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

nature human behaviour

There are amendments to this paper

Psychophysical scaling reveals a unified theory of visual memory strength

Mark W. Schurgin ✉, John T. Wixted and Timothy F. Brady ✉

Almost all models of visual memory implicitly assume that errors in mnemonic representations are linearly related to distance in stimulus space. Here we show that neither memory nor perception are appropriately scaled in stimulus space; instead, they are based on a transformed similarity representation that is nonlinearly related to stimulus space. This result calls into question a foundational assumption of extant models of visual working memory. Once psychophysical similarity is taken into account, aspects of memory that have been thought to demonstrate a fixed working memory capacity of around three or four items and to require fundamentally different representations—across different stimuli, tasks and types of memory—can be parsimoniously explained with a unitary signal detection framework. These results have substantial implications for the study of visual memory and lead to a substantial reinterpretation of the relationship between perception, working memory and long-term memory.

Working memory is typically conceptualized as a fixed capacity system with a discrete number of items, each of which is represented with a certain degree of precision[1,2]. It is thought to be a core cognitive system[3,4], with individual capacity differences strongly correlating with measures of broad cognitive function such as fluid intelligence and academic performance[5,6]. As a result, many researchers are deeply interested in understanding and quantifying working memory capacity and understanding the connections between working memory and long-term memory.

Continuous feature spaces are often used to investigate memory, as they allow the precise quantification of information stored in memory[2,7,8]. One prominent method involves researchers presenting a set of stimuli to remember and then probing one item after a delay, asking participants to report the target by clicking on a circular stimulus report wheel (Fig. 1a). The data are typically analysed using the circular difference between the true stimulus and reported stimulus, which is then modelled to quantify memory performance[7,8]. Because errors that arise in this task have a 'fat tail' (that is, there are more far away errors than you might expect; Fig. 1b), the dominant models of working memory draw critical distinctions between fundamentally different kinds of memory errors: those caused by limits in how many items are represented versus how precisely they are represented[7] or those caused by items encoded with high precision versus extremely low precision[8].

Here, we present evidence that these small versus large errors are not distinct kinds of errors and do not represent multiple psychological constructs being measured (for example, precision versus guessing). Instead, we demonstrate that these responses arise fundamentally from a single process. To describe this new conceptualization of memory, we begin with working memory for colour as our main case study and then expand the model to encompass working memory for faces (a multi-feature stimulus space) and long-term memory for real-world objects.

The model we propose is a straightforward extension of standard signal detection-based accounts of memory, with the fundamental insight of our framework being the nature of the psychophysical similarity function that explains how familiarity spreads. Consider the simplest case of memory: being asked to remember just a single colour. When you encode this colour (for example, red), it will now have notably enhanced familiarity. Thus, if you are later asked to distinguish the colour you saw from a foil colour (for example, red versus green), the colour you saw will probably be more familiar. However, due to noise that corrupts the familiarity signals, this will not always be the case, and on some trials, green might feel more familiar than red.

The critical insight of our model is that when you see red, it does not boost only familiarity associated with red. Instead, a gradient of familiarity will spread to other colours according to a fixed psychophysical similarity function, with considerable activity spreading to similar colours (for example, pink will also feel familiar), but with much less spreading to dissimilar colours (for example, yellow, blue and green will lead to virtually no boost in familiarity). If asked to hold this colour in mind, these initial familiarity signals will be corrupted by noise, and when memory is probed (for example, if people are asked to report what colour they saw on a colour wheel), people will report the colour of the response option that currently has maximum familiarity. Although the encoded colour is most likely to generate the maximum familiarity signal, competition from other colours (especially from similar colours) ensures that this will not always be the case, and the more noise that accumulates, the more likely a very dissimilar colour will be reported. Notably, in this model, memory is not simply a point representation ("I think this item is red") but instead an entire population of familiarity signals (similar to neural models[9–11]). (We have built an interactive demonstration of this model at https://bradylab.ucsd.edu/tcc/ to explain it dynamically.)

According to the model, the way familiarity spreads is a fixed perceptual property—one that can be independently measured using a conventional psychophysical similarity function. Once the nature of the familiarity gradient for a given stimulus space is measured, memory is simply modelled by taking this fixed property of the stimuli and adding noise, with d' being the only memory-based parameter of the model. This model thus uniquely explains the complex shape of error data with only a single free parameter (memory strength, d') and permits parameter-free generalization across different tasks (that is, without any free parameters, using

Department of Psychology, University of California San Diego, La Jolla, CA, USA. ✉e-mail: mschurgin@ucsd.edu; timbrady@ucsd.edu

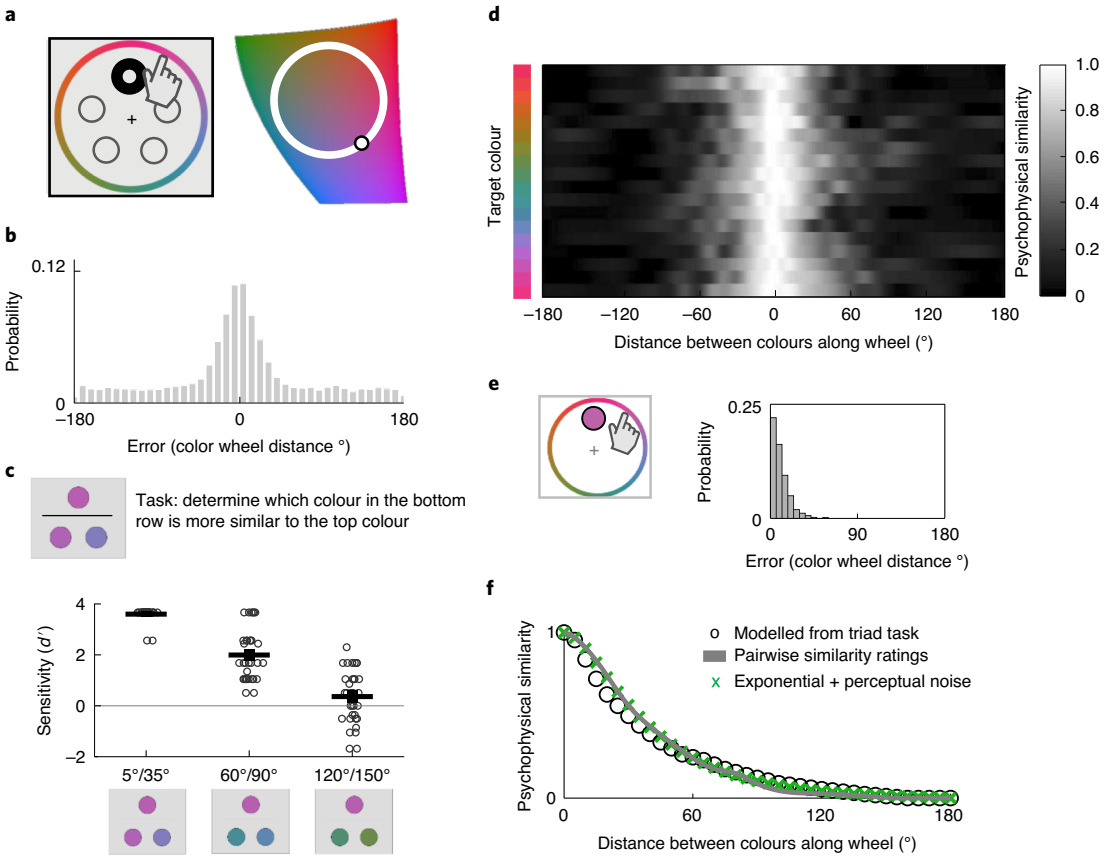


Fig. 1 | Measuring visual memory and measuring psychophysical similarity. a, An example of a continuous response memory task. A widely used method in working memory is to select a colour circle from a slice of colour space (right; black circle), show memory items drawn from this circle and then, during the test, probe the contents of a memory item by presenting the entire continuous circle to participants and asking them to give a response (left). Similar response wheels are used for other features, such as face identity. **b**, A histogram of results generally observed for such tasks, traditionally plotted as a function of distance in degrees of error along the response wheel. There are long, fat tails of errors far from 0 that are often interpreted as evidence for distinct memory states (for example, guesses or items encoded with very low precision). **c**, Top: in a triad psychophysical scaling task, n=40 participants had to say which of two colours in the bottom row was more similar to the top (target) colour. Bottom: despite the difference between the two choice colours always being 30° on the colour wheel, sensitivity (d') dramatically decreased as the choices became more distant from the target (ANOVA, $F(12,384)=71.8$; $P<0.00001$; $\eta^2=0.69$). Error bars show within-participant s.e.m. and dots represent individual participants. See Extended Data Fig. 1 for the full data. **d**, We can use the data from another similarity task (that is, a simple pairwise Likert rating of similarity (n=50)) to infer the global psychophysical distance of colours at different physical distances along the colour wheel. Here we plot these data for sets of target colours, demonstrating previously observed local non-uniformities in colour space as the small differences across rows[12]. Critically, all of these rows demonstrate a much larger global structure that is separate from this local structure: overall similarity falls in an approximately exponential manner. **e**, Some aspects of this similarity must derive from perceptual discrimination failures (for example, there are not really 360 independent colours on the colour wheel). To estimate this underlying perceptual noise, we use a continuous report task where participants must match a visible colour using the same colour wheel (n=40). **f**, We can plot the global psychophysical function (averaged over all target colours) using the triad task or the Likert task. Both are very similar and show the same underlying shape. Consistent with previous work, we find that this similarity function is exponential once perceptual noise is taken into account (for example, an exponential convolved with the measured perceptual noise function provides an excellent fit to these data).

only measured memory strength and similarity values from different participants). Because this model operates in a signal detection framework, as most models of long-term memory do, it also suggests that a unified framework can be used to understand the nature of mnemonic representations and decision-making across working memory and long-term memory.

Results

Psychophysical similarity. The most critical component of our proposed model is the psychophysical similarity function that explains how familiarity spreads within a stimulus space (for example, across the colour wheel). While previous work has documented local inhomogeneities in the structure of stimulus spaces[12–14], we were primarily interested in the global structure of similarity: for a stimulus 10° away on the colour wheel from a target colour (regardless of what the target colour is), how similar is this colour to the target on average? Thus, we measured how similarity scales with distance measured in terms of degrees along the colour wheel (see Methods section 'Fixed-distance triad experiment'). To do so, we tested how accurately participants could determine which of two test colours was closer in colour space to a target colour using a triad task[15,16]. This is a perceptual task, but it is analogous to the working memory situation where participants have a target colour in mind and are asked to compare other colours to that target. We found that with a fixed 30° distance between two colour choices, participants are significantly more accurate at determining which colour is closer to the target when the two colours are close to the target in colour space compared with when they are far from the target

(Fig. 1c and Extended Data Fig. 1; analysis of variance (ANOVA), $F(12,384) = 71.8$; $P < 0.00001$; $\eta^2 = 0.69$). In other words, in a purely perceptual task, participants largely could not tell whether a colour 120° or 150° from the target was closer to the target, whereas this task was trivial if the colours were 5° and 35° from the target. This demonstrates a strong nonlinearity in perceptual similarity.

To compute a full psychophysical similarity function, we utilized the just-described triad task with additional distance pairs (see Methods section 'Psychophysical scaling triad experiment'). We then applied the maximum likelihood difference scaling (MLDS) technique[16] that is commonly used for perceptual scaling to estimate how differences between colour stimuli are actually perceived. The estimated psychophysical similarity function fell off in a nonlinear, exponential-like fashion with respect to distance (Fig. 1f). In colour space, it was also well matched by a smoother measure that required substantially less data; namely, the pairwise subjective similarity ratings of colours at different distances along the colour wheel using a Likert scale (see Methods section 'Likert colour similarity experiment'; Fig. 1f).

While there were also small local inhomogeneities (Fig. 1d), we were primarily interested in the fact that the global structure of similarity space was strongly nonlinear, in agreement with decades of work suggesting that psychological similarity is globally exponential (for example, the universal law of generalization[17,18]), with confusions for very similar colours also caused by perceptual noise[19] (measured here using a perceptual matching task; see Methods section 'Perceptual matching experiment' and Fig. 1e,f).

A key implication of these similarity scaling results is that the linear axis of error along the response wheel (for example, −180° to 180°) that was previously used to analyse working memory capacity does not capture the psychological representation of the stimuli. This poses a serious challenge to existing memory models, as their parameters are derived assuming linear similarity (that is, treating the axis of error in degrees as a linear scale). However, this axis is not linear, even in a perceptual task: since participants are essentially incapable of discerning whether an item 120° or 180° from the target in colour space is more similar to the target, it is not surprising that they confuse these colours equally often with the target in memory.

Incorporating psychophysical similarity into a signal detection model. Psychophysical scaling formalizes how similar two stimuli are perceived to be and is the first aspect of our proposed model. The next aspect is that signals are corrupted by noise, which we formalize using signal detection theory.

In particular, the model we propose here is fundamentally the same longstanding signal detection model used across decades of research on long-term memory and perception[20–22], modified to take into account psychophysical similarity. The basis of signal detection theory is that when deciding among each of the colours at test, participants rely on a noisy, cue-dependent familiarity signal for each colour, and the colour that generates the maximum familiarity signal is selected (Fig. 2). The stronger the maximum signal, the higher the confidence in the selected colour.

Our model differs from a standard model of the n-alternative forced choice (n-AFC) only in the usage of the psychophysical similarity measure. In a standard signal detection model of an n-AFC task, it is generally assumed that exactly one item has been previously seen, so its familiarity is centred on d', whereas the other $n-1$ items are equally unfamiliar and therefore centred on zero familiarity[21]. However, when memory is tested using a continuous stimulus space, it would be implausible to assume that a colour 1° away in colour space from the target would have no added familiarity and would have noise that is totally uncorrelated with the target.

Thus, in our model, the mean memory signal for a given colour x on the colour wheel, denoted d_x, is based on that colour's separately measured similarity to the target (that is, $d_x = d'f(x)$), where d' is the model's only free parameter (memory strength) and $f(x)$ is the empirically determined psychophysical similarity function (that is, a measurement, taken from different participants, of the similarity structure of the colour space). The noise added to each colour is also correlated between nearby colours according to the empirically measured proportion of how often colours at that distance are confused in a perceptual matching task (Fig. 1e), although this is not critical for fitting continuous report error distributions (Extended Data Fig. 2).

Because of the nonlinear similarity function, colours in the physical distance range around >90° all cluster near $f(x) \approx f(x)_{\min}$ such that $d_x \approx 0$ for $x = 90°$ to 180°. Thus, when participants encode a colour (for example, purple), it increases the average familiarity signal in the purple channel and also in nearby (similar-to-purple) channels while having almost no effect in dissimilar colour channels (Fig. 2b). The familiarity signals in each channel are then corrupted by noise, and the resulting reports are based on this noisy signal. In the case of continuous reports, people theoretically report the colour with maximum familiarity.

Importantly, this target confusability competition (TCC) model can explain all of the key features of visual working memory. In particular, it accurately characterizes memory performance across a variety of domains, including different set sizes, encoding times and delays (Fig. 3 and Supplementary Fig. 1). Previous cognitive models of visual working memory allow for many ways in which memory for an individual item can vary (for example, guess rate, precision and variation in precision[7,8,23]). In contrast, TCC holds that these experimental manipulations affect only a single fundamental underlying parameter (the memory strength parameter, d'), and that the complex changes in the shape of the error distribution arise not from multiple parameters, but simply from the similarity function combined with the nonlinearity inherent in selecting only your strongest familiarity value for report. Thus, the fact that manipulations of set size, delay and encoding time (22 different manipulations in total) result in distributions that can be accurately characterized with only a single varying parameter is strong evidence in favour of TCC, as is the fact that it describes the data extremely well despite being markedly simpler than alternative theories. It is markedly simpler because it proposes a unified generative process for all responses instead of requiring different states to generate different subsets of responses (as in the encoding variability or lack of represented items proposed by previous models[7,8,23]), and because it replaces free parameters (such as precision) with independently measured values (such as similarity, which is independently measured and fixed for all participants and conditions; Extended Data Fig. 4).

The measured nonlinear similarity function is critical to the ability of TCC to fit the data. While reporting the colour that is maximally familiar does, on its own, introduce a nonlinearity that favours the strongest signals, this alone is not sufficient to explain the data (Extended Data Fig. 3). Instead, the explanatory value of TCC comes from the combination of the nonlinear similarity function and signal detection theory.

