mixture model allows this model to set the 'guess rate', but it leaves the precision of memory unknown. Thus, mixture models, while capable of fitting the data the same as TCC for a certain precision parameter (since ultimately they can predict any distribution TCC can, as they are much more flexible), do not make a unique prediction. Making strong predictions is the most critical test of a model[26] and can be formalized using a Bayes factor, which provides strong evidence in favor of TCC in this case. Similar logic applies in the experiment taking 180 degree 2-AFC and generalizing to continuous report and other n-AFC conditions.



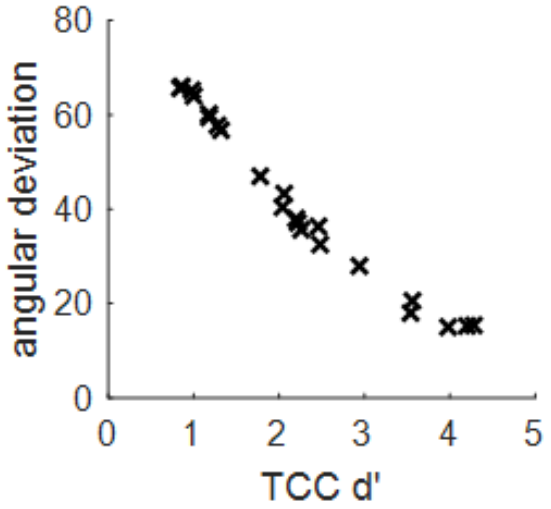
Supplementary Figure 3. Simulation of mixture model vs. TCC fits. (**A**) We generated data from both TCC (*d'*) and the standard mixture model (precision [SD] and guessing), performing 50 simulations of 2000 trials worth of data each for each of the models (consistent with the amount of group data in the main experiments), and then fit both models to the generated data to see which yielded a higher log-likelihood. With no penalty for complexity — simply using log likelihood — for data generated by TCC, the standard mixture model fit all data with a *d'* < 1 better than TCC itself. Thus, for data generated by TCC, the standard mixture model, being considerably more flexible than TCC in the range of distributions it can fit, fits the data about as well — and in some cases, better — than TCC. When fitting data generated by the mixture model, TCC was dispreferred at all values in terms of fit, and strongly dispreferred for huge swaths of potential mixture model parameters. This is because the mixture model can generate a huge variety of distributions that TCC cannot mimic. The same is true, but even more so, for the 3-parameter variable precision model, which can fit an even much larger range of distributions than even the standard 2-parameter mixture model. Only a miniscule part of the distributions predicted by the 3-parameter variable precision model can even be approximated by TCC, and this model can perfectly mimic TCC. (**B**) Same data, with BIC instead of log-likelihood. Taking into account model complexity increases the preference for TCC in TCC-generated data and creates a very slight TCC preference in mixture model data with simulated "guess rates" very near 1.0, where the two models make identical predictions in terms of error (of equal responding to all options); though note the two models make differing predictions about confidence at these values, predicting different ROCs. In general, with this amount of data, BIC appears well-calibrated, accurately recovering the appropriate model in nearly all cases and with a stronger preference for the relevant models where they diverge from each other more.



Supplementary Figure 4. Analysis of previous literature measuring the most widely used model parameters currently used to analyze working memory performance. Gray dots are values reported in papers found in the literature; the dashed black curve is a LOESS (local regression) smoothed version of these points. The solid blue curve reflects the average "guess" and "SD" parameters when fitting the mixture model to data generated by TCC, as a function of the *d'* of TCC. The blue shading shows 2 standard deviations when each participant has 100 trials/condition. Despite claiming to independently model multiple parameters, this entire diverse set of data points falls near the trade-off between these parameters predicted when fitting data sampled from the TCC with the 2-parameter model — in other words, one parameter is sufficient to capture much of the data observed in working memory tasks (data that has previously been thought to require at least two — and often 3 parameters — to explain). Note that the region in Supplementary Figure 4 TCC predicts is also the only region of Supplementary Figure 3 where the TCC can fit data generated from the mixture model. In addition, note that some of these papers use different color wheels than the one we use to generate the similarity function, and thus some of the deviation from the TCC prediction line — minor as it is — is caused by using an "incorrect" TCC prediction (e.g., using a prediction from an incorrect stimulus space).



