bioRxiv preprint doi: https://doi.org/10.1101/325472; this version posted August 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Psychophysical Scaling Reveals a Unified Theory of Visual Memory Strength

Mark W. Schurgin[1]*, John T. Wixted[1], Timothy F. Brady[1]

[1] Department of Psychology
University of California, San Diego

* Correspondence to:
Mark Schurgin
9500 Gilman Drive #0109
La Jolla, CA 92093
Email: mschurgin@ucsd.edu

bioRxiv preprint doi: https://doi.org/10.1101/325472; this version posted August 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Abstract

Almost all models of visual memory implicitly assume that 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 non-linearly related to stimulus space. This result calls into question a foundational assumption of virtually all 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 ~3-4 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 significant implications for the study of visual memory and lead to a substantial reinterpretation of the relationship between perception, working memory and long-term memory.

bioRxiv preprint doi: https://doi.org/10.1101/325472; this version posted August 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Working memory is typically conceptualized as a fixed capacity system, with a discrete number of items, each 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.

Continuous feature spaces are often used to investigate working memory, as they allow the precise quantification of information stored in memory. In one prominent paradigm, researchers present a set of stimuli to remember and then probe 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 analyzed using the circular difference between the true stimulus and reported stimulus, which is then modeled to quantify memory performance[7,8]. Such models appear to make critical distinctions between distinct kinds of memories: either how many items are represented vs. how precisely they are represented[7] or between items encoded with high precision vs. extremely low precision[8].

Here we show that taking into account the globally non-linear function of psychophysical similarity between test stimuli opens the door to a new model of visual working and visual long-term memory, one that is both more parsimonious and more powerful than existing models. The new model suggests that all memories arise fundamentally from a single process, rather than from different memory states or degrees of encoding precision, and it uniquely permits parameter-free generalization across different stimulus spaces and different tasks. The model operates in a signal detection framework, as most models of long-term memory do, suggesting a unified framework can be used to understand the nature of mnemonic representations and decision-making across working memory and long-term memory. We begin with working memory for color as our main case study and then expand the psychophysical similarity framework and resulting model to encompass working memory for faces (a multi-feature stimulus space) and long-term memory for real-world objects.

Results

Psychophysical scaling. While previous work has documented local inhomogeneities in stimulus space[9-11], we measured the global structure of the working memory color space. That is, we measured how similarity scales with distance measured in terms of degrees along the color wheel. To do so, we tested how accurately participants could determine which of two test colors was more similar to a target color using a triad task[12,13]. This is a perceptual task, but it is analogous to the working memory situation where participants have a target color in mind and are asked to compare other colors to that target. We found that with a fixed 30° distance between two color choices, participants are significantly more accurate at determining which color is closer to the target when the two colors are nearby in color space compared to when they are far from the target (Fig. 1C, S1; 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 color 120°or 150° from the target was closer to the target, whereas this task is trivial if the colors are 5° and 35° from the target. To compute a full psychophysical similarity function, we utilized the just-described triad task with additional distance pairs. We then applied the maximum likelihood difference scaling technique[13] (MLDS) commonly used for perceptual scaling to estimate how differences between color stimuli are actually perceived. The estimated psychophysical similarity function falls off in a nonlinear, exponential-like fashion with respect to distance (Fig. 1F). In color space, it is also well-matched by a smoother measure that requires substantially less data, namely, the pairwise similarity ratings of colors at different distances along the color wheel using a Likert scale (Fig. 1F, S1).

bioRxiv preprint doi: https://doi.org/10.1101/325472; this version posted August 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Thus, while there are small local inhomogeneities (Fig 1D, S5), that fact is orthogonal to the argument we advance here. Our theory is based on the completely separate fact that the global structure of similarity space is deeply non-linear, in agreement with decades of work suggesting psychological similarity is globally exponential (e.g., the universal law of generalization[14,15]), with additional local confusions caused by perceptual noise[16] (measured here using a perceptual matching task; Fig 1E, F).