While the main evidence in favour of TCC is its ability to parsimoniously characterize the effects of qualitatively different experimental manipulations (Fig. 3 and Supplementary Table 1) and to make precise predictions across tasks and stimuli (see below), we also compared the fit provided by TCC with the fit provided by mixture models of visual working memory, including the standard two-parameter mixture model that interprets performance as arising from distinct concepts of capacity and precision[7] and a three-parameter version of the mixture model that allows for variable precision[23]. Despite being simpler and having fewer parameters, TCC was just as good at predicting held-out data in a cross-validation test and was reliably preferred in every participant

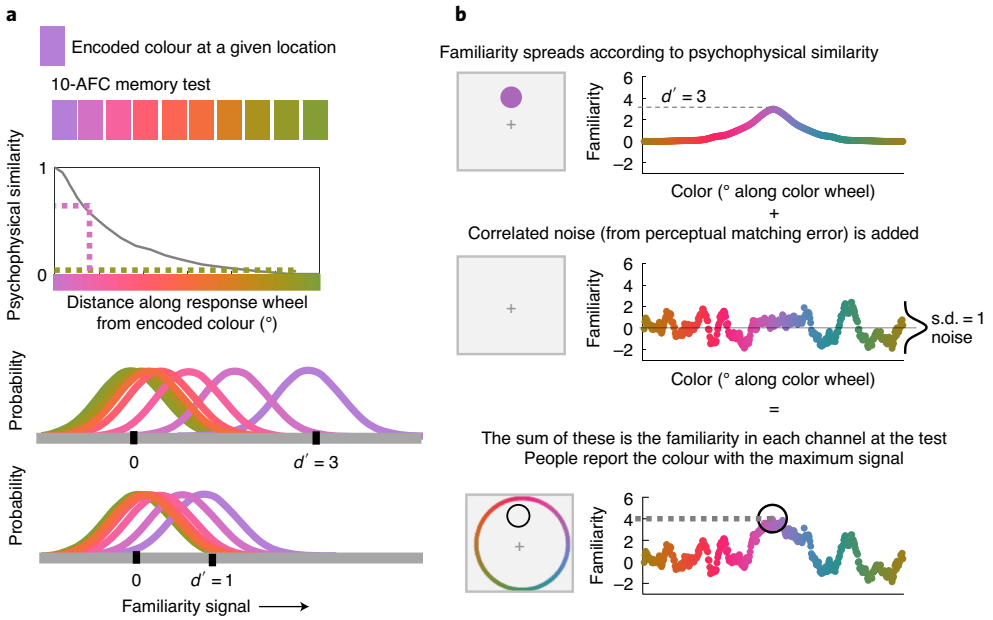


Fig. 2 | The target confusability competition (TCC) model of visual memory. a, Our TCC model applied to a hypothetical 10-AFC memory test. In standard two-alternative long-term recognition memory experiments, unseen items vary in familiarity, which is modelled as a normal distribution. Previously encoded items elicit higher familiarity on average, modelled (in the simplest case) as a normal distribution with a mean of d', where d' indicates how many standard deviations of memory strength are added to seen items. When asked what they remember, people pick whichever colour elicits higher familiarity on that trial. To generalize to a 10-AFC, we thus only need to specify the average familiarity strength of every lure. Usually, all nine lures are assumed to have a mean of zero, with no added familiarity, when modelling such tasks[21]. However, in a continuous space, this is not plausible. Thus, in TCC, we propose that familiarity spreads according to similarity: the mean of each lure's familiarity distribution is simply its similarity to the target. For example, if the target is purple, other purples will have boosted familiarity as well, and thus people will choose a slightly different purple lure much more often than an entirely unrelated lure such as green. Examples of d'=3 and d'=1 illustrate the idea that when memory for the target colour is weaker, more of the lure distributions cluster near the target, and at d'=1, all of the distant colours are in a position to sometimes 'win the competition' by having the highest familiarity, but will do so on average equally often, creating a long fat tail. The 10-AFC logic provided here can then simply be adapted to 360-AFC to model continuous report, but with the added knowledge that very similar colours also have correlated noise (measured using the perceptual matching function); that is, there are not 360 independent colours on the colour wheel. **b**, An alternative way of plotting the same model is to consider a single trial, rather than the distribution of memory strengths across trials. When we encode a purple colour, with memory strength d'=3, the familiarity of purple as well as similar colours is increased (according to the measured psychophysical similarity function). Then, we add s.d.=1 noise to each colour channel. The resulting familiarity values, after being corrupted by noise, guide participants' decisions. In a continuous report task, people simply report the colour that generates the maximum familiarity value. For an interactive explanation of the model, see https://bradylab.ucsd.edu/tcc/.

across set sizes when using metrics preferring simpler models (Supplementary Table 2). This was true even though TCC fits are based on aggregated similarity functions from a different group of participants, suggesting that the global structure of the psychophysical similarity function is largely a fixed aspect of a given stimulus space. Taking into account colour-specific similarity functions (for example, Fig. 1d) or individual differences in similarity scaling should further improve the fit of the model (Extended Data Fig. 5), and would be necessary for comparing the model with others that do take into account such information, but here we focus on the general case of treating all colours and participants as sharing a similarity function.

While memory strength varies according to a variety of different factors (Fig. 3), many researchers have been particularly interested in the influence of set size. TCC shows that at a given encoding time and delay, d' (theoretically an interval-scale measure of memory strength[21,24]) decreases according to a power law as the set size changes (Extended Data Fig. 6), broadly consistent with fixed resource theories of memory[25,26]. Critically, memory strength decreases most at low set sizes (for example, one to three), suggesting that limits of working memory may be best studied across lower set sizes, in contrast with the majority of the field, which seeks to pressure capacity via high set sizes to understand the nature of working memory.

TCC accurately predicts connections between qualitatively different tests of working memory that mixture models claim are impossible. Ultimately, evaluating theories based on model comparisons of fit—when all models fit the data well, as is the case here—is not as useful as investigating what they accurately predict[27]. TCC makes a precise and unique prediction that since all responses are generated from the same underlying process, measuring d' in any way that avoids floor and ceiling performance—even using only two maximally dissimilar 180° away colours in a two-alternative forced-choice (2-AFC) task—is sufficient to accurately predict (with no free parameters) memory performance involving more similar colours and/or more response options (including continuous report). This is in direct contrast with the inability of mixture models and variable precision models to make such predictions. Such models claim that memory varies in multiple fundamentally distinct ways (that is, precision and guessing can both vary, or the distribution of precisions can vary), and clearly, a single measure of accuracy cannot possibly measure more than one fundamental distinct property of memory.

Specifically, existing models insist that such predictions should not be possible because they claim that heterogeneity between items is crucial to explaining large versus small errors. That is, existing models claim that fundamentally distinct items and memory states explain close-to-target responses on the colour wheel (for example,

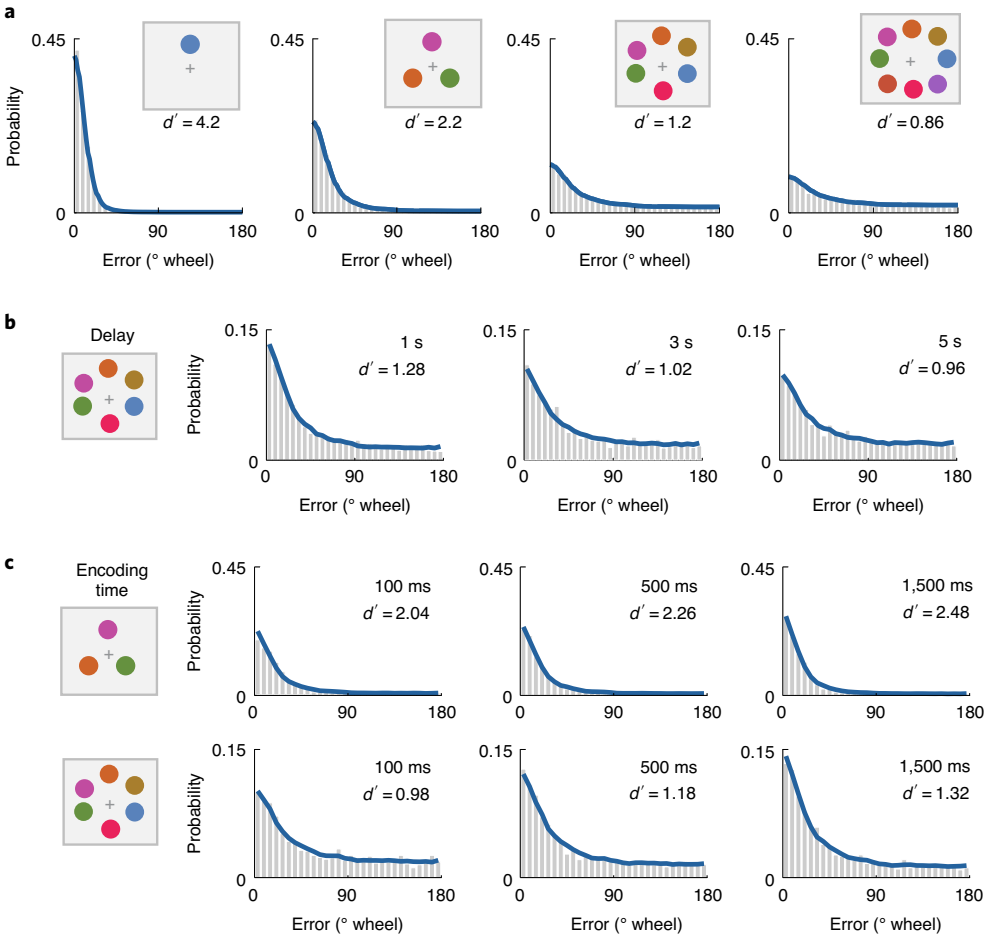


Fig. 3 | TCC accurately fits a variety of working memory data. a, TCC fits to group data at set sizes of one, three, six and eight (n=20). Despite no concept of unrepresented items or guessing or poorly encoded items, and adopting for the sake of simplicity the assumption that all items are encoded equally (that is, with the same d'), TCC fits even the larger set size data accurately because of the noisy nature of the signal detection process combined with the nonlinear psychophysical similarity function. **b**, TCC fits to n=20 group data with varying delay (only a set size of six is shown; the remainder of the data are shown in Supplementary Fig. 1). **c**, TCC fits to n=20 group data across different encoding times (only two set sizes are shown; see Supplementary Fig. 1). Across several key manipulations of visual working memory (set size, delay and encoding time), which drastically alter the response distributions, TCC accurately captures (with only a single free parameter d') the response distribution typically attributed to multiple parameters or psychological states by existing frameworks and models of working memory. Only a subset of the delay and encoding time fits are plotted here, but all fits are accurate, as demonstrated by the Pearson correlation between the binned data and model fits as a function of set size (Supplementary Table 1). Note that d' of the fit to the group data, as plotted, is not the same as the average of individual participant d' values, as used in the model comparisons.

precision errors for remembered items or high-precision items) versus responses far away from the target (for example, guesses or low-precision items). Thus, existing models inherently assume that a singular measure of how well participants can discriminate 180° changes (for example, was it red or green?), which measures only information about items that cause large errors, cannot, even in principle, measure the properties of the items that cause small errors. In contrast, TCC says that all responses to more similar colours are directly predictable using the fixed similarity function, and that memory varies in only one way (memory strength); thus, such a 2-AFC task is sufficient to measure memory performance.

In two experiments, we tested TCC's prediction that a single measured d' is sufficient to characterize memory performance across a variety of tasks that are currently thought to tap different memory processes. In both experiments, we asked participants to perform a memory task involving a 2-AFC test with maximally dissimilar colours (two options: 0° away from the target colour versus 180° away from the target colour). We used the data from this 2-AFC task to compute d' in the standard way (denoted $d'_{180°}$) and then used TCC—with this exact d'—to compute parameter-free predictions for a variety of other conditions. We intermixed all of the conditions, including conditions that required participants to remember the precise colour they saw, so that participants could not rely on a categorical memory strategy in the maximally distinct 2-AFC task.

In one experiment involving a 2-AFC task (Fig. 4), we used TCC with fixed $d'_{180°}$, to predict how well participants could discriminate the target from more similar foils (for example, to predict $d'_{12°}$ from a 2-AFC task involving the colour they saw versus a colour only 12° away). With no free parameters, memory performance was accurately predicted over the entire range of intermediate foil similarities (Fig. 4c). TCC accomplished this with no free parameters because it specifies how the perceptual similarity of the two colours on a 2-AFC task (measured in a separate psychophysical procedure) should impact memory performance (see also Kahana and Sekuler[28] and Nosofsky[19]). In contrast, mixture models, based on the distinct concepts of guessing and precision, anticipate no particular relationship between performance on a 2-AFC task involving maximally dissimilar foils and performance on a 2-AFC task involving

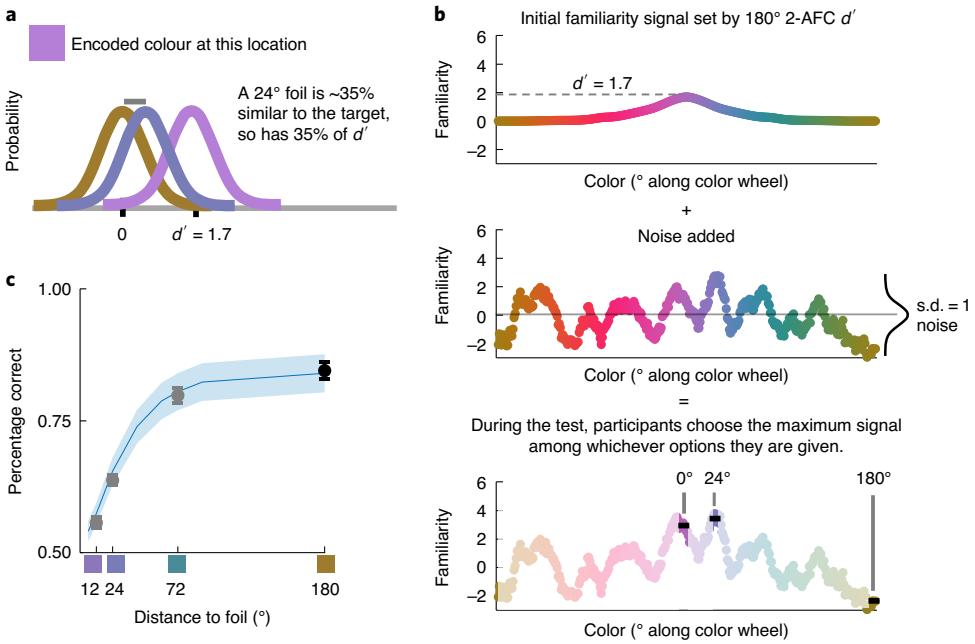


Fig. 4 | TCC predicts how memory performance changes for more similar foils. a, Since TCC states that visual working memory performance is determined by simply *d′* (memory signal strength) once perceptual similarity is known for a given feature space, it makes novel predictions that no other theory of working memory can make. In particular, it predicts that a *d′* measured with a 180°, maximally dissimilar foil (that is, $d'_{180°}$) should be completely sufficient to predict all of memory performance, unlike models in which errors to maximally dissimilar foils arise from different processes from errors to similar foils (for example, where errors to maximally dissimilar foils arise solely from guessing (in some models) or from extremely poorly encoded items (in other models)). For example, after measuring $d'_{180°}$, TCC predicts that since a 24° foil is ~35% similar to the target, discriminability on a 2-AFC task in which the foil is 24° away from the target should be ~35% of $d'_{180°}$ (but note that correlated noise makes this more complex for very similar foils). **b**, On a single trial, this prediction can be visualized in a straightforward way. If we know the target was encoded with $d'_{180°}$=1.7, then TCC makes a strong prediction about how this familiarity spreads to other colours and how it is corrupted by noise. In continuous report, the decision rule is to report the maximum of the resulting colour channel familiarity responses. In 2-AFC, the decision rule—based on the exact same underlying colour channel responses—is to choose the highest-familiarity signal of the response options. Thus, in this example trial, the participant in a 2-AFC task would choose the 0° target over a 180° foil, but would choose a 24° foil over the 0° target. Because TCC specifies this entire generative process, it makes precise predictions about how often people will make errors to different distance foils. **c**, Predicted percentage of correct responses for different distances of colours from the target (blue)—a prediction based only on performance from the 180° condition (black) with no free parameters. Blue shading indicates TCC prediction (based on 180° condition only). When comparing participants' performance at different foil distances (grey; *n*=60), we demonstrate that TCC accurately predicts performance across different foil distances.

more similar foils. Two-parameter mixture models can use 180° 2-AFC performance only to measure the guess rate, leaving precision unspecified. Thus, with only 180° 2-AFC performance in hand, these models are able to predict a wide range of possible outcomes on 2-AFC tasks with more similar foils, depending on the unknown factor of memory precision (Supplementary Fig. 2). Note that precision, unlike similarity, is thought to be changed by memory strength and to differ across participants; thus, precision measures are not constrained by fixed perceptual similarity data that TCC can utilize so effectively. Because the mixture model predictions are largely unconstrained, TCC is strongly preferred to mixture models by a Bayes factor model comparison (group Bayes factor preference for TCC > 200:1; individual participants: $t(54) = 11.19$; $P < 0.001$; $d_z = 1.51$; confidence interval (CI) = 2.9:1 to 4.2:1).

In a second experiment, we went further, showing that TCC—again using only measured $d'_{180°}$ from a 2-AFC task and separately measured perceptual similarity between the response option colours in different participants—can accurately predict performance when there are more than two response options, up to and including continuous report, again with no free parameters (Fig. 5). In this experiment, we once again found a strong preference for TCC's prediction over the mixture model models in generalizing from 2-AFC to continuous report, which is the only condition the mixture model can be fit to (group Bayesian information criterion (BIC) preference for TCC > 650:1; individual participants: $t(51) = 7.64$; $P < 0.001$; $d_z = 1.06$; CI = 9.5:1 to 16.2:1). We also found that 2-AFC *d′* measured in the standard way (that is, $d'_{180°}$) maps directly to TCC's *d′*, which explains the full continuous report distribution (Fig. 5b). The lopsided Bayes factors arise because TCC precisely predicts the outcomes (outcomes that, when tested, are empirically observed), whereas competing models necessarily claim that the 2-AFC data are insufficient to completely measure memory since they do not measure the precision of memory.

Thus, with TCC, measuring only how well participants can distinguish between far apart test items (0° versus 180°) using a 2-AFC task is sufficient to predict the distribution of responses from a continuous report task and to predict 2-AFC performance for distinguishing targets and foils of varying similarity (so long as the 2-AFC task is not at the ceiling or floor). Together, these experiments provide compelling evidence against previous models of visual working memory in which the tails of the continuous report distribution (the only aspect of performance that is theoretically measured with 180° foils in 2-AFC) are fundamentally distinct from the centre of the distribution.

In other words, in the competing models, responses in the tails of the distribution result from guesses or low precision, whereas

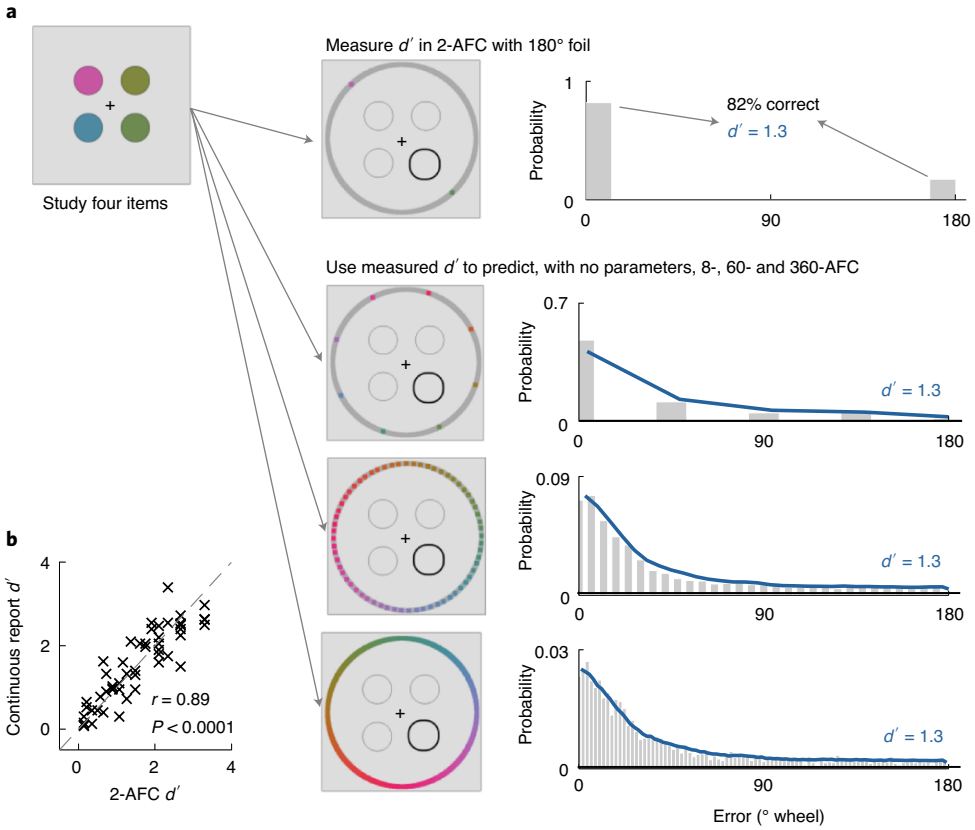


Fig. 5 | TCC allows zero-free-parameter generalizations from 2-AFC all the way to continuous report. a, According to TCC, the *d*′ in a 2-AFC task is fundamentally the same *d*′ in continuous report tasks (or any other AFC task). Thus, unlike other models, TCC makes a strong prediction that *d*′, as measured with a 180° foil ($d'_{180°}$) is completely sufficient to predict all of memory across any number of options presented at the test, including being completely sufficient to predict the entire distribution of errors in continuous report (since, ultimately, this distribution does not arise from distinct psychological states, but simply from combining the fixed similarity structure of the stimulus space with memory strength). To test this prediction, *n*=60 participants encoded items into memory and were then tested using 2-AFC, 8-AFC, 60-AFC or continuous report (360-AFC). During 2-AFC trials, the foil was always 180° away, which we used to calculate $d'_{180°}$. We then used TCC, with this measured *d*′, but with no free parameters, to accurately predict 8-, 60- and 360-AFC performance. The accuracy of these predictions provides further evidence that there is no need for forgotten or low-precision items to account for the tail of continuous report distributions. Instead, for a given stimulus space, the continuous report distribution is modulated by memory strength but is otherwise always the same shape, determined by the shape of the similarity function for that stimulus space. **b**, We can also independently estimate *d*′ from the continuous report data and from the 2-AFC data. We find a strong participant-level correspondence between TCC's continuous report estimate of *d*′ and *d*′ estimated from the 2-AFC task in the traditional way (that is, $d'_{180°}$) (Pearson's *r*=0.89; *P*<0.001; CI=0.81 to 0.93), in line with what is expected simply from the noise ceiling of these measurements. Each point is a participant mean.

the central responses result from high-precision memories. If these models were correct, it should not be possible for TCC to make such accurate predictions across tasks using a single *d*′ and no free parameters. The fact that TCC can make such accurate predictions allows the reintegration of a huge literature on change detection with very distinct foils, with important theoretical and clinical implications[29], as it shows that measuring *d*′ with maximally distinct foils is sufficient to understand memory response distributions—there is no separate concept of 'precision' that is being missed in such tasks.