Supplementary Figure 5. Analysis of previous literature measuring the most widely used model parameters currently used to analyze working memory performance. Existing working memory data from high set sizes (4+) is often claimed to provide evidence for 'slots' or for the existence of very low precision items, with these items that are unrepresented or poorly represented giving rise to the long tails of the distribution. By contrast, TCC predicts such long tails with no sense of unrepresented or poorly represented items. Here we show how TCC predicts that mixture model parameters from the standard two parameter mixture model should change as a function of *d'* in the TCC model. The blue line and all of the data points are the same as Supplementary Figure 4, but with the data points now labeled by set size and only "high" set sizes (>=4) plotted, as these are the points where traditional models claim many items must be unrepresented or extremely poorly represented. Note that the vast majority of the points are better fit by the straightforward TCC model — which simply assumes all items are equally well represented — than by models that add some proportion of 'unrepresented' items to TCC (plotted in green; note that as expected, these models selectively change the predicted 'guess rate' parameter). For a slot model prediction with 3 items represented, nearly 50% of items should be unrepresented at set size 6, and this is clearly incompatible with the previous data as well as the data we report in the main manuscript. In general, the parameters found in the previous literature are perfectly consistent with the basic TCC prediction with no added assumptions about unrepresented items or poorly represented items. Note that the two set size 6 points outlined in yellow come from the original Zhang and Luck[7] paper that introduced mixture models to this literature and used them to argue for slots. The fact that they are an outlier on this plot may be the reason those authors proposed a model that argues that only 'guess rate' but not 'standard deviation' changes as a function of set size.



Supplementary Figure 6. Plot of the best fit TCC d' vs. the circular standard deviation of the error data (a circular analog of the standard deviation; as computed with MATLAB's circular statistics toolbox circ_std function) for all 22 datasets from Fig. 3. For data like the current data where there is nearly no location-based confusions ('swaps'), the simpler analysis of this descriptive statistic (circular standard deviation, or more formally the angular deviation) is linearly related to d' for d' less than approximately 3.0, and thus, for data not near ceiling, may be an adequate substitute for fitting the full TCC. This is useful because the circular standard deviation is just a descriptive statistic of the data and thus does not require the collection of similarity data or perceptual confusability data. Note that just as with percent correct — which is approximately linear with d' when far from ceiling, but becomes deeply non-linear near ceiling — the d' curve begins to bend near ceiling. This is because improving from 95% correct to 99% correct requires a very large change in d', and similarly, improving your performance in continuous report when it is already very good requires a large change in memory strength. In theory the same should be true near floor, although these 22 datasets do not clearly demonstrate that because there is little data with d'<1.0. However, for data away from ceiling and floor and with little or no 'swaps', computing circular standard deviation may be sufficient to summarize data in a framework compatible with TCC.

Supplementary Tables

Supplementary Table 1 TCC's fit to binned color memory errors (Fig. 3). All correlations are Pearson correlations.

Set size experiment	
Set size 1	r=0.998, p<0.001, CI=(0.997, 0.999)
Set size 3	r=0.996, p<0.001, CI=(0.993, 0.998)
Set size 6	r=0.984, p<0.001, CI=(0.969, 0.991)
Set size 8	r=0.976, p<0.001, CI=(0.954, 0.987)

Delay experiment	*1 sec delay*	*3 sec delay*	*5 sec delay*
Set size 1	r=0.997, p<0.001, CI=(0.994, 0.998)	r=0.998, p<0.001, CI=(0.995, 0.999)	r=0.995, p<0.001, CI=(0.990, 0.997)
Set size 3	r=0.993, p<0.001, CI=(0.986, 0.996)	r=0.992, p<0.001, CI=(0.985, 0.996)	r=0.994, p<0.001, CI=(0.988, 0.997)
Set size 6	r=0.989, p<0.001, CI=(0.979, 0.994)	r=0.971, p<0.001, CI=(0.946, 0.985)	r=0.986, p<0.001, CI=(0.973, 0.992)

Encoding time experiment	*100ms encoding*	*500ms encoding*	*1.5 sec encoding*
Set size 1	r=0.992, p<0.001, CI=(0.984, 0.996)	r=0.997, p<0.001, CI=(0.994, 0.998)	r=0.998, p<0.001, CI=(0.996, 0.999)
Set size 3	r=0.971, p<0.001, CI=(0.945, 0.985)	r=0.991, p<0.001, CI=(0.983, 0.995)	r=0.995, p<0.001, CI=(0.990, 0.997)
Set size 6	r=0.975, p<0.001, CI=(0.952, 0.987)	r=0.993, p<0.001, CI=(0.987, 0.997)	r=0.990, p<0.001, CI=(0.981, 0.995)

Supplementary Table 2. TCC's fit to color memory data is reliably preferred by model comparison metrics that emphasize simplicity (e.g., BIC) across all set sizes compared to mixture models and variable precision mixture models. It provides a similar fit to these models when using leave-one-out cross validation on log likelihood, as both TCC as well as the two mixture models predict effectively the same distribution of errors when fit with N-1 error points (as N=2000 error datapoints >> the number of parameters for all models). Fitting to the group data rather than individual subjects gives BIC values at set size 1,3,6 and 8 of -24, -56, -26, -25 for TCC vs. standard mixture model (all very strong evidence favoring TCC), and BIC values of -2, -23, -15, -19 for TCC vs. variable precision model (e.g., both models fit set size 1 data well — the least distinct set size, since there are no long tails — but all others are very strong evidence in favor of TCC). Note that, as shown in Supplementary Figure 3, model recovery using BIC is well calibrated using this number of trials.

BIC avg. (S.E.M.); negative favors TCC	*Set size 1*	*Set size 3*	*Set size 6*	*Set size 8*
TCC - Mixture model	-3.64 (1.67)	-6.48 (0.95)	-6.08 (0.88)	-4.77 (0.67)
TCC - variable precision mixture model	-7.85 (1.14)	-10.65 (0.60)	-11.21 (0.67)	-10.82 (0.63)

Leave one out cross validation log likelihood difference (S.E.M.); positive favors TCC	*Set size 1*	*Set size 3*	*Set size 6*	*Set size 8*
TCC - Mixture model	1.54 (1.71)	1.22 (0.80)	0.14 (0.83)	0.07 (0.47)
TCC - variable precision mixture model	0.43 (1.32)	0.10 (0.43)	-0.31 (0.70)	0.21 (0.59)

Supplementary Table 3. TCC applied to face memory. As with colors, TCC is reliably preferred by model comparison metrics that emphasize simplicity (e.g., BIC) across all set sizes compared to mixture models and variable precision mixture models. Also, as with color, it provides a similar fit to these models when using leave-one-out cross validation on log likelihood, as both TCC as well as the two mixture models predict effectively the same distribution of errors when fit with N-1 points (as N >> the number of parameters for all models). Fitting to the group data rather than individual subjects gives BIC values at set size 1 and 3 of -177 and -24 for TCC vs. standard mixture model (all very strong evidence favoring TCC), and BIC values of -53, -10 for TCC vs. variable precision model (all very strong evidence in favor of TCC). Note that, as shown in Supplementary Figure 3, model recovery using BIC is well calibrated using this number of trials.

BIC avg. (S.E.M.); negative favors TCC	*Set size 1*	*Set size 3*
TCC - Mixture model	-8.1 (0.7)	-5.3 (0.4)
TCC - variable precision mixture model	-11.4 (0.5)	-10.8 (0.3)

Leave one out cross validation log likelihood difference (S.E.M.); positive favors TCC	*Set size 1*	*Set size 3*
TCC - Mixture model	2.5 (0.46)	0.51 (0.45)
TCC - variable precision mixture model	0.87 (0.41)	-0.05 (0.36)

Supplementary Table 4. Data points used in the literature review collected from a total of 14 papers.