Generalization across different stimulus spaces. So far, we have focused largely on colour space, which is the dominant way visual working memory is studied[7]. However, TCC is not limited to colour and can be applied to any stimulus space. To demonstrate its generality, we applied TCC to the case of face identity, since it is a complex stimulus space that contains multiple low- and high-level features. Using a previously created face identity continuous report procedure[30], we collected memory data for set sizes one and three. We also measured the psychophysical similarity function and the accuracy of perceptual matching on this face space (Fig. 6). Again, we found that the TCC fit observed memory data extremely well across both set sizes one and three (see Fig. 6) and fit reliably better than existing mixture models (Supplementary Table 3).

Thus, TCC accounts for data across multiple stimulus spaces. As long as the perceptual similarity space of the stimuli is accurately measured using psychophysical scaling (see Supplementary Discussion), TCC's straightforward signal detection account, with only a single *d*′ parameter, accurately captures the data.

Generalization across different memory systems. To demonstrate TCC's applicability to multiple memory systems, not just visual working memory, we fit data from a visual long-term memory continuous report task with colours. Unlike the previous datasets, these data had been previously reported in the literature[31]. Participants performed blocks where they sequentially saw 40 real-world objects that were randomly coloured. Then, after a delay, they reported the colour of the object using a colour wheel (as in Brady et al.[32]). Some items were seen only once and some were repeated twice in the same colour within a block (Fig. 6d). Again, we found that TCC fit the observed memory data extremely well across both the

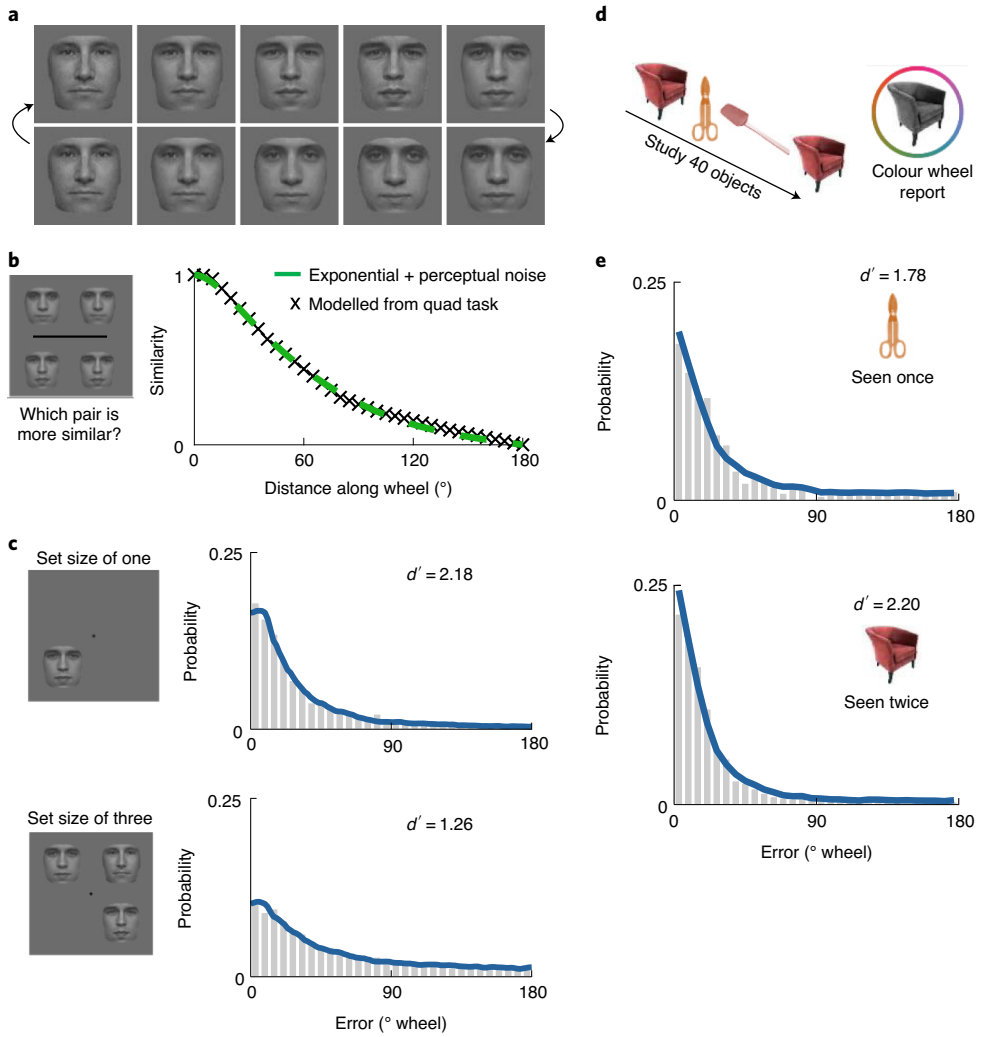


Fig. 6 | TCC generalizes to face identity and visual long-term memory. a, Examples from a previously used continuous face space[30]. **b**, Using a quad similarity task to reduce relational encoding, and the same MLDS method and perceptual matching task as with colour, we collected a psychophysical distance function for face identity (n=102). **c**, TCC fits to working memory data (n=50) using face identity at set sizes of one and three (one: r=0.997; P<0.001; CI=0.993 to 0.998; three: r=0.985; P<0.001; CI=0.971 to 0.992). TCC accurately captures face identity data, demonstrating its generalizability across diverse stimulus spaces. **d**, To show generalization to other memory systems, we fit data on a visual long-term memory continuous report task with colours[31]. Thirty participants performed blocks of memorizing 40 items. Then, after a delay, they reported the colours of the items using a colour wheel. Some items were seen only once and some were repeated twice in the same colour within a block. **e**, TCC fits to visual long-term memory data for items seen only once and for items repeated twice (once: r=0.978; P<0.001; CI=0.958 to 0.988; twice: r=0.991; P<0.001; CI=0.983 to 0.995). TCC accurately captures visual long-term memory data, suggesting that the psychological similarity function is a constraint on both working and long-term memory systems. Note that long-term memory performance in this task probably depends on a two-part decision: item memory and source memory (for example, the object itself and then its colour). This two-part decision is related to the processes of recollection and familiarity and probably introduces heterogeneity in memory strength into the colour memory reports. Here, where item memory was consistently strong and colour memory was the main factor, this did not affect the fits of TCC, but in other data where heterogeneity in the strength of item memory was greater, variability in d' between items would probably need to be accounted for.

unrepeated and repeated items (Fig. 6e). Thus, unlike working memory modelling frameworks, which propose system-specific mechanisms (for example, population coding combined with divisive normalization[10]), TCC naturally fits data from both visual working memory and long-term memory with the same underlying similarity function and signal detection process applicable across both memory systems.

Implications of TCC. *No objective guessing.* One particularly important implication of TCC's fit to the data with just a single parameter is that it implies that there is little-to-no objective guessing in working memory. This provides evidence against a fixed capacity limit where participants only remember around three or four items[1,2], and is consistent with more continuous conceptions of working memory[4]. In particular, while colours far from the target in colour space sometimes win the competition (for example, have maximal familiarity after noise is added), this is not because the target was fundamentally unrepresented or varied hugely in encoded memory strength trial to trial. In a stochastic competition, the strongest representation does not always win. Moreover, the target will be more likely to lose the competition the weaker its representation is. Critically, in TCC, at least as proposed so far, the target is always represented (that is, people's familiarity signals are never unaffected by what they just saw 1 s ago (as in $d'=0$)).

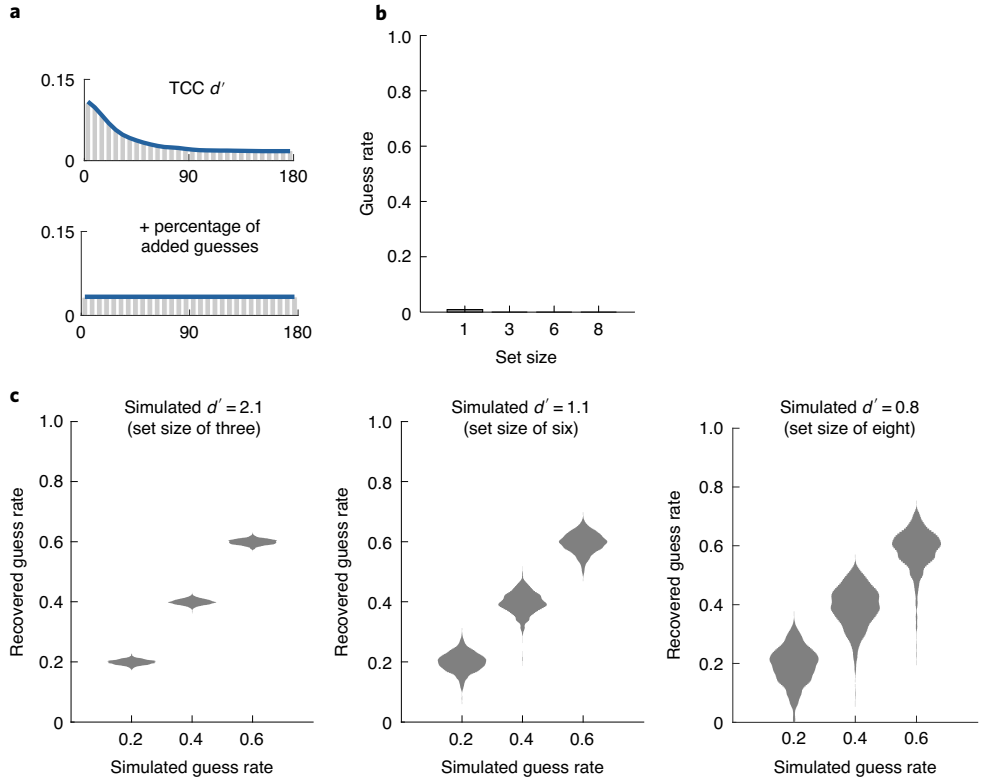


Fig. 7 | TCC measures of 'objective guessing' in visual working memory. a, To validate whether TCC could detect objective guessing (that is, a separate psychological state with no information) if present in the data, we considered a mixture of responses from TCC plus objective guessing, creating a mixture model of TCC and a uniform distribution. **b**, Although model comparison strongly preferred TCC with no guessing, we nevertheless fit a hybrid TCC+guessing model (two parameters) to real participant data. We found that the guessing parameter in real data is estimated at ~0 across all set sizes. **c**, However, when fitting the hybrid TCC+guessing model to simulated data, we observed accurate recovery of guessing if present in the data. Even for 20% of guesses added to a set size of eight, d' levels were accurately recovered and never estimated as 0. Violin plots show the entire distribution of recovered parameters. Furthermore, model comparison metrics—even those, such as BIC, designed to prefer simpler models—preferred the hybrid model with the guessing parameter in every simulation with guessing added (all BIC>30:1 in favour of the hybrid model). This provides strong evidence that there is little objective guessing in visual working memory data and that our modelling with TCC would be able to detect any substantial number of added no-information responses if they were present.

While these conclusions follow from the excellent fits of the straightforward one-parameter TCC model to a wide variety of data (data widely thought to provide prima facie evidence for the existence of unrepresented items) and from the generalization of maximally dissimilar 2-AFC performance to other conditions, to evaluate this claim further, we assessed a two-parameter hybrid model based on TCC but mixed with objective guessing. This hybrid model assumes that only a subset of items are represented and that the remainder have $d'=0$. Focusing on the highest set sizes (six and eight), we found that such a model was dispreferred in model comparisons in 100% of participants compared with TCC (BIC: set size of six: $t(19)=-41.99$; $P<0.001$; $d_z=9.39$; CI = 6.2:1 to 6.9:1; set size of eight: $t(19)=-16.09$; $P<0.001$; $d_z=3.60$; CI = 5.3:1 to 6.9:1), and BIC was well calibrated for these model comparisons (Supplementary Fig. 3). Furthermore, while this hybrid model accurately recovered its own parameters from simulated hybrid data, showing that it detects objective guessing if it is present (Fig. 7c), when fit to empirical data, it estimated guessing rates near 0 in every set size in group data (Fig. 7b) and a guess rate <5% in the majority of individual participants at every set size. Thus, although some items may occasionally have a d' of 0 (perhaps because they were completely unattended during encoding), it appears to happen too infrequently to appreciably affect the fit, and it happens far less often than required for slot models of working memory that suppose that four to five of the eight items are always entirely unrepresented[2]. The simulation results show that it is possible to detect random guesses if they are present in the data, but TCC finds no evidence for such objective guessing in real data. Critically, however, as with any standard signal detection model, TCC naturally accounts for the subjective feeling of guessing/low confidence[21] that arises when memories tend to be weak, such as at high set sizes (Extended Data Figs. 7 and 8).

Mixture models are not measuring distinct psychological states. The dominant quantitative model of visual memory is the mixture model, which claims to measure two distinct psychological concepts from continuous report error data: (1) how precisely people remember items that they have in mind (for example, precision or variability in precision); and (2) how often people have an item in mind (likelihood of retrieval, or its opposite, guess rate). The fundamental claim that there are two distinct ways memory can fail (that is, loss of precision or loss of discrete items) permeates a huge variety of the literature in working memory, attention[33], iconic memory[34] and long-term memory[35]. TCC makes a counterclaim: the fact that manipulations of set size, delay and encoding time that hold the stimulus space constant (for example, use of a particular colour wheel) can be fit by varying a single memory strength parameter, and the fact that measuring how well people

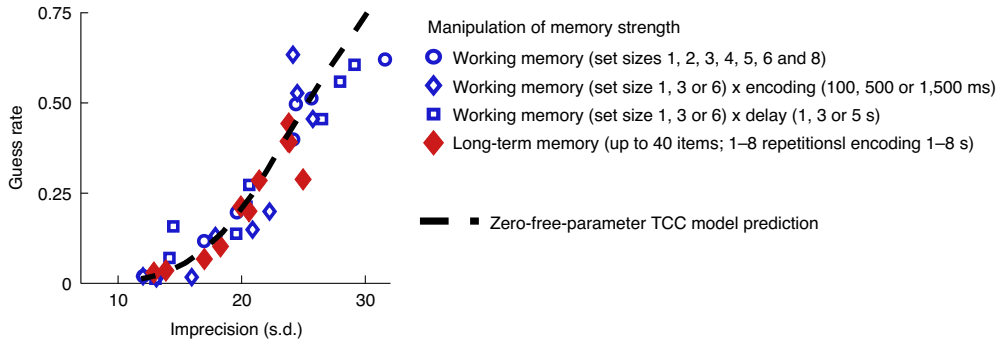


Fig. 8 | State-trace plot of mixture model parameters is in line with TCC prediction. The currently dominant conception of memory arises from mixture models claiming that memory varies in at least two psychologically distinct ways: the precision of memory and the number of represented items (modelled as the guess rate). TCC makes a strong counter prediction: that if the stimulus space, and thus psychophysical similarity function, is held constant, memory report distributions vary in only one way (that is, memory strength). Thus, TCC claims that the particular manipulation (encoding, set size or delay) used to change memory strength should not selectively change one mixture model parameter or another (for example, encoding changing the precision or high set sizes affecting only the guess rate, and so on), but that both should always change together. To visualize this, we show a state–trace plot of mixture model parameters across a wide range of manipulations of working memory (from the current paper) and long-term memory (from Miner et al.[31]), with one point per condition. We find that despite the huge number of different ways we vary memory strength, all of the points lie on a single line, consistent with only a single parameter being varied, and that this line is extremely well predicted by the zero-free-parameter prediction of TCC. TCC can only predict an extremely small part of the possible space that the mixture model can predict, and only a very particular relationship between the two mixture model parameters, and the data from all of these conditions land on this line. This provides strong evidence against mixture models measuring two distinct parameters and in favour of the TCC conception of memory.

can distinguish only maximally distinct comparisons (such as red versus green) is sufficient to characterize memory, appears to falsify the idea that memory changes in two or three psychologically distinct ways (for example, precision versus guess rate). Another way to test this is to fit the mixture model, which purports to measure two distinct parameters, to data from a single stimulus space (for example, from a single colour wheel) and to ask whether the state–trace plot shows evidence of a single way memory changes or multiple ways[36]. Figure 8 shows this plot for all data from the current paper (for example, the 22 conditions shown above, plus the other experiments) and from all of the conditions in Miner et al.[31], which provided the long-term memory data fit above. As can be clearly seen in this plot, the two parameters always change together: while not linear in their relationship, they are nearly perfectly related, and their relationship is well predicted by the zero-free-parameter prediction of TCC (for example, TCC's prediction across a range of d' values). The nonlinear relationship accounts for most cases in which people have found evidence to dissociate the two parameters (see Supplementary Discussion). This is further evidence that TCC's single parameter conception of performance is correct and that mixture models are not measuring distinct psychological constructs (see also Supplementary Figs. 4 and 5 and Supplementary Table 4, which use data from the literature, although without holding the stimulus space constant as here).