SD	Guess	Paper	Set size	Notes	Digitized
13.9	0.01	Zhang & Luck 2008	SS1	None	In Paper
19.4	0.05	Zhang & Luck 2008	SS2	None	In Paper
21.9	0.17	Zhang & Luck 2008	SS3	None	In Paper
22.3	0.62	Zhang & Luck 2008	SS6	None	In Paper
20.8	0.16	Zhang & Luck 2008	SS3	None	In Paper
23.3	0.59	Zhang & Luck 2008	SS6	None	In Paper
22.9	0.26	Zhang & Luck 2009	SS3	Retention interval 1 second	In Paper
24.4	0.26	Zhang & Luck 2009	SS3	Retention interval 4 seconds	In Paper
24.4	0.39	Zhang & Luck 2009	SS3	Retention interval 10 seconds	In Paper
14.5	0.001	Bays, Catalao & Husain 2009	SS1	100ms, Collapsed with swaps	X
19.3	0.105	Bays, Catalao & Husain 2009	SS2	100ms, Collapsed with swaps	X
23.7	0.33	Bays, Catalao & Husain 2009	SS4	100ms, Collapsed with swaps	X
24.7	0.51	Bays, Catalao & Husain 2009	SS6	100ms, Collapsed with swaps	X
13.5	0.001	Bays, Catalao & Husain 2009	SS1	500ms, Collapsed with swaps	X
17.9	0.08	Bays, Catalao & Husain 2009	SS2	500ms, Collapsed with swaps	X
20	0.281	Bays, Catalao & Husain 2009	SS4	500ms, Collapsed with swaps	X
26.1	0.383	Bays, Catalao & Husain 2009	SS6	500ms, Collapsed with swaps	X
13.2	0.0245	Bays, Catalao & Husain 2009	SS1	2000ms, Collapsed with swaps	X
16.45	0.0565	Bays, Catalao & Husain 2009	SS2	2000ms, Collapsed with swaps	X
21.3	0.207	Bays, Catalao & Husain 2009	SS4	2000ms, Collapsed with swaps	X
23.2	0.375	Bays, Catalao & Husain 2009	SS6	2000ms, Collapsed with swaps	X
18.79	0.23	Fougnie, Aslplund & Marois, 2010	SS3	Single feature	X
25.28	0.3	Fougnie, Aslplund & Marois, 2010	SS3	Conjunction (w/ orientation)	X
21.5	0.18	Fougnie, Aslplund & Marois, 2010	SS3	Single feature	X
22.78	0.52	Fougnie, Aslplund & Marois, 2010	SS3	Conjunction (w/ orientation)	X
24.7	0.53	Zhang & Luck 2011	SS4	Low Precision, None	X
26.24	0.51	Zhang & Luck 2011	SS4	High Precision, None	X
22.53	0.39	Zhang & Luck 2011	SS4	Low Precision, Feedback provided	X
20.87	0.39	Zhang & Luck 2011	SS4	High Precision, Feedback provided	X
26	0.43	Zhang & Luck 2011	SS4	Low Precision, Payoff provided	X
24.67	0.46	Zhang & Luck 2011	SS4	High Precision, Payoff provided	X
11.6	0.011	Fougnie, Suchow & Alvarez 2012	SS1	None	In Paper
17.4	0.141	Fougnie, Suchow & Alvarez 2012	SS3	None	In Paper
23.7	0.216	Fougnie, Suchow & Alvarez 2012	SS5	None	In Paper
21.06	0.2	Brady & Alvarez 2015	SS1	None	Lab Data
22.6	0.24	Brady & Alvarez 2015	SS3	None	Lab Data
25.7	0.58	Brady & Alvarez 2015	SS6	None	Lab Data
22.7	0.203	Fougnie et al 2016	SS5	Only included single report condition	In Paper
24	0.197	Fougnie et al 2016	SS5	Only included single report condition	In Paper
26.3	0.342	Xie & Zhang, 2016	SS5	None	X
23.8	0.29	Suchow, Fougnie, Alvarez 2016	SS3	Random report	In Paper
12.9	0.015	Swan, Collins & Wyble, 2016	SS1	Pre-surprise	In Paper
15.3	0.032	Swan, Collins & Wyble, 2016	SS1	Post-surprise	In Paper
21.3	0.25	Bocincova et al. 2017	SS4	None	In Paper
16.9	0.05	Bocincova et al. 2017	SS2	None	In Paper
19.6	0.117	Wee et al 2013	SS3	Short delay (1 sec)	X
22.6	0.174	Wee et al 2013	SS3	Long delay (10 sec)	X
22.8	0.32	Wang et al, 2016	SS5	On-probe, 200ms SOA	X
24.1	0.39	Wang et al, 2016	SS5	Off-probe, 200ms SOA	X
23.05	0.41	Wang et al, 2016	SS5	On-probe, 400ms SOA	X
25.24	0.37	Wang et al, 2016	SS5	Off-probe, 400ms SOA	X
20.04	0.13	Fougnie, Aslplund & Marois, 2010	SS3	Single feature 6 features	X
25.06	0.15	Fougnie, Aslplund & Marois, 2010	SS3	Conjunction (w/ orientation) 6 features	X
20.21	0.09	Fougnie, Aslplund & Marois, 2010	SS3	Single feature 6 features	X
22.77	0.16	Fougnie, Aslplund & Marois, 2010	SS3	Conjunction (w/ orientation) 6 features	X