Discussion

Most previous theories and models of visual working memory have not considered the relationship between stimuli and the psychological similarity of those stimuli. In the absence of psychophysical scaling and without regard for its theoretical implications, the use of these models has led to what we show are illusory independent estimates of guessing/capacity and precision, and to arguments for limited capacity characterized by so-called discrete failures of working memory, attention[33], iconic memory[34] and long-term memory[35]. Indeed, claims about selective deficits in clinical populations[37–39], and even about the nature of consciousness[33], have been made based on dissociations between model-based estimates of precision and guessing. Here, we have shown that these apparent dissociations are an illusion of modelling the data without taking into account the nonlinear way that familiarity spreads in stimulus space. When this fixed perceptual similarity structure is taken into account, TCC provides a unifying theory of visual memory strength—one that is capable of bridging distinct tasks and stimulus conditions that would not be possible using previous models and that undermines the interpretation of apparent discrete failures of attention and memory[33–35,37–39].

While TCC rejects the idea that the distribution of responses collected from continuous report is explained primarily by items that are remembered or not (or items that are encoded with extremely different precisions[8]), this does not mean that some variability between items is not present in working memory tasks. Psychophysical scaling can naturally account for many stimulus-specific variability effects (for example, some colours being more distinct than others; Extended Data Fig. 5) by using separate similarity functions for each target colour. Furthermore, in light of the signal detection framework of TCC, much of the existing evidence for variable precision does not actually provide direct evidence of variability in the d' parameter of the TCC model. Many aspects of variability between items arise in TCC naturally from the independent noise added to different items that is at the heart of signal detection theory, such as the effect of varying confidence on continuous report data or allowing participants to choose their best item for report (Extended Data Figs. 7 and 8). Thus, it remains an open question to what extent d' varies between items and trials. In TCC, if such variation needs to be accounted for, this would be done by moving to an unequal variance signal detection model, whereas the current modelling has used a purely equal variance model. Critically, however, we show that mixing in items that are unrepresented ($d'=0$) is inconsistent with the data. Thus, any variability in d' that does exist across items probably does not include an appreciable role for items with $d'=0$.

Many models of working memory focus almost exclusively on how memory strength changes with set size, taking this as the central factor in how much understanding of working memory they have achieved. We take a fundamentally different view, seeing our measure of memory strength (d') as a measure of discriminability that is probably modulated by many factors, and which has a shared structure not only in working memory, where set size matters so much, but also in long-term memory, which appears to

follow fundamentally the same rules of memory confusability and a similar decision process (Fig. 8; Miner et al.[31]). Notably, we find that while set size modulates memory strength in the current work, there are many other factors that affect memory strength nearly as much. For example, increased delay decreases d' (more noise accumulates even with the same signal) and increased encoding time improves d' (more signal relative to the same noise). Similarly, in some situations, other factors such as location noise (for example, swaps; Bays et al.[26]) and ensemble coding[40,41] seem to play a major role in memory errors. Thus, while we find an approximately power law-like relationship between set size and d' (Supplementary Fig. 4), we are hesitant to assume that there will be a fixed law for how set size relates to memory errors, and note that previous work that claims to find such rules[7–9] has almost never examined whether those rules hold when manipulating other factors that will also independently impact memory strength, such as encoding time and delay.

In addition, in the current work, we present a straightforward version of the TCC model that does not account for all possible factors. For example, it is possible to make different predictions for different target colours, taking into account category effects (for example, Extended Data Fig. 5). In addition, while in the current data we see almost no swaps or location-based confusions (because we use long encoding times and placeholders), it is of course possible to implement a swap parameter in TCC (as in Williams et al.[42]) or to explicitly model the psychophysical similarity structure of location and therefore make parameter-free location confusion predictions. Similarly, hierarchical models of ensemble coding and grouping, or other forms of integration across items, could potentially be implemented using TCC as the basis of memory responses. If there is substantial integration across items or across time in a particular paradigm, more complex models such as these would be needed because TCC's item-based prediction about error distributions would no longer be a valid assumption.

While TCC is a theory about the fundamental nature of the underlying memory signal in visual working and long-term memory tasks, and about how this signal is used to make decisions, there are many potential cognitive and neural explanations (shared or independent across systems) that may instantiate the model. Indeed, in long-term memory, signal detection models have often been conceptualized in relation to neural measures, including both neuroimaging[43] and single-unit recording[44].

The central feature of TCC is the psychophysical similarity measurement, which provides the basis for the straightforward signal detection model. This similarity function is naturally understood using models of efficient coding[18] or population coding[10]. For example, the idea that far away items in feature space are all approximately equally similar arises naturally from population codes—if individual neurons' tuning functions only represent a small part of colour space (for example, 15° on the colour wheel), there would be extremely limited overlap in the population of neurons that code for any two colours even a medium distance apart on the wheel. There would also be correlated noise between nearby colours, as we assume in TCC.

Thus, the current model is in many ways related to existing models of working memory based on population codes[9,10]. Indeed, the similarities between the framework of population coding and the cognitive model proposed here offer important promise for bridging across levels of understanding in neuroscience, with population coding implementations of TCC possible[45,46]. However, compared with existing population-based models[10], the cognitive basis of the current model—with the measured scaling function following the well-known cognitive laws of similarity[17,19]—allows us to fit data with an extremely simple one-parameter model that allows generalization across tasks and draws strong connections to signal detection theory and long-term memory that are not apparent when thinking about population coding alone without this cognitive basis.

In addition, framing our model in terms of signal detection theory allows a very general model of the decision process, compared with population coding models for which the decision process is based on variability in spikes in a fixed neural population[10], which are difficult to reconcile with data from high-level stimuli such as faces (which are probably encoded in many distinct populations) and data from long-term memory (which are not stored online in a fixed neural population).

Previous work has shown that psychophysical similarity metrics are probably distinct for different stimuli in the same stimulus space (for example, memory varies across colours[12,13]; Extended Data Fig. 5). The underlying space on which the exponential similarity function is imposed may be designed to take advantage of efficient coding of environmental regularities[47] such that the more frequent the stimuli the more neural resources we devote, giving improved discriminability and predictable memory biases[48]. Taking this into account may allow a simple parameterization of not only the average similarity function but the particular functions for individual stimuli (as in Fig. 1d). In addition, psychophysical similarity may not be a fixed property but may be dependent on how the current environment affects discriminability[49,50]. For example, memory biases are altered when discriminability is affected by adaptation or contextual effects[48].

Some previous models of visual working memory have, like TCC, rejected the idea that the fat tails in the error distribution (Fig. 1) arise from unrepresented items[8,9]. For example, models such as the variable precision model[8] hold that items vary in the precision with which they are encoded, and this heterogeneity between items is critical to explaining the shape of the error distribution (that is, extremely poorly represented items, rather than completely unrepresented items, explain the tail of the error distribution). As in TCC, this model holds that there is not in fact a completely uniform, flat tail in the distribution, and assumes that items vary in representational fidelity (like the independent noise for different items in TCC).

However, in other ways, the two models differ substantially. The variable precision models, like other previous memory models, rely on the assumption that the response axis can usefully be thought of as linear. In contrast, we have shown that similarity and memory confusability are deeply nonlinear along this axis, in agreement with decades of work suggesting that psychological similarity is globally exponential (for example, the universal law of generalization[17,18]). This results in critical differences between the variable precision model and TCC. In particular, in the variable precision model, the latent distribution of precisions is an unknown that is taken to vary between situations, whereas TCC uses the insight that similarity is nonlinear and relatively fixed to greatly simplify the model of the error distribution (allowing, for example, the generalizations from 180° 2-AFC that are not possible in the variable precision model).

Finally, TCC provides a compelling connection between working memory and long-term recognition memory, which is often conceptualized in a signal detection framework. In particular, it can be naturally adapted to explain a number of findings that are in common between the working memory and long-term memory literatures but have been difficult to explain with previous working memory models, such as the relationship between confidence and accuracy[51,52] (Extended Data Figs. 7 and 8) and the ability of participants to respond correctly when given a second chance, even if their first response was a guess or low-precision response[53]. Thus, despite research on working and long-term memory operating largely independent of one another, TCC provides a unified framework for investigating the distinctions and similarities in memory across both domains by emphasizing that competition and perceptual confusability between items is a major limiting factor across both working memory and long-term memory.

Methods

All of the conducted studies were approved by the Institutional Review Board at the University of California, San Diego, and all participants gave informed consent before beginning the experiment. All colour experiments used a circle in CIE $L^*a^*b^*$ colour space, centred in the colour space at $L=54$, $a=21.5$ and $b=11.5$, with a radius of 49. For online experiments, this was converted to RGB using an assumed equal energy ("E") whitepoint. All sample sizes were decided a priori, and were similar to those in previous publications[7–9,32]. Approximately half of the data were generated by experiments run in the laboratory, with the others conducted using Amazon Mechanical Turk. Mechanical Turk users form a representative subset of adults in the United States[54], and data from Mechanical Turk are known to closely match data from the laboratory on visual cognition tasks[40,55], including providing extremely reliable and high agreement on colour report data[41]. Any systematic differences between the laboratory studies (in which we collected most of the memory data) and the Mechanical Turk studies (in which we collected most of the similarity data) would decrease the appropriateness of the similarity function for fitting the memory data, hurting the fit of TCC. Data collection and analysis were performed with knowledge of the conditions of the experiments. All statistical tests were two tailed.

Fixed-distance triad experiment. A total of 40 participants on Mechanical Turk judged which of two colours presented was more similar to a target colour. The target colour was chosen randomly from 360 colour values that were evenly distributed along a circle in the CIE $L^*a^*b^*$ colour space, as described above. The pairs of colours were chosen to be 30° apart from one another, with the offset of the closest colour to the target being chosen with an offset of either 0, 5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 120 or 150° (for example, in the 150° offset condition, the two choice colours were 150 and 180° away from the target colour; in the 0° offset condition, one choice exactly matched the target and the other was 30° away).

Participants were asked to make their judgements solely based on intuitive visual similarity and to repeat the word 'the' for the duration of the trial to minimize the use of verbal strategies. Each participant completed 130 trials, including ten repeats of each of the 13 offset conditions, each with a different distance to the closest choice colour to the target, and trials were conducted in a random order. The trials were not speeded, and the colours remained visible until participants chose an option. To be conservative about the inclusion of participants, we excluded any participant who made an incorrect response in any of the ten trials where the target colour exactly matched one of the choice colours, leading to exclusion of seven of the 40 participants, and based on our a priori exclusion rule, we excluded any participants whose overall accuracy was two standard deviations below the mean, leading to no additional exclusions. In addition, based on an a priori exclusion rule, we excluded trials with reaction times of <200 or >5,000 ms, which accounted for 1.75% (s.e.m. = 0.5%) of the trials. The data from a subset of offset conditions are plotted in Fig. 1c, and the full dataset is plot in Extended Data Fig. 1.

Psychophysical scaling triad experiment. A total of 100 participants on Mechanical Turk judged which of two presented colours was more similar to a target colour, as in the fixed-distance triad experiment. However, the pairs of colours now varied in offset from each other and from the target, to allow us to accurately estimate the entire psychophysical distance function. In particular, the closest choice item to the target colour could be one of 21 distances away from the target colour: 0, 3, 5, 8, 10, 13, 15, 20, 25, 30, 35, 45, 55, 65, 75, 85, 100, 120, 140, 160 or 180°. If we refer to these offsets as o_i, such that o_1 is 0° offset and o_{21} is 180° offset, then given a first-choice item of o_i, the second-choice item was equally often o_{i+1}, o_{i+2}, o_{i+3}, o_{i+4} or o_{i+8} degrees from the target colour (excluding cases for which such options were >21).

Participants were asked to make their judgements solely based on intuitive visual similarity, and to repeat the word 'the' for the duration of the trial to prevent the usage of words or other verbal information. Each participant completed 261 trials, including three repeats of each of the possible pairs of offset conditions, and the trials were conducted in a random order. The trials were not speeded, and the colours remained visible until participants chose an option. Following our a priori exclusion rule, we excluded any participant whose overall accuracy was two standard deviations below the mean ($M=77.5\%$), leading to the exclusion of eight of the 100 participants. In addition, based on an a priori exclusion rule, we excluded trials with reaction times <200 or >5,000 ms, which accounted for 1.7% (s.e.m. = 0.26%) of trials.

To compute psychophysical similarity from these data, we used a modified version of the model proposed by Maloney and Yang[16]: the MLDS method. Rather than using this model to measure the distance between, for example, red and green, we adapted it to measure the appropriate psychophysical scaling of similarity between colours as a function of their distance between colours along the wheel rather than their absolute colour. In particular, any given trial had a target colour, S_i, plus two options in answer to which is more similar, S_j and S_k. Let $l_{ij}=S_j-S_i$ (the distance between S_i and S_j on the colour wheel, which is always in the set [0,3,5,…180]), and let ψ_{ij} represent the psychophysical similarity to which this distance corresponds. If people made decisions without noise, they should pick item j if and only if $\psi_{ij}>\psi_{ik}$. We added noise by assuming participants' decisions were affected by Gaussian error, such that they picked item j if $\psi_{ij}+\varepsilon>\psi_{ik}$. We set the standard deviation of the Gaussian ε noise to 1, consistent with a signal detection model. Thus, the model had 20 free parameters, corresponding to the similarity scaling values for each possible distance length (for example, how similar a distance of 5 or 10° on the colour wheel really was to participants), and then we fit the model using maximum likelihood search (fmincon in MATLAB). Thus, these scaled values for each interval length most accurately predicted observers' similarity judgements, in that equal intervals in the scaled space were discriminated with equal performance. Once the scaling was estimated, we normalized the psychophysical scaling parameters so that psychophysical similarity ranged from 0 to 1.

We did not test all of the possible pairings, but simply a subset (five different offsets), because collecting more pairs would not have improved the estimate of the psychophysical scaling function much, if at all, since the pairs we used overlapped enough without using all of them. Each possible pairing provided an estimate of a slope on the psychophysical similarity graph. For each pair, the relevant part of the x axis was known, and people's d' at discriminating each pair (that is, determining which was closer between the target + 10° and the target + 45°) was an estimate of the y axis difference/slope in that range (that is, the difference in psychophysical similarity between those two points). Having 21 (distances) × 5 (offsets from those distances) = 105 such slope estimates, with some covering wide ranges of the x axis and some covering small ranges, and each well estimated, was sufficient to constrain the global shape of the function when using the MLDS method.

Likert colour similarity experiment. A total of 50 participants on Mechanical Turk judged the similarity of two colours presented simultaneously on a Likert scale, ranging from 1 (least similar) to 7 (most similar). The colours were chosen from a wheel consisting of 360 colour values that were evenly distributed along the response circle in the CIE $L^*a^*b^*$ colour space. The pairs of colours were chosen by first generating a random start colour from the wheel and then choosing an offset to the second colour from the set 0, 5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 120, 150 or 180°. Participants were given instructions by showing them two examples. In example 1, the two colours were identical (0° apart on the colour wheel) and participants were told that they should give trials such as this a 7. In example 2, the two colours were maximally dissimilar (180° apart on the colour wheel) and participants were told that they should give this trial a 1. No information was given about how to treat intermediate trials. Participants were asked to make their judgements solely based on intuitive visual similarity, and to repeat the word 'the' for the duration of the trial to prevent the usage of words or other verbal information. Each participant took part in 140 trials, including ten repeats of each of the 14 offset conditions, each with a different starting colour, and trials were conducted in a random order. The trials were not speeded, and the colours remained visible until participants chose an option. Two participants were excluded for failing a manipulation check (requiring a similarity of >6 for trials in which the colours were identical). Based on an a priori exclusion rule, we excluded trials with reaction times <200 or >5,000 ms, which accounted for 3.0% (s.e.m. = 0.4%) of trials. The similarity between two colours separated by x° was measured using a seven-point Likert scale, where $S_{min}=1$ and $S_{max}=7$. To generate the psychophysical similarity function, we simply normalized these data to range from 0 to 1, giving a psychophysical similarity metric, such that $f(x)=((S_x-S_{min})/(S_{max}-S_{min}))$.

Perceptual matching experiment. A total of 40 participants on Mechanical Turk were shown a colour and had to match this colour, either using a continuous report colour wheel (100 trials) or choosing among 60 options (100 trials; spaced 6° apart on the colour wheel, always including the target colour). The 60-AFC version was designed to limit the contribution of motor noise, since the colours in this condition were spaced apart and presented as discrete boxes that could not easily be misclicked. Colours were generated using the same colour wheel as the other experiments, and participants were given unlimited time in which to choose the matching colour. The colour and colour wheel/response options remained continuously visible until participants clicked to lock in their answer. The colour was presented at one of four locations centred around fixation (randomly), approximately matching the distance to the colour wheel and variation in the position used in the continuous report memory experiments. One participant's data were lost due to experimenter error and two participants were excluded for an average error rate greater than two standard deviations away from the mean.

To convert these data into a perceptual correlation matrix, which asked how likely participants were to confuse a colour x degrees away in a perception experiment, we relied on the 60-AFC data alone, since these data received no contribution from motor noise and so were solely a measure of perceptual noise. However, using the continuous report data instead resulted in no difference in any subsequent conclusions, as the contribution of motor noise in that task appeared to be minimal. To create the perceptual correlation matrix, we created a normalized histogram across all participants of how often they made errors of each size up to 60° errors (−60, −54… 0, … 54, 60), and then linearly interpolated between these to obtain a value of the confusability for each degree of distance. We then normalized this to range from 0 to 1.

Modelling data using the TCC model. The TCC model is explained interactively at https://bradylab.ucsd.edu/tcc/. In general, the model is typical of a signal detection model of long-term memory, but adapted to the case of continuous report, which we treat as a 360-AFC for the purposes of the model. The analysis of such data focuses on the distribution of errors people make, measured in degrees along the response wheel, x, where correct responses have $x=0°$ error and errors range up to $x=\pm180°$, reflecting the incorrect choice of the most distant item from the target on the response wheel (Fig. 1b). In the TCC model, when probed on a single item and asked to report its colour: (1) each of the colours on the colour wheel generates a memory-match signal m_x, with the strength of this signal drawn from a Gaussian distribution, $m_x \sim N(d_x, 1)$; (2) participants report whichever colour x has the maximum m_x; (3) the mean of the memory-match signal for each colour, d_x, is determined by its psychophysical similarity to the target according to the measured function ($f(x)$), such that $d_x = d'f(x)$ (Fig. 2); and (4) the noise is correlated across nearby colours according to confusability in a perceptual matching task. To obtain the psychophysical similarity function, $f(x)$, we use the smooth function estimated from the Likert similarity experiment, although the triad task-modelled similarity function predicts fundamentally the same results (Extended Data Fig. 4).

According to the model, the mean memory-match signal for a given colour x on the working memory task is given by $d_x = d'f(x)$, where d' is the model's only free parameter. When $x=0$, $f(x)=1$, so $d_0=d'$. By contrast, when $x=180$, $f(x)=0$, so $d_{180}=0$. Then, as noted above, at test, each colour on the response wheel generates a memory-match signal, m_x, conceptualized as a random draw from that colour's distribution centred on d_x. That is, if the noise is uncorrelated between nearby colours, $m_x \sim N(d_x, 1)$. The response (r) on a given trial is made to the colour on the wheel that generates the maximum memory-match signal, $r=\text{argmax}(m)$.

Thus, the full code for sampling an absolute value of error from such a TCC-like (uncorrelated noise) model is only two lines of MATLAB:

```
memoryMatchStrengths = randn(1, 180) + similarityFunction * dprime;
[~, memoryError] = max(memoryMatchStrengths);
```

This model fits the data well as it is (see Extended Data Fig. 2), but as specified so far, this model assumes that 360 independent colour probes elicit independent noisy memory-match signals. The shapes of the distributions the model predicts are effectively independent of how many colour channels we assume, so this number is not important to TCC's ability to fit working memory data, but the d' value in the model does change depending on the number of colour channels used. Thus, to make the d' value in TCC comparable to real signal detection d' values, it is important to consider how many colour channels people are accessing.

Rather than make this a discrete decision (for example, there are 30 independent colours on the colour wheel, so people consider 30 channels), we instead estimated the covariance between nearby channels in a continuous manner. The familiarity value of colour 181 and 182 on the wheel cannot possibly be fully independent, since these two colours are perceptually indistinguishable. Following this intuition, we made a simple assumption: the amount of shared variance in the noise between any two colour channels is simply how often colours at that distance are confused in a perceptual matching task. Thus, $p(x)$, the correlation in the noise between any two colours x apart on the colour wheel, is given by C_x/C_0, where C_x is how often colours x degrees away from the target are chosen in the perceptual matching task (with these values interpolated from the histogram of errors; see Methods section 'Perceptual matching experiment'). On average, colours 1° away were chosen about 96% as often as the correct colour in the matching task, so the noise between any two channels 1° apart was assumed to share 96% of its variance; with 82% at 5°, and so on. Thus, having measured both the similarity function and the perceptual matching matrix, to sample from the full (correlated noise) TCC model, we used MATLAB code that was nearly as straightforward as in the uncorrelated model:

```
memoryMatchStrengths = mvnrnd(similarityFunction *
dprime, percepCorrMatrix);
[~, memoryError] = max(memoryMatchStrengths);
```

Thus, in the full version of TCC, the mean of the memory-match signal for each colour, d_x, is determined by its psychophysical similarity to the target according to the measured function $f(x)$, which is taken to be symmetrical for the fitting based on the averaged similarity data, such that $d_x = d'f(|x|)$, for x values $[-179,180]$. The covariance between colours (R) is given by the perceptual confusability of colours at that distance, $p(x)$, which is also taken to be symmetrical:

$$R = \begin{pmatrix} p(0) & p(1) & p(2) & \dots & p(180) & p(179) & \dots & p(2) & p(1) \\ p(1) & p(0) & p(1) & \dots & p(179) & p(180) & \dots & p(3) & p(2) \\ p(2) & p(1) & p(0) & \dots & p(178) & p(179) & \dots & p(4) & p(3) \\ \dots & & & & & & & & \dots \\ p(180) & p(179) & p(178) & \dots & p(0) & p(1) & \dots & p(178) & p(179) \\ p(179) & p(180) & p(179) & \cdots & p(1) & p(0) & \dots & p(177) & p(178) \\ \dots & & & & & & & & \dots \\ p(1) & p(2) & p(3) & \dots & p(179) & p(178) & \dots & p(1) & p(0) \end{pmatrix}$$

To use the perceptual correlation data as the covariance in the correlated model, because there might not always be a perfect correlation matrix (for example, it might not be perfectly symmetrical, as it was based on real data), we first computed R and then iteratively removed negative eigenvalues from this matrix and forced it to be symmetrical until it was a valid correlation matrix. This resulted in only minimal changes compared with the raw perceptual correlations inferred from the perceptual confusability data.

Then, let $(X_{-179},\dots, X_{180})$ be a multivariate normal random vector with mean d, unit variance and correlation matrix R. The winning memory strength (m; that is, subjective confidence) and reported colour value, r, are then the max and argmax, respectively, of this vector:

$$m = \max(X_{-179}, \dots, X_{180})$$
$$r = \text{argmax}(X_{-179}, \dots, X_{180})$$

and the error, e, is the circular distance from r to 0. The distribution of m is in theory directly computable[56], but we rely on sampling from this distribution for the fits in the current paper (see below).

Although also not important to the fit of the current data, the model can also be adapted to include a motor error component. Whereas existing mixture models predict the shape of the response distribution directly and thus confound motor error with the standard deviation of memory (see Fougnie et al.[57] for an attempt to de-confound these), our model makes predictions about the actual item that participants wish to report. Thus, if participants do not perfectly pick the exact location of their intended response on a continuous wheel during every trial, a small degree of Gaussian motor error can be assumed to be included in responses; for example, the response on a given trial, rather than being $\text{argmax}(X_{-179},\dots, X_{180})$, probably includes motor noise of some small amount (for example, 2°):

$$r \sim N(\text{argmax}(X_{-179}, \dots, X_{180}), 2°)$$

Thus, for accuracy to the real generative model of responses, in the model fitting reported in the present paper, we included a fixed normally distributed motor error with s.d. = 2°, although we found that the results were not importantly different if we did not include this in the model.

For fits using the uncorrelated noise model, fits of the d' parameter of the model to datasets were performed using the MemToolbox[58], making use of maximum likelihood (see code on OSF). For fits of the correlated model, which it is difficult to compute a likelihood function for but straightforward to sample from, we relied on sampling 500,000 samples from the model's error at each of a range of d' values (0 to 4.5 in steps of 0.02) and slightly smoothing the result to obtain a PDF for the model at each d' value. The uncorrelated noise version of TCC, which can be directly maximized, results in the same fits as the correlated version, with d' linearly scaled by ~0.65 (see Extended Data Fig. 2). Thus, it is also possible to fit the correlated noise version by fitting the uncorrelated version through maximum likelihood with the appropriate adjustment to d', and doing so results in the same fits.

Continuous colour report data as a function of set size. The continuous colour report data used for fitting the model were collected in the laboratory, to allow a larger number of trials per participant. A total of 20 participants performed 400 trials of a memory experiment, with 100 trials at set sizes of one, three, six and eight (plus four practice trials). The display consisted of eight placeholder circles. Colours were then presented for 1,000 ms, followed by an 800-ms inter-stimulus interval (ISI). For set sizes below eight, the colours appeared at random locations with placeholders in place for any remaining locations (for example, at set size three, the colours appeared at three random locations with placeholders remaining in the other five locations). Colours were constrained to be at least 15° apart in colour space along the response wheel. After the ISI, a target item was probed by marking a placeholder circle with a thicker outline, and participants were asked to respond on a continuous colour wheel to indicate what colour had been presented at that location. The response wheel was held constant from trial to trial. Error was calculated as the number of degrees on the colour wheel between the probed item and the response. No participants were excluded.

Continuous report memory as a function of delay. A total of 20 participants in the laboratory completed a colour working memory task similar to the previous set-size experiment, but with the following exceptions. Participants performed 12 blocks of 75 trials (900 trials in total). Each block contained an equal number of trials at set sizes of one, three and six. The display consisted of six placeholder circles. Colours were presented for 500 ms and followed by a delay of either 1,000, 3,000 or 5,000 ms. The delay time was blocked, and participants were informed of the delay time for that block at the beginning of each block. Each combination of the three set sizes and three delays was used in 100 trials. One participant was excluded for having performance greater than two standard deviations worse than average (across all conditions), leaving a final sample of 19.

Continuous report memory as a function of encoding time. A total of 20 participants in the laboratory completed a colour working memory task identical to the delay experiment, but with the following exceptions. Participants performed

12 blocks of 75 trials. Each block contained an equal number of trials at set sizes of one, three and six. Colours were presented for either 100, 500 or 1,500 ms. The encoding time was blocked, and participants were informed of the encoding time for that block at the beginning of each block. Following encoding, there was a 1,000-ms delay before a target item was probed. Each combination of the three set sizes and three encoding times was used in 100 trials. No participants were excluded.

Model comparisons with mixture models. For all model comparisons, we created new versions of mixture models designed to be directly comparable with TCC. In particular, to make predictions derived from mixture models comparable with those derived from TCC (which specifies a probability of response discretely for each 1° of the wheel, not over a continuous distribution), we used discrete versions of the two- and three-parameter mixture models in which the probabilities of the data were normalized over each of 360 possible integer error values (not over the continuous space of errors).

We performed two types of model comparison: one to simply assess the fit of the model to the data; and one designed to penalize more complex models. In particular, we first performed a cross-validation procedure to assess the fit of each model[59]. Specifically, we fit the TCC and the two-parameter and variable precision mixture models to data from each set of $n-1$ trials separately for each participant and set size, and then evaluated the log-likelihood of this model using data from the single held-out trial. We then assessed the reliability of this likelihood difference across participants separately for each set size. TCC and mixture models provided relatively comparable fits (see Supplementary Table 2), which was to be expected given that the mixture model can almost perfectly accurately mimic TCC (see Supplementary Fig. 3), and given that the amount of data used to fit the models was much greater than the number of parameters in either model (which ranged from one to three), so cross-validation provided effectively no penalty for complexity.

We then compared how well the competing models (TCC, the two-parameter mixture model and the three-parameter variable precision mixture model) fit data from individual participants for the colour report data when using an explicit penalty for the greater complexity of the mixture models. In particular, we assessed BIC separately for each set size and each participant. We found a strong preference for TCC over both mixture models when penalizing complexity (see Supplementary Table 2). Note that this was true even though TCC fits were based on aggregated similarity functions from a different group of participants, collected in a different way (online versus in the laboratory), suggesting that the global structure of the psychophysical similarity function is largely a fixed aspect of a given stimulus space. Ideally, for model comparison purposes TCC would be fit with a similarity function specific to each individual target colour (which can be done and predicts the appropriate deviations; see Extended Data Fig. 5), which would almost certainly improve the fit of TCC even further with no added parameters (because the added complexity would simply be more measured perceptual data). However, in the current fits, we relied solely on averaged similarity to demonstrate how it is the global, not local, structure of the similarity space that is critical to the fit of TCC.

2-AFC at different foil similarities. A total of 60 participants on Mechanical Turk completed 200 trials of a four-item working memory task. On each trial, they saw four colours randomly chosen from the colour wheel (subject to the constraint that no two colours were within 15° of each other). The colours were presented for 1,000 ms; then, after an 800-ms delay, each participant had to answer a 2-AFC memory probe about one of the colours. The foil colour in the 2-AFC could be offset from the target 180, 72, 24 or 12° (50 trials per condition). These conditions were interleaved so that participants needed to maintain detailed memories of the colour on every trial, since conceivably if only 180° foils were present for a block or in an entire experiment, participants would be likely to encode only categorical, not perceptual, information. The response options were presented at appropriate locations along a full colour wheel; for example, the 180° foils were presented 180° apart on the screen and the 12° foils were presented 12° apart on the screen, to visually indicate the distance between the target and foil in colour space. The response wheel was rotated from trial to trial.

Performance was scored as the number correct out of 50 at each offset of the memory foil. Five participants were excluded for below-chance performance in the maximally easy 180° offset condition, leaving $n=55$ participants.

To assess the predictions of TCC for these data in a way amenable to the use of Bayes factors, we took the number correct out of 50 in the 180° foil condition and used this to calculate a probability distribution over d' values (for example, any given d' predicts, according to the binomial function, a likelihood over all numbers of correct responses). In TCC, a given d' value for 180° foils predicts d' for all other offsets straightforwardly, although for the correlated noise TCC, performance is not simply d' modulated by similarity (for similar foils, the correlated noise plays a role). Thus, to predict performance, we sampled from the model repeatedly; for example, for 24° foils, in MATLAB notation:

$$\text{memoryMatchStrengths} = \text{mvnrnd}(\text{similarityFunction} * \text{dprime}_{180}, \text{percepCorrMatrix}, 50);$$
$$\text{isCorrect} = \text{memoryMatchStrengths}_{0\text{deg}} > \text{memoryMatchStrengths}_{24\text{deg}}$$

In other words, to assess performance in the 24° offset condition, we assumed that responses were generated according to the argmax of only these two values:

$$r = \text{argmax}(X_0, X_{24})$$

To preserve all uncertainty, we marginalized over the distribution of d' values implied by the number of correct trials in the 180° foil case and used this to make a prediction about the distributions of correct answers expected for each of the other offset conditions. This allows us to understand the likelihood of each participant's performance in the other conditions given their 180° foil performance in TCC.

To assess the likelihood of performance at different offsets in the mixture model framework of Zhang and Luck[7], we used performance at the 180° foil conditions to assess the guess rate of participants (guess rate $= 1-(2\times\text{percent correct}_{180}-1)$) in the standard way (for example, Brady et al.[60]). However, in this framework, 180° foils leave an unknown free parameter: memory precision cannot be assessed using such foils, and thus is free to vary. Thus, to predict the likelihood of each performance level at each other foil offset, we needed to marginalize over the unknown precision parameter. To minimize assumptions about this, we used the same prior on precisions that van den Berg et al.[8] used when fitting both the standard mixture model and their own variable precision model—a uniform prior over the concentration parameter of the von Mises from 0–200. For any given guess rate and precision, we then calculated the percentage of the PDF that was closest to each 2-AFC response option at each offset to generate a likelihood for the data (as in MemToolbox[58]). To calculate Bayes factors, we used a grid of values for both the d' in TCC and for the precision in the mixture model, using steps of 1 in the precision and steps of 0.01 in d', and we assessed the summed log-likelihood of each of the three other offsets (for example, not including the 180° condition) as our final data likelihood.

2-AFC generalization to *n*-AFC and continuous report. A total of 60 participants on Mechanical Turk completed 200 trials of a four-item working memory task. On each trial, they saw four colours randomly chosen from the colour wheel (subject to the constraint that no two colours were within 15° of each other). The colours were presented for 1,000 ms; then, after an 800-ms delay, each participant had to answer a probe about one of the colours. This probe could be a 2-AFC (with 180° different foil), an 8-AFC (with the choices equally spaced around the colour wheel, and always including the target), a 60-AFC (similarly equally spaced) or continuous report (360-AFC). These conditions were interleaved so that participants needed to maintain detailed memories of the colour on every trial, since conceivably if only 180° foils were present for a block or in an entire experiment, participants would be likely to encode only categorical, not perceptual, information. The response options were presented at appropriate locations along a full colour wheel; for example, the 2-AFC foils were presented 180° apart on the screen and the 60-AFC foils were presented 6° apart on the screen, to visually indicate the distance between the target and foils in colour space. The response wheel was rotated from trial to trial.

Performance was scored as the number correct out of 50 at each offset of the memory foil. One participant's data were lost, and seven participants were excluded for below-chance performance in the maximally easy 2-AFC 180° offset condition, leaving $n=52$ participants.

The simplest metric is simply to compare the d' computed from 2-AFC performance (where p is percent correct, $(d' = \frac{\phi^{-1}(p)-\phi^{-1}(1-p)}{\sqrt{2}})$ with the d' from fitting TCC to the continuous report data. These are extremely strongly related (Fig. 5b).

To assess the predictions of TCC for these data in a way amenable to the use of Bayes factors, we again took the number correct out of 50 in the 2-AFC 180° foil condition and used this to calculate a distribution over d' values (for example, any given d' predicts, according to the binomial function, a likelihood over all numbers of correct responses). In TCC, a given d' value for 180° foils predicts d' for all other n-AFCs (including 360-AFC) straightforwardly, by simply first choosing the maximum out of the relevant foil options that are present; for example, at 8-AFC:

$$r = \text{argmax}(\ldots, X_{-45}, X_0, X_{45}, \ldots)$$

To preserve all uncertainty, we marginalized over the distribution of d' values implied by the number correct in the 180° foil case and used this to make a prediction about the distributions of responses to each foil expected for each of the other n-AFC conditions. This allowed us to understand the likelihood of each participant's performance in the other conditions given their 180° foil performance in TCC.

To assess the likelihood of performance in continuous report given performance in the 2-AFC task, in the mixture model framework of Zhang and Luck[7], we used performance at the 180° foil conditions to assess the guess rate of participants (guess rate $= 1-(2\times\text{percent correct}_{180}-1)$) in the standard way (for example, Brady et al.[60]). However, in this framework, 180° foils again left an unknown free parameter: memory precision cannot be assessed using such foils, and thus is free to vary. Thus, to predict the likelihood of each performance level at each other foil offset, we needed to marginalize over the unknown precision parameter. To minimize assumptions about this, we used the same prior on precisions that van den Berg et al.[8] used when fitting both the standard

mixture model and their own variable precision model—a uniform prior over the concentration parameter of the von Mises from 0–200. For any given guess rate and precision, we then calculated the likelihood of participants' continuous report performance under these parameters. To calculate Bayes factors, we used a grid of values for both the d' in TCC and the precision in the mixture model, using steps of 1 for precision and steps of 0.01 for d'. We assessed the log-likelihood of TCC and the mixture model only in the continuous report case, having fit the parameter(s) using only the data from the 2-AFC 180° condition.

Face identity continuous report data. We utilized the same continuous report task, but adapted the stimulus space to face identity using the continuous face identity space and continuous response wheel created by Haberman et al.[30]. In particular, as described in that work, the faces were 360 linearly interpolated identity morphs, taken from the Harvard Face Database, of three distinct male faces (A–B–C–A; see Fig. 6), generated using MorphAge software (version 4.1.3; Creaceed). Face morphs were nominally separated from one another in identity units, which corresponded to steps in the morph space. Before morphing, face images were luminance normalized. In our memory task, we used set sizes of one and three, showing either one or three faces at once, and the encoding display was shown for 1.5 s due to the increased complexity of the face stimuli and task difficulty. Participants on Mechanical Turk ($n=50$) completed 180 trials. The first 20 trials were practice trials and not included in the analysis. A total of 14 participants were excluded for having near-chance performance levels ($d'<0.50$) at a set size of three, although including all participants with $d'\geq 0$ did not affect our conclusions or the fit of TCC.

Face identity similarity quad task. A total of 102 participants on Mechanical Turk judged which of two pairs of faces presented were more distinct (which pair had constituent items that were more different from each other). On each trial, we chose two pairs of faces, with the first item in each pair being randomly chosen and the second item in each pair always having an offset of 0, 5, 10, 20, 40, 60, 80, 100, 140 or 180° away. Altogether, they completed 18 trials of each kind, giving a total of 180 trials each.

Participants were asked to make their judgements solely based on intuitive visual similarity, rather than the use of knowledge of faces or using verbal labels. We excluded participants whose overall performance level was more than two standard deviations below the mean, resulting in a final sample of $n=85$.

To compute psychophysical distance from these data, we used a similar model as for colours, based on the model proposed by Maloney and Yang[16] (that is, the MLDS method). In particular, any given trial had two pairs of faces, where their face wheel values were S_i, S_j and S_k, S_l. Let $l_{ij}=S_j-S_i$ (the length of the physical interval between S_i and S_j, which is always in the set [0,5,10…180]) and let ψ_{ij} represent the psychophysical similarity to which this distance corresponds. If people made decisions without noise, they should have picked pair i,j if, and only if, $\psi_{ij}>\psi_{kl}$. We added noise by assuming that participants' decisions were affected by Gaussian error, such that they picked pair i,j if $\psi_{ij}+\varepsilon>\psi_{kl}$. We set the standard deviation of the Gaussian ε noise to 1, so that the model had nine free parameters, corresponding to the psychophysical scaling values for each possible interval length (for example, how similar a distance of 5 or 10° really was to participants), and then we fit the model using the maximum likelihood search (fmincon in MATLAB). Thus, these scaled values for each interval length most accurately predicted observers' judgements in that equal intervals in the scaled space were discriminated with equal performance. Once the scaling was estimated, we normalized the psychophysical scaling parameters so that psychophysical similarity ranged from 0 to 1.

Face identity perceptual matching. A total of 40 participants on Mechanical Turk were shown a face and were asked to match this face using a continuous report wheel (100 trials). Because the contribution of motor noise appeared to be minimal in the colour matching task (relative to perceptual error) and because showing 60 faces simultaneously would be challenging, we used only a continuous report wheel (no 60-AFC). Faces were generated from the same continuous face space used in the other experiments and participants had unlimited time to choose the matching face. The face and face wheel/response options remained continuously visible until participants clicked to lock in their answer. The face was presented at one of four locations centred around fixation (randomly), approximately matching the distance to the face wheel and variation in position used in the continuous report memory experiments. Seven participants were excluded for below-chance error rates.

To convert these data into a perceptual correlation matrix (asking how likely participants were to confuse a face x degrees away in a perception experiment), we created a normalized histogram across all participants of how often they made errors of each size (in bins of 5°: −180, −175, … 180) and then linearly interpolated between these to obtain a value of the confusability for each degree of distance. We then normalized this to range from 0 to 1.

Visual long-term memory colour report task. Long-term memory data from Fig. 6 were taken from Miner et al.[31] (experiment 2a). A total of 30 participants in the laboratory at the University of California, San Diego performed five blocks of a long-term memory experiment. In each block, they memorized real-world objects' colours; then, after a brief delay, they were shown a sequence of memory tests. Each block's study session consisted of 20 items of distinct categories seen only once and ten items also of distinct categories seen twice, for a total of 40 presentations of coloured objects. Each presentation lasted 3 s, followed by a 1-s inter-stimulus interval. During the test, 20 old objects were presented (ten seen once and ten seen twice) and 20 new objects of distinct categories were presented. Participants saw each object in grayscale and made an old/new judgement; then, if they reported that the item was old, they reported its colour using a continuous colour wheel. As described by Miner et al.[31], six participants were excluded per the criterion used in that paper.

Note that long-term memory performance in this task probably depends on a two-part decision–item memory and source memory (for example, the object itself and then its colour). This two-part decision is related to the processes of recollection and familiarity that can be modelled in various ways[61], and probably introduces notable heterogeneity into the colour memory strength, since some items will have weak item memories, preventing the retrieval of colour information. TCC provides a strong fit here, and to the other long-term memory data plotted in Fig. 8, without addressing this, probably due to the fact that item memory in all of these studies was very strong (only a small number of categorically distinct items needed to be remembered). Future research should clarify how TCC connects to distinctions between recollection and familiarity and the extent to which heterogeneity in d' between items in long-term memory must be assumed for fitting a wider variety of tasks.

Literature analysis. To assess our model's prediction that previously observed trade-offs between different psychological states are measuring the same underlying parameter (d'), we conducted a literature analysis of data from colour working memory research. In particular, we examined the two parameters most commonly reported by those fitting mixture models to their data: precision (in terms of s.d.) and guessing.

We searched for papers in mid-2018 that used these mixture model techniques by finding papers that cited the most prominent mixture modelling toolboxes: Suchow et al.[58] and Bays et al.[26]. We used liberal inclusion criteria to obtain as many data points as possible. Our inclusion criteria were papers that cited either of these toolboxes and reported data where: (1) there was some delay between the working memory study array and test; (2) the instructions were to remember all of the items; (3) s.d. or guess values were reported or graph axes were clearly labelled; (4) participants were healthy and between the ages of 18 and 35 years; and (5) the colours used were widely spaced, discriminable colours from the CIE $L^*a^*b^*$ colour space. Note that even slight changes in the colour wheel used between papers (or the addition of noise to stimuli[7]) change the perceptual confusability of the stimuli and therefore ideally call for a different similarity function to be measured and therefore a different prediction from TCC about the relationship between guess rate and s.d. However, in the current literature analysis, we simply assumed that these were the same for all papers. For papers that did not report s.d. or guess values in the text or tables, these values were obtained by digitizing figures with clear axis labels[62].

These inclusion criteria resulted in a diverse set of data points, including studies using sequential or simultaneous presentation, feedback versus no feedback, cues versus no-cues, varying encoding times (100–2,000 ms) and variable delays (1–10 s). A total of 14 papers and 56 data points were included (Supplementary Table 4). In general, TCC provides a strong fit to these existing data given the heterogeneity in methods (Supplementary Fig. 4) and these data are also consistent with the idea that there is no added guessing at high set sizes (Supplementary Fig. 5).

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

All relevant data for this manuscript are available at https://osf.io/j2h65/?view_only=fdd51dd775a945508c7cbbf25b662692.

Code availability

All relevant analysis code for this manuscript is available at https://osf.io/j2h65/?view_only=fdd51dd775a945508c7cbbf25b662692.

Received: 25 February 2020; Accepted: 28 July 2020;
Published online: 7 September 2020

References

1. Cowan, N. Metatheory of storage capacity limits. *Behav. Brain Sci.* **24**, 154–176 (2001).
2. Luck, S. J. & Vogel, E. K. Visual working memory capacity: from psychophysics and neurobiology to individual differences. *Trends Cogn. Sci.* **17**, 391–400 (2013).
3. Baddeley, A. Working memory: looking back and looking forward. *Nat. Rev. Neurosci.* **4**, 829–839 (2003).

4. Ma, W. J., Husain, M. & Bays, P. M.Changing concepts of working memory. *Nat. Neurosci.* **17**, 347–356 (2014).
5. Fukuda, K., Vogel, E., Mayr, U. & Awh, E. Quantity, not quality: the relationship between fluid intelligence and working memory capacity. *Psychonomic Bull. Rev.* **17**, 673–679 (2010).
6. Alloway, T. P. & Alloway, R. G. Investigating the predictive roles of working memory and IQ in academic attainment. *J. Exp. Child Psychol.* **106**, 20–29 (2010).
7. Zhang, W. & Luck, S. J.Discrete fixed-resolution representations in visual working memory. *Nature* **453**, 233–235 (2008).
8. Van den Berg, R., Shin, H., Chou, W. C., George, R. & Ma, W. J. Variability in encoding precision accounts for visual short-term memory limitations. *Proc. Natl Acad. Sci. USA* **109**, 8780–8785 (2012).
9. Bays, P. M. Noise in neural populations accounts for errors in working memory. *J. Neurosci.* **34**, 3632–3645 (2014).
10. Bays, P. M. Spikes not slots: noise in neural populations limits working memory. *Trends Cogn. Sci.* **19**, 431–438 (2015).
11. Serences, J. T. Neural mechanisms of information storage in visual short-term memory. *Vis. Res.* **128**, 53–67 (2016).
12. Bae, G. Y., Olkkonen, M., Allred, S. R., Wilson, C. & Flombaum, J. I. Stimulus-specific variability in color working memory with delayed estimation. *J. Vision* **14**, 7 (2014).
13. Allred, S. R. & Flombaum, J. I. Relating color working memory and color perception. *Trends Cogn. Sci.* **18**, 562–565 (2014).
14. Pratte, M. S., Park, Y. E., Rademaker, R. L. & Tong, F.Accounting for stimulus-specific variation in precision reveals a discrete capacity limit in visual working memory. *J. Exp. Psychol. Hum. Percept. Perform.* **43**, 6–17 (2017).
15. Torgerson, W. S. *Theory and Methods of Scaling* (Wiley, 1958).
16. Maloney, L. T. & Yang, J. N. Maximum likelihood difference scaling. *J. Vision* **3**, 5 (2003).
17. Shepard, R. N. Toward a universal law of generalization for psychological science. *Science* **237**, 1317–1323 (1987).
18. Sims, C. R. Efficient coding explains the universal law of generalization in human perception. *Science* **360**, 652–656 (2018).
19. Nosofsky, R. M. Similarity scaling and cognitive process models. *Annu. Rev. Psychol.* **43**, 25–53 (1992).
20. Tanner, W. P.Jr & Swets, J. A.A decision-making theory of visual detection. *Psychol. Rev.* **61**, 401–409 (1954).
21. Macmillan, N. A. & Creelman, C. D. *Detection Theory: A User's Guide* 2nd edn (Erlbaum, 2005).
22. Wilken, P. & Ma, W. J. A detection theory account of change detection. *J. Vision* **4**, 11 (2004).
23. Fougnie, D., Suchow, J. W. & Alvarez, G. A. Variability in the quality of visual working memory. *Nat. Commun.* **3**, 1229 (2012).
24. Loftus, G. R. & Bamber, D. Weak models, strong models, unidimensional models, and psychological time. *J. Exp. Psychol. Learn. Mem. Cogn.* **16**, 16–19 (1990).
25. Smith, P. L., Lilburn, S. D., Corbett, E. A., Sewell, D. K. & Kyllingsbæk, S. The attention-weighted sample-size model of visual short-term memory: attention capture predicts resource allocation and memory load. *Cogn. Psychol.* **89**, 71–105 (2016).
26. Bays, P. M., Catalao, R. F. & Husain, M. The precision of visual working memory is set by allocation of a shared resource. *J. Vision* **9**, 7 (2009).
27. Roberts, S. & Pashler, H.How persuasive is a good fit? A comment on theory testing. *Psychol. Rev.* **107**, 358–367 (2000).
28. Kahana, M. J. & Sekuler, R. Recognizing spatial patterns: a noisy exemplar approach. *Vis. Res.* **42**, 2177–2192 (2002).
29. Gold, J. M., Wilk, C. M., McMahon, R. P., Buchanan, R. W. & Luck, S. J.Working memory for visual features and conjunctions in schizophrenia. *J. Abnorm. Psychol.* **112**, 61–71 (2003).
30. Haberman, J., Brady, T. F. & Alvarez, G. A.Individual differences in ensemble perception reveal multiple, independent levels of ensemble representation. *J. Exp. Psychol. Gen.* **144**, 432–446 (2015).
31. Miner, A. E., Schurgin, M. W. & Brady, T. F. Is working memory inherently more 'precise' than long-term memory? Extremely high fidelity visual long-term memories for frequently encountered objects. *J. Exp. Psychol. Human Percept. Perform.* **46**, 813 (2020).
32. Brady, T. F., Konkle, T., Gill, J., Oliva, A. & Alvarez, G. A. Visual long-term memory has the same limit on fidelity as visual working memory. *Psychol. Sci.* **24**, 981–990 (2013).
33. Asplund, C. L., Fougnie, D., Zughni, S., Martin, J. W. & Marois, R. The attentional blink reveals the probabilistic nature of discrete conscious perception. *Psychol. Sci.* **25**, 824–831 (2014).
34. Pratte, M. S. Iconic memories die a sudden death. *Psychol. Sci.* **29**, 877–887 (2018).
35. Richter, F. R., Cooper, R. A., Bays, P. M. & Simons, J. S.Distinct neural mechanisms underlie the success, precision, and vividness of episodic memory. *eLife* **5**, e18260 (2016).
36. Dunn, J. C. & Kalish, M. L. *State–Trace Analysis* (Springer, 2018).
37. Zokaei, N. et al. Visual short-term memory deficits associated with GBA mutation and Parkinson's disease. *Brain* **137**, 2303–2311 (2014).
38. Rolinski, M. et al. Visual short-term memory deficits in REM sleep behaviour disorder mirror those in Parkinson's disease. *Brain* **139**, 47–53 (2015).
39. Pertzov, Y. et al. Binding deficits in memory following medial temporal lobe damage in patients with voltage-gated potassium channel complex antibody-associated limbic encephalitis. *Brain* **136**, 2474–2485 (2013).
40. Brady, T. F. & Alvarez, G. A. Hierarchical encoding in visual working memory: ensemble statistics bias memory for individual items. *Psychol. Sci.* **22**, 384–392 (2011).
41. Brady, T. F. & Alvarez, G. A. Contextual effects in visual working memory reveal hierarchically structured memory representations. *J. Vision* **15**, 6 (2015).
42. Williams, J., Brady, T. & Störmer, V. S. Guidance of attention by working memory is a matter of representational fidelity, not a privileged status for one or more items. Preprint at *PsyArXiv* https://doi.org/10.31234/osf.io/c4t92 (2019).
43. Henson, R. N. A., Rugg, M. D., Shallice, T. & Dolan, R. J. Confidence in recognition memory for words: dissociating right prefrontal roles in episodic retrieval. *J. Cogn. Neurosci.* **12**, 913–923 (2000).
44. Rutishauser, U. et al. Representation of retrieval confidence by single neurons in the human medial temporal lobe. *Nat. Neurosci.* **18**, 1041–1050 (2015).
45. Marr, D. *Vision: A Computational Investigation Into the Human Representation and Processing of Visual Information* (Henry Holt and Company, 1982).
46. Bays, P. M. Correspondence between population coding and psychophysical scaling models of working memory. Preprint at *BioRxiv* https://doi.org/10.1101/699884 (2019).
47. Wei, X. X. & Stocker, A. A.A Bayesian observer model constrained by efficient coding can explain 'anti-Bayesian' percepts. *Nat. Neurosci.* **18**, 1509–1517 (2015).
48. Wei, X. X. & Stocker, A. A. Lawful relation between perceptual bias and discriminability. *Proc. Natl Acad. Sci. USA* **114**, 10244–10249 (2017).
49. Krauskopf, J. & Gegenfurtner, K. R. Color discrimination and adaptation. *Vis. Res.* **32**, 2165–2175 (1992).
50. Giesel, M., Hansen, T. & Gegenfurtner, K. R. The discrimination of chromatic textures. *J. Vision* **9**, 11 (2009).
51. Rademaker, R. L., Tredway, C. H. & Tong, F. Introspective judgments predict the precision and likelihood of successful maintenance of visual working memory. *J. Vision* **12**, 21 (2012).
52. Wixted, J. T. & Wells, G. L.The relationship between eyewitness confidence and identification accuracy: a new synthesis. *Psychol. Sci. Public Interest* **18**, 10–65 (2017).
53. Fougnie, D., Brady, T. F. & Alvarez, G. A. If at first you don't retrieve, try, try again: the role of retrieval failures in visual working memory. *J. Vis.* **14**, 851–851 (2014).
54. Difallah, D., Filatova, E. & Ipeirotis, P. Demographics and dynamics of mechanical Turk workers. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining* 135–143 (ACM, 2018).
55. Brady, T. F. & Tenenbaum, J. B.A probabilistic model of visual working memory: incorporating higher order regularities into working memory capacity estimates. *Psychol. Rev.* **120**, 85–109 (2013).
56. Nadarajah, S., Afuecheta, E. & Chan, S. On the distribution of maximum of multivariate normal random vectors. *Commun. Stat. Theory Methods* **48**, 2425–2445 (2019).
57. Fougnie, D., Asplund, C. L. & Marois, R. What are the units of storage in visual working memory? *J. Vision* **10**, 27 (2010).
58. Suchow, J. W., Brady, T. F., Fougnie, D. & Alvarez, G. A. Modeling visual working memory with the MemToolbox. *J. Vision* **13**, 9 (2013).
59. Myung, J. I. & Pitt, M. A. in *Steven's Handbook of Experimental Psychology and Cognitive Neuroscience* 4th edn, Vol. 5 (eds Wixted, J. & Wagenmakers, E.-J.) 85–118 (John Wiley & Sons, 2018).
60. Brady, T. F., Konkle, T., Alvarez, G. A. & Oliva, A. Visual long-term memory has a massive storage capacity for object details. *Proc. Natl Acad. Sci. USA* **105**, 14325–14329 (2008).
61. Wixted, J. T. & Mickes, L.A continuous dual-process model of remember/know judgments. *Psychol. Rev.* **117**, 1025–1054 (2010).
62. Rohatgi, A. WebPlotDigitizer (2011); https://automeris.io/WebPlotDigitizer/
63. Thurstone, L. L. A law of comparative judgment. *Psychol. Rev.* **34**, 273–286 (1927).
64. Rotello, C. M. in *Learning and Memory: A Comprehensive Reference* 2nd edn, Vol. 2 (eds Byrne, J. H. & Wixted, J. T.) 201–226 (Elsevier, 2017).

Acknowledgements

We thank V. Störmer, R. Rademaker, J. Wolfe, M. Robinson, D. Fougnie and T. Konkle for comments on these ideas and on the manuscript, and Y. H. Chung and B. Hawkins for help with data collection. For funding, we also acknowledge NSF CAREER

(BCS-1653457; to T.F.B.). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

M.W.S. jointly conceived of the model with J.T.W. and T.F.B. M.W.S. and T.F.B. designed the experiments. T.F.B. wrote the code, ran the model and analysed the output data. M.W.S., J.T.W. and T.F.B. wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41562-020-00938-0.

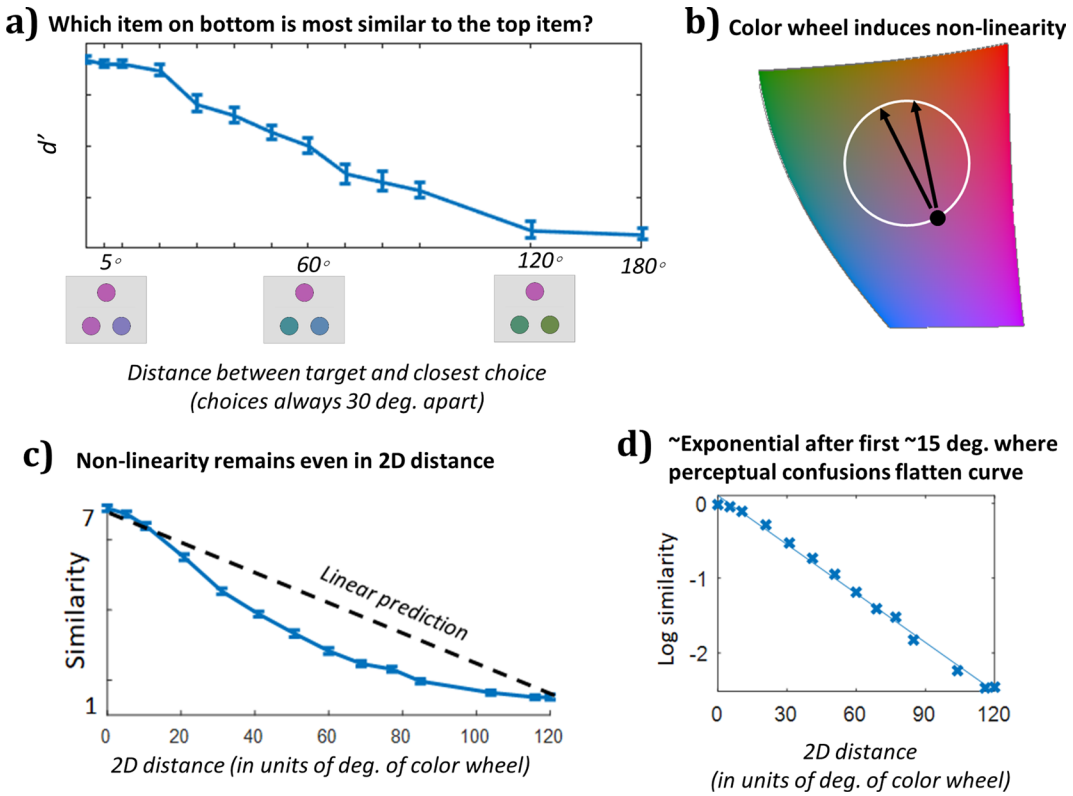
Supplementary information is available for this paper at https://doi.org/10.1038/s41562-020-00938-0.

Correspondence and requests for materials should be addressed to M.W.S. or T.F.B.

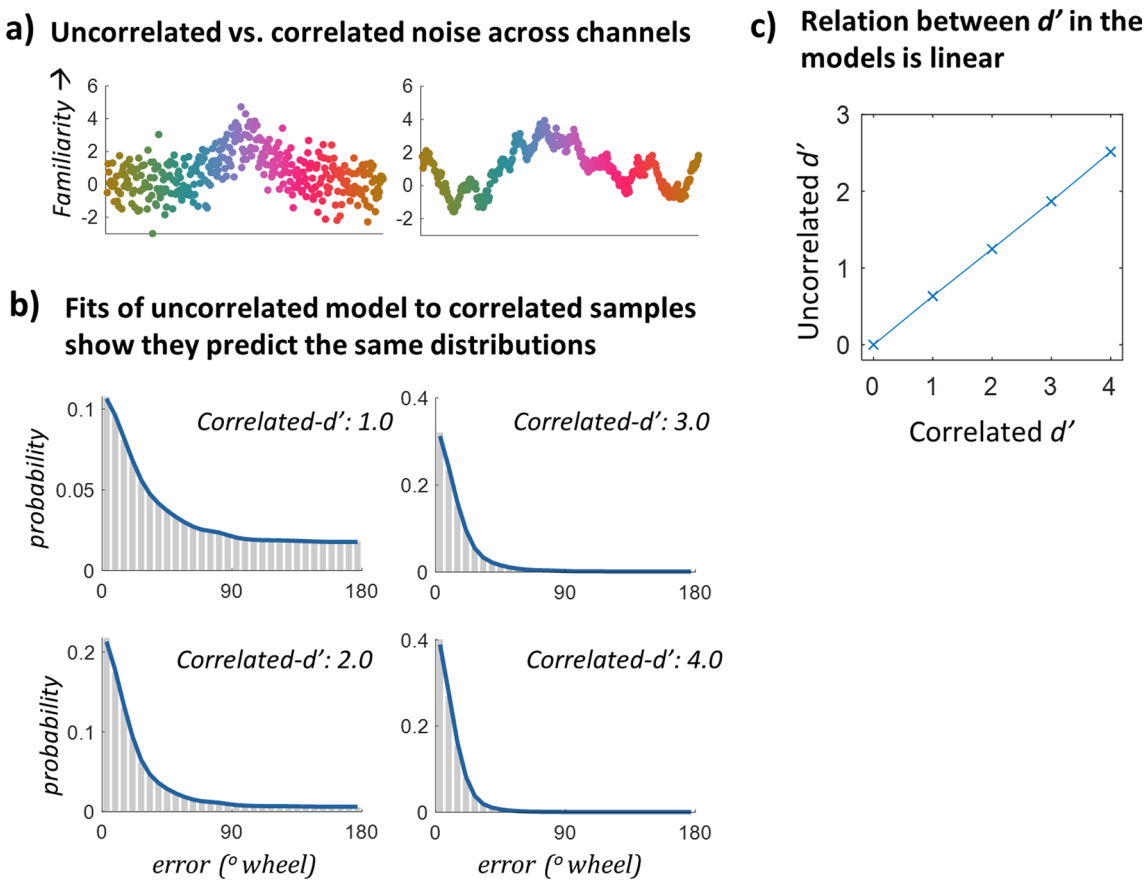
Peer review information Primary Handling Editor: Marike Schiffer.

Reprints and permissions information is available at www.nature.com/reprints.

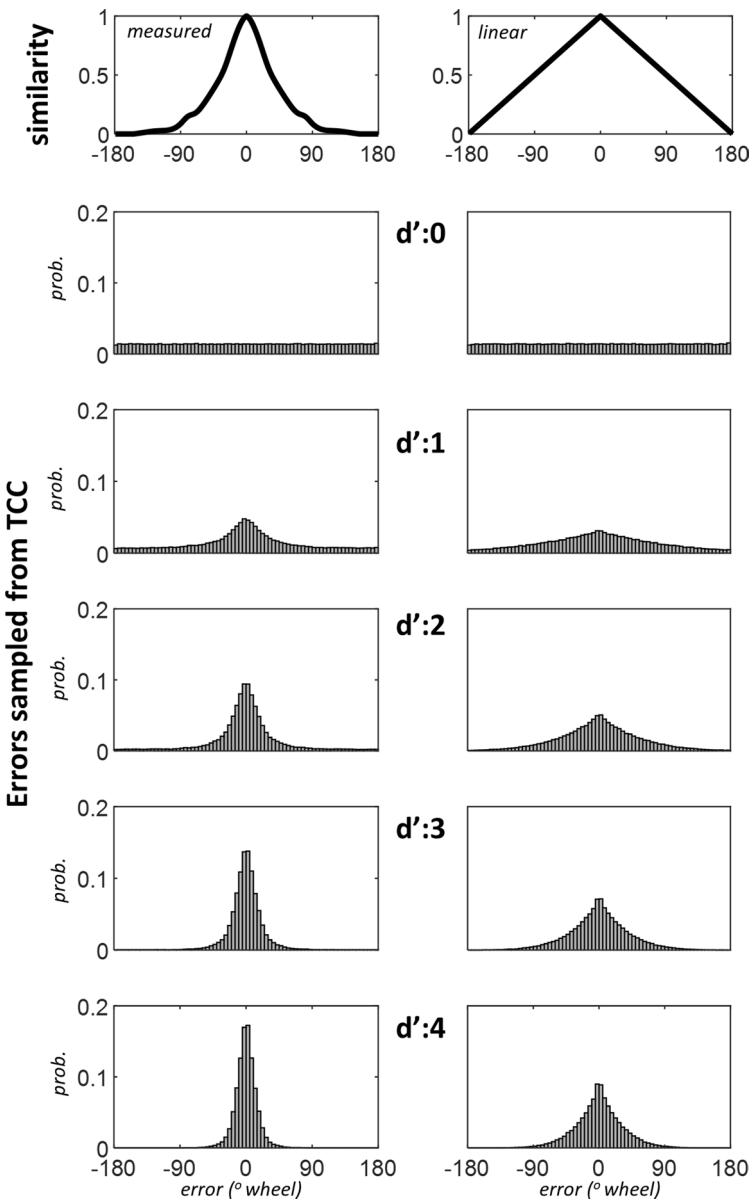
Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



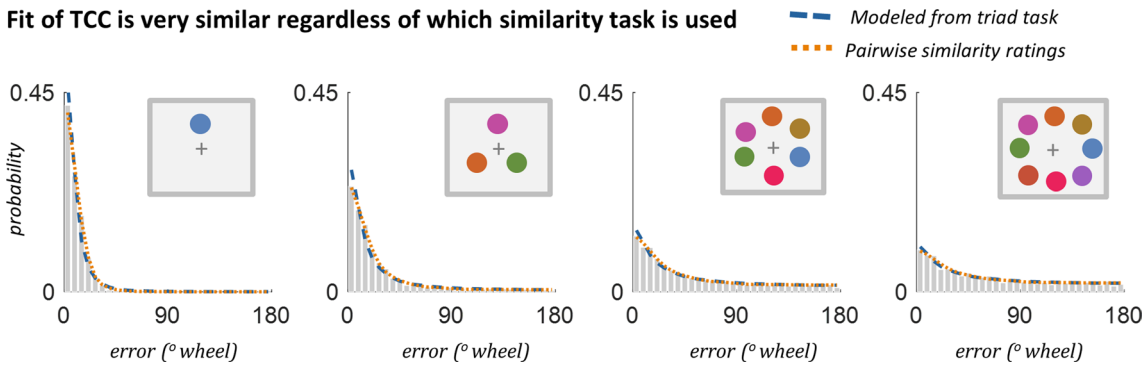
Extended Data Fig. 1 | Similarity as a function of distance in color space. a, Data from all distances in the fixed distance triad task (Fig. 1c). On each trial, there was a target color, here always at 0°, and participants' task was to choose which of two other colors was closer to the target color in color space. The two choice colors always differed by 30°. The x-axis shows the closer color of the two choice colors. That is, the 150° label on the x-axis reflects performance on a condition where the two choices were 150° and 180° away from the target color. As shown with a subset of this data in Fig. 1c, increasing distance from the target results in a decreased ability to tell which of two colors is closer to the target in color space. This shows the non-linearity of color space with respect to judgments of color similarity. Note that this function does not depict the actual psychophysical similarity function: Roughly speaking, the *d'* estimate in this graph is the estimate of instantaneous slope (over a 30 deg. range) in the similarity function in Fig. 1f. **b**, Despite being conceived of as a color wheel in many memory experiments, in reality, participants internal representation of color–and thus the confusability between colors–ought to be a function of their linear distance in an approximately 3D color space, not their angular distance along the circumference of an artificially imposed wheel. Since the colors are equal luminance, we can conceive of this on a 2D plane. Thus, on this plane the confusability of a color "180 degrees away" on the wheel is only slightly lower than one "150 degrees away" on the wheel, since in 2D color space it is only slightly further away. This simple non-linearity from ignoring the global structure of the color 'wheel' partially explains the long tails observed in typical color report experiments, although it does not explain the full degree of this non-linearity, which is additionally attributable to psychophysical similarity being a non-linear function even of distance across 2D color space. **c**, The similarity function remains non-linear even in 2D color space. Distances here are scaled relative to the color wheel rather than in absolute CIELa*b* values., for example, an item 180 degrees opposite on the color wheel is "120" in real distance since if the distance along the circumference is 180, 120 is the distance across the color wheel. **d**, Plotted on a log axis, the similarity falls off approximately linearly, indicating that similarity falls of roughly exponentially with the exception of colors nearby the target. The non-exponential fall-off near the 0 point reflects perceptual noise/lack of perceptual discriminability between nearby colors. As shown in Fig. 1, when you convolve measured perceptual noise with an exponential function, this provides a very good fit to the similarity function, consistent with a wide-variety of evidence about the structure of similarity and generalization[19].



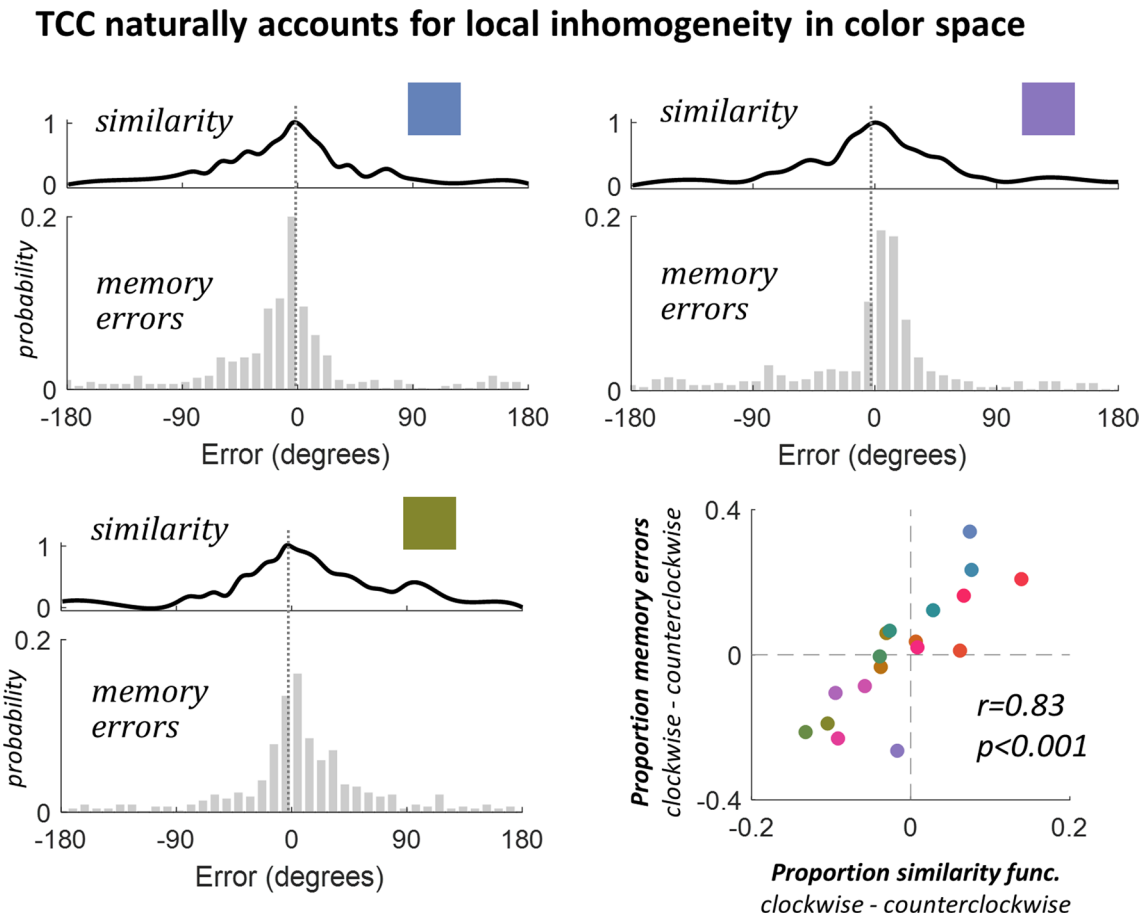
Extended Data Fig. 2 | Simulations of uncorrelated vs. correlated noise versions of TCC. In the main text, we report d' from a version of TCC where noise in similar color channels is correlated, based on measured perceptual confusions. However, this decision to correlate the noise of nearby colors is not critical, as shown in this simulation of uncorrelated vs. correlated noise versions of TCC. Only the correlated-noise TCC produces true *d'* values–those that are interchangeable with *d'* you'd estimate from a same/diff task with the same stimuli. However, the simpler uncorrelated noise TCC predicts the exact same distributions of errors in continuous report, and the d' values between the correlated and uncorrelated noise models are linearly related by a factor of ~0.65. Thus, in many cases it may be useful to fit the uncorrelated TCC to data and then adjust the *d'* rather than fitting correlated noise TCC. This also means that for color, similarity alone without perceptual confusion data can be used to make linear (but not exact) predictions about confusability in n-AFC tasks outside the range of perceptual confusion (approx. 15 deg).



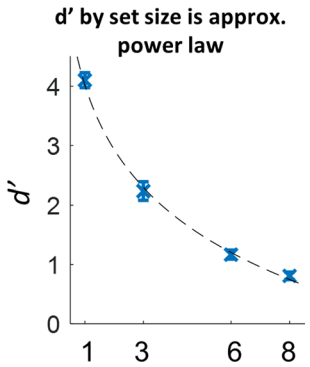
Extended Data Fig. 3 | Simulations comparing the measured psychological similarity function to a linear similarity function. Simulations show data sampled from TCC, using either the measured psychological similarity function or a linear similarity function. Given a linear similarity function, it is clear TCC does not predict response distributions similar to human performance – accurate memory fits are critically dependent on the well-known exponential-like shape of similarity functions. Notice also how the max rule from the signal detection decision process plays a major role in the shape of the distributions. Since people pick the strongest signal, the distribution of max signals is peakier than the underlying signals themselves (which always follows the similarity function).



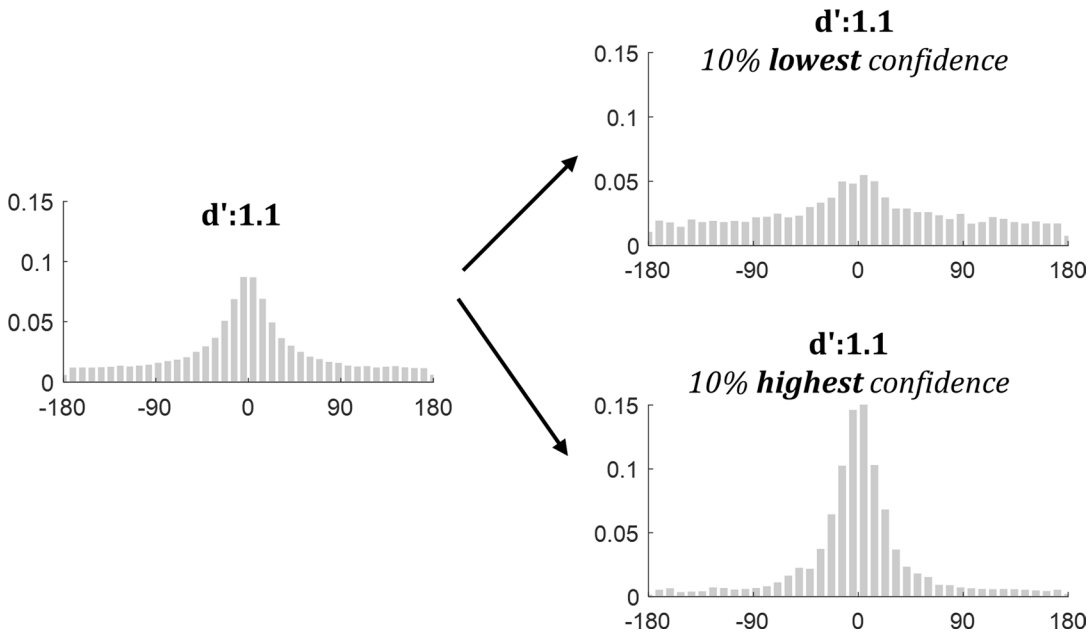
Extended Data Fig. 4 | Comparisons of fit to memory data for different measured similarity functions. Comparison of fit to memory data for similarity functions reported in main text. In the current data for color, both the model-based triad psychophysical scaling data and the Likert similarity rating produce extremely similar data (see Fig. 1). Thus, they all produce similar fits to the memory data (shown here are the set size data). It is important to note that depending on the number of trials, a large number of data points (that is subjects) may be necessary in order to obtain reliable estimates of a given stimulus space in the triad and quad scaling tasks (we use the quad task for face similarity). The Likert task requires considerably less data to estimate, and it was in agreement with the results of the triad task for colors, so we rely on it as our primary measure of similarity in the current fits. However, depending on the stimulus space, observers may utilize different strategies in such subjective similarity tasks (particularly for spaces, like orientation, where it is obviously a linear physical manipulation), and ultimately an objective task like the quad task may be best to understand the psychophysical similarity function. This is why for the face space task we used the quad similarity task. The task used to estimate similarity is important in that it is important 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. How best to ensure that participants rely on absolute intervals is represented in a large literature dating to Thurstone[63] and Torgerson[15].



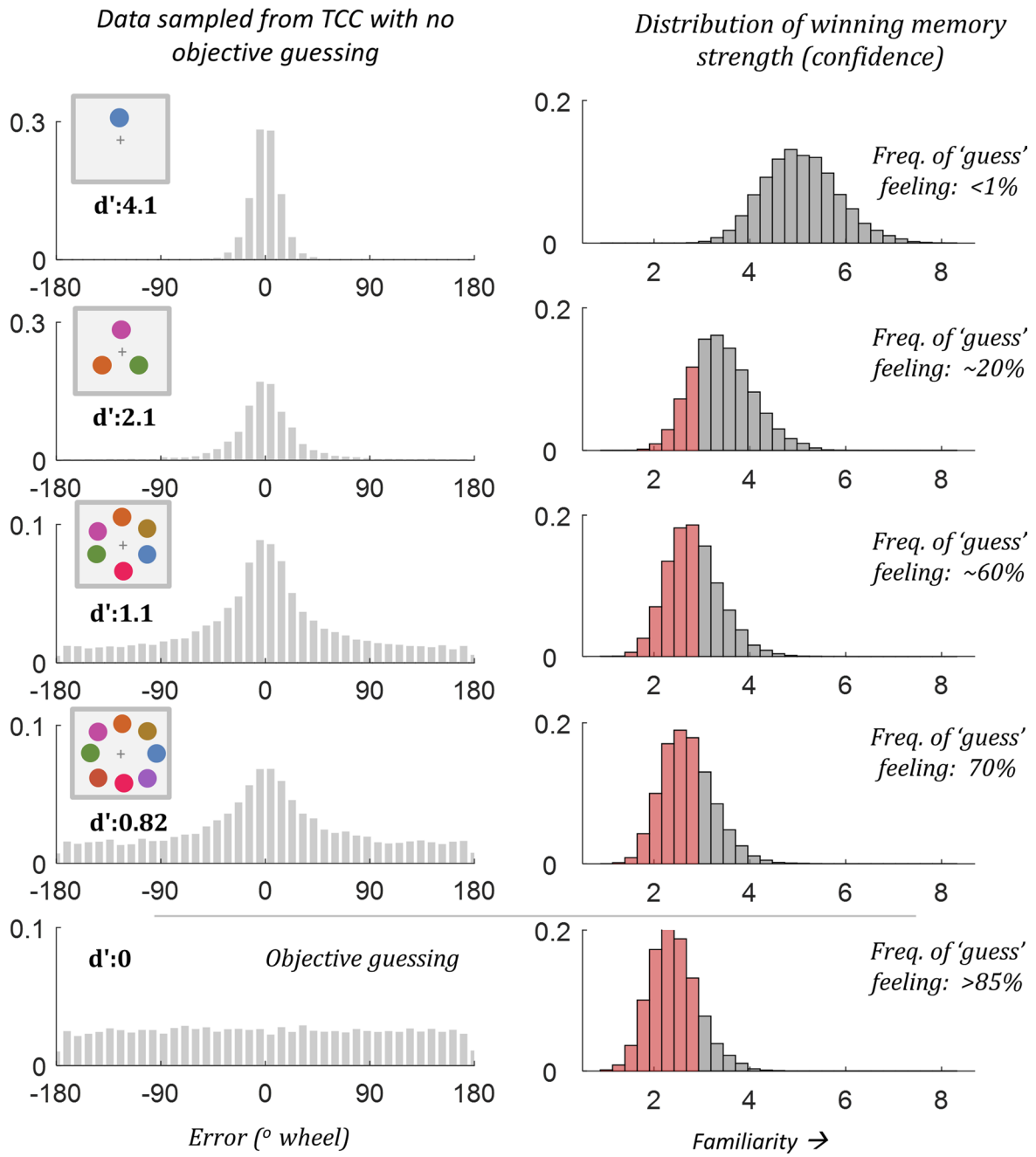
Extended Data Fig. 5 | Non-uniformities across color space in memory and similarity. Non-uniformities in memory and similarity for set size data reported in the main text. Many stimulus spaces contain non-uniformities, which may affect subsequent working memory performance. Indeed, Bae et al.[12] discovered non-uniformities in working memory for color, where responses for targets tend to be more precise for some colors than others and can be biased towards nearby categorical anchors (that is red, blue, yellow, etc). While many assume randomizing target colors in working memory should account for potential biases arising from a non-uniform feature space, others have suggested these differences may have broader consequences than previously considered[13,14]. A key advantage of TCC is that by taking into account the psychophysical similarity function, non-uniformities within whatever feature space being probed can be automatically captured if psychophysical similarity data is measured separately from each relevant starting point in the feature space (for example, Fig. 1d). In the current work, we mostly use only a single psychophysical similarity estimate averaged across possible starting points and fit memory data averaged across starting points. However, this is not necessary to the TCC framework, and is only a simplification–if we wish to fit memory data averaged across all targets, we should use similarity averaged across all targets (or use the particular similarity function relevant to each item on each trial). Here we show that rather than using a psychophysical similarity function that averages over all targets, one can also use similarity specific to each possible target, which differ and have predictable consequences for memory in our set size experiment. For example, the propensity of errors (at set size 1, 3, 6 and 8) in the clockwise vs. counterclockwise direction for a given target color is directly predicted by the similarity function–even when very similar colors have more similar colors in opposite directions (top row), and this is true across all color bins (bottom right). Thus, using target-specific similarity functions naturally captures potential non-uniformities or biases within a feature space with no change in the TCC framework.



Extended Data Fig. 6 | d' as a function of set size. Data from the set size experiment reported in the main text. While memory strength varies according to a variety of different factors, many researchers have been particularly interested in the influence of set size. TCC shows that at a fixed encoding time and with a fixed delay, memory strength (d') decreases according to a power law as set size changes, broadly consistent with fixed resource theories of memory[25,26]. However, capacity cannot be fixed globally, as the total "capacity" appears to smoothly change with encoding time and delay and differs for different stimuli.



Extended Data Fig. 7 | Variation in representational fidelity with the same d′ by separating on strength of strongest memory signal. Simulation from TCC illustrating how signal detection can predict variance in representational fidelity as a function of confidence even with a fixed d′ (see also[42]). Some studies used to support variability of information across individual items or trials have done so by using a confidence metric[51]. While variability and confidence are distinct from one another, in a large amount of research they are inextricably linked. An interesting advantage and implication of signal detection-based models is that they naturally predict confidence data[64]. In particular, the strength of the winning memory match signal is used as the measure of memory strength–and confidence–in signal detection models of memory. Thus, even with a fixed *d*′ value for all items, TCC naturally predicts varying distributions relative to confidence. This likely explains some of the evidence previously observed in the literature that when distinguishing responses according to confidence, researchers found support for variability in precision among items / trials. Note that this occurs in TCC even though *d*′ is fixed in this simulation–that is, all trials are generated from a process with the same signal-to-noise ratio. Thus, variability in responses as a function of confidence (or related effects, like improved performance when participants choose their own favorite item to report[23]) are not evidence for variability in *d*′ in TCC, but simply a natural prediction of the underlying signal detection process. Of course, it is possible *d*′ may also vary between items, which remains an open question.



Extended Data Fig. 8 | Simulation of expected confidence as a function of set size in TCC. Participants in a set size 8 working memory experiment often feel like they do not remember an item and are "guessing", leading to a wide variety of models that predict people know nothing about many items at high set sizes and truly are objectively guessing. However, as noted in Extended Data Fig. 7, signal detection naturally accounts for varying confidence, and so can easily account for this subjective feeling of guessing even though in fact, models like TCC predict that people are almost never responding based on no information at all about the item they just saw. In particular, confidence in signal detection is based on the strength of the winning memory signal. Imagine that the subjective feeling of guessing occurs whenever your memory match signal is below some threshold (here, arbitrarily set to 2.75). This would lead to people never feeling like they are guessing at set size 1, and nearly always feeling like they are guessing if they objectively closed their eyes and saw nothing. However, this would also make people feel like they are guessing a large part of the time at set size 6 and 8, even though this data is simulated from TCC–and the generative process always contains information about all items. This is the key distinction in signal detection models between the subjective feeling of guessing and the claim that people are objectively guessing.

nature research

Corresponding author(s): Timothy Brady

Last updated by author(s): Jun 30, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Studies deployed on Mechanical Turk were coded using HTML and Javascript. Studies collected in-lab were coded and run on MATLAB using the Psychophysics Toolbox (Brainard, 1997; Pelli, 1997).
Data analysis	All analyses were run on MATLAB 2016B with some analyses using the MemToolbox for MATLAB (Suchow et al. 2013). The code used to support the findings are described in mathematical and MATLAB notation in the methods section and available on OSF.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The data and analysis code for all figures and analyses is vailable on OSF.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	Experiment 1 (Fixed Distance Triad): Quantitative experimental study. Participants judged which of two colors presented were more similar to a target color. Experiment 2 (Psychological Scaling Triad): Quantitative experimental study. Participants judged which of two colors presented were more similar to a target color, as in the fixed distance triad experiment. Experiment 3 (Color Similarity): Quantitative experimental study. Participants judged the similarity of two colors presented simultaneously on a Likert scale, ranging from 1 (least similar) to 7 (most similar). Experiment 4 (Perceptual matching): Quantitative experimental study. Participants were shown a color and had to match this color using a continuous report wheel. Experiment 5 (Continuous Color Report Set Size): Quantitative experimental study. Participants completed a visual working memory task, where they were presented either 1, 3, 6, or 8 colors to remember and after a delay, an item was probed and participants reported the color using a continuous color report wheel. Experiment 6 (Continuous Color Report Delay): Quantitative experimental study. Participants completed a visual working memory task, where they were presented either 1, 3, or 6 colors to remember and after a varying delay (1s, 3s, 5s), an item was probed and participants reported the color using a continuous color report wheel. Experiment 7 (Continuous Color Report Encoding Time): Quantitative experimental study. Participants completed a visual working memory task, where they were presented either 1, 3, or 6 colors (presented for either 100, 500, or 1500ms) to remember and after a delay, an item was probed and participants reported the color using a continuous color report wheel. Experiment 8 (2AFC Different Foil Similarities): Quantitative experimental study. Participants completed a visual working memory task, where they were presented 4 colors to remember and after a delay, had to answer a 2AFC memory probe about one of the colors. The foil of the 2AFC varied from the target 180, 72, 24, or 12 degrees. Experiment 9 (2AFC generalization to n-AFC): Quantitative experimental study. Participants completed a visual working memory task, where they were presented 4 colors to remember and after a delay, an item was probed and participants reported the color either with a 2-AFC, 8-AFC, 60-AFC or full continuous report (360-AFC). Experiment 10 (Face Identity Continuous Report): Quantitative experimental study. Participants completed a visual working memory task, where they were presented either 1 or 3 face identities to remember and after a delay, an item was probed and participants reported the face identity of the target using a continuous face identity report wheel. Experiment 11 (Face Identity Quad Task): Quantitative experimental study. Participants judged which of two pairs of faces presented were more similar. Experiment 12 (Face Identity Perceptual Matching): Quantitative experimental study. Participants were shown a face and had to match this color using a continuous report wheel. Experiment 13 (Visual Long-Term Memory Color Report): Quantitative experimental study. Participants encoded images of real-world objects embedded in specific colors during an encoding block. At test, old and new objects were presented, and participants judged the object as old or new. If they reported the item was old, they reported its color using a continuous color wheel.
Research sample	Experiment 1 (Fixed Distance Triad): N=40 participants on Amazon Mechanical Turk participated. Mechanical Turk users form a representative subset of adults in the United States (Berinsky, Huber, & Lenz, 2012; Buhrmester, Kwang, & Gosling, 2011), and data from Turk are known to closely match data from the lab on visual cognition tasks (Brady & Alvarez, 2011; Brady & Tenenbaum, 2013). including providing extremely reliable and high-agreement on color report data (Brady & Alvarez, 2015). Experiment 2 (Psychological Scaling Triad): N=100 participants on Mechanical Turk participated. Experiment 3 (Color Similarity): N=50 participants on Mechanical Turk participated. Experiment 4 (Perceptual Matching): N=40 participants on Amazon Mechanical Turk participated. Experiment 5 (Continuous Color Report Set Size): N=20 participants in the lab at UC San Diego participated. Experiment 6 (Continuous Color Report Delay): N=20 participants in the lab at UC San Diego participated. Experiment 7 (Continuous Color Report Encoding Time): N=20 participants in the lab at UC San Diego participated. Experiment 8 (2AFC Different Foil Similarities): N=60 participants on Mechanical Turk participated. Experiment 9 (2AFC generalization to n-AFC): N=60 participants on Mechanical Turk participated. Experiment 10 (Face Identity Continuous Report): N=50 participants on Mechanical Turk participated. Experiment 11 (Face Identity Quad Task): N=102 participants on Mechanical Turk participated. Experiment 12 (Face Identity Perceptual Matching): N=40 participants on Mechanical Turk participated. Experiment 13 (Visual Long-Term Memory Color Report): N=30 participants in the lab at UC San Diego participated.
Sampling strategy	All sample sizes were decided a priori. All studies used convenience sampling of either users from Amazon Mechanical Turk or undergraduates from the University of California, San Diego.
Data collection	Experiments 1-4, 8-9, 11-12: Experiments were deployed online via Mechanical Turk. Participants computer screens showed stimuli, and responses were collected via keyboard or mouse. Experiment 5-7, 13: The study took place in a dimly lit sound-attenuated room. Stimuli were presented on a Macintosh iMac computer, and responses were collected via keyboard or mouse.
Timing	All of the included studies (except for the LTM data reported in a previous paper) were collected between September 2017 - July 2019
Data exclusions	In studies with continuous report memory tasks, we excluded participants <2 std below the mean in overall d' across conditions. This did not result in the exclusion of any participants with the exception of one participant in the delay experiment who had chance level performance at all delays/set sizes. In all studies with objective perceptual tasks, we used the same exclusion rules: excluding trials with reaction times <200ms or >5000ms or any participants whose overall accuracy was 2 standard deviations below the mean.

In our subjective similarity experiment, we included manipulation check trials where the two items they were judging the similarity for were identical, and excluded any participant whose mean rating on these trials was not >6 (on a 1-7 scale).

In particular:

Experiment 1: . To be conservative about the inclusion of participants, we excluded any participant who made an incorrect response in any of the 10 trials where the target color exactly matched one of the choice colors, leading to the exclusion of 7 of the 40 participants, and based on our a priori exclusion rule, excluded any participants whose overall accuracy was 2 standard deviations below the mean, leading to the exclusion of 0 additional participants. In addition, based on an a priori exclusion rule, we excluded trials with reaction times <200ms or >5000ms, which accounted for 1.75% (SEM:0.5%) of trials.
Experiment 2: Using our a priori exclusion rule, we excluded any participant whose overall accuracy was 2 standard deviations below the mean (M=77.5%) leading to the exclusion of 8 of the 100 participants. In addition, based on an a priori exclusion rule, we excluded trials with reaction times <200ms or >5000ms, which accounted for 1.7% (SEM:0.26%) of trials.
Experiment 3: Following our a priori exclusion rule, we excluded trials with reaction times <200ms or >5000ms, which accounted for 3.0% (SEM:0.4%) of trials. 2 participants were excluded for failing the manipulation check.
Experiment 4: 1 participant's data was lost due to experimenter error and following our priori exclusion rule 2 participants were excluded for an average error rate greater than 2 standard deviations away from the mean.
Experiment 5: No participants were excluded.
Experiment 6: One participant was excluded for being <2 std below the mean in overall d'.
Experiment 7: No participants were excluded.
Experiment 8: Following our a priori exclusion rule, 5 participants were excluded for below chance performance in the maximally easy 180 deg. offset condition, leaving N=55 participants.
Experiment 9: One participant's data was lost, and following our priori exclusion rule 7 participants were excluded for below chance performance in the maximally easy 2-AFC, 180 deg. offset condition, leaving N=52 participants.
Experiment 10: No participants were excluded.
Experiment 11: Following our a priori exclusion rule, we excluded participants whose overall performance level was more than 2 standard deviations below the mean, resulting in a final sample of N=85.
Experiment 12: 7 participants were excluded for below chance error rates.
Experiment 13: As described in Miner et al. (in press) 6 participants were excluded.

Non-participation: No participants dropped out or declined participation.

Randomization: Every study is within-subject so no randomization of participants to groups was required

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics: For studies conducted in-lab, participants were undergraduates from University of California, San Diego. All participants at UCSD reported normal color vision and were between the ages of 18-35 years old.

For studies conducted online, Mechanical Turk users form a representative subset of adults in the United States (Berinsky, Huber, & Lenz, 2012; Buhrmester, Kwang, & Gosling, 2011), and data from Turk are known to closely match data from the lab on visual cognition tasks (Brady & Alvarez, 2011; Brady & Tenenbaum, 2013).

Recruitment: For studies conducted in-lab, participants were recruited via the Sona Systems online portal, where psychology undergraduate students can participate in studies for extra credit. Studies conducted on Mechanical Turk performed recruitment only by posting HITs, as is standard.

Ethics oversight: The studies were approved by the UC San Diego IRB.

Note that full information on the approval of the study protocol must also be provided in the manuscript.