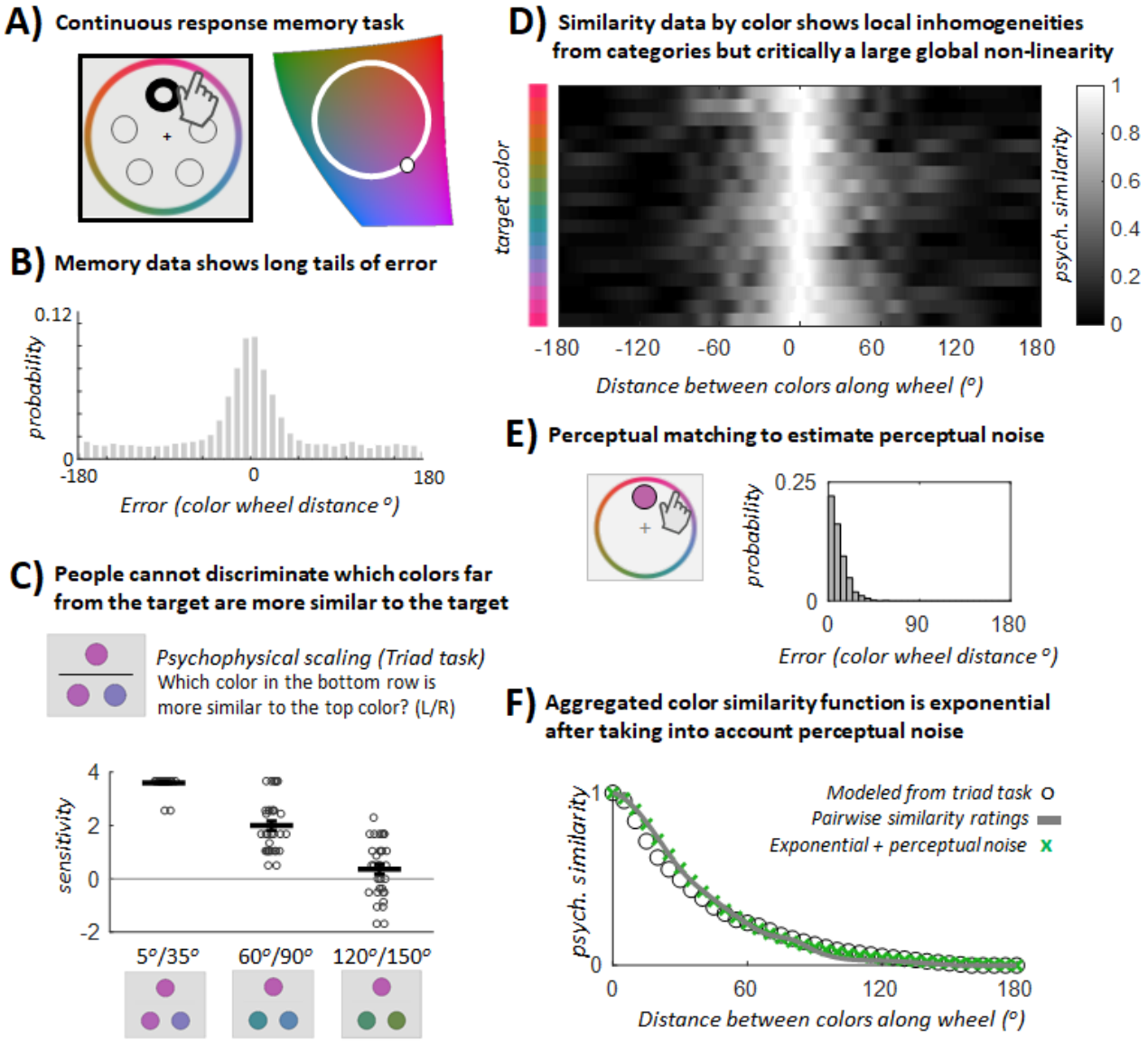

Figure 1. *(**A**) A widely used method in working memory is to select a color circle from a slice of color space, show memory items drawn from this circle, and then, at test, probe the contents of a memory item by presenting the entire continuous circle to participants to make a response. Similar response 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 is a 'long tail' of errors far from 0 that is often interpreted as evidence for distinct memory states (e.g., guesses or items encoded with very low precision). (**C**) In a triad psychophysical scaling task, participants had to say which of two colors in the bottom row was more similar to the top (target) color. Despite the difference between the two choice colors always being 30° on the color wheel, sensitivity (d') dramatically*

bioRxiv preprint doi: https://doi.org/10.1101/325472; this version posted August 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

decreased as the choices became more distant from the target. Error bars are within-subject S.E.M. and dots represent individual trials. ***(D)*** *We can use the data from another similarity task, a simple pairwise Likert rating of similarity, to infer the global psychophysical distance of colors at different physical distances along the color wheel. Here we plot this data for sets of target colors, demonstrating previously observed local non-uniformities in color space as the small differences across rows (see Bae et al.[9]).Critically, all of these row demonstrate a much larger global structure, separate from this local structure: overall similarity falls in an approximately exponential manner. This insight about the shape of the global psychological similarity function has not been taken into account by previous models or theories of visual working memory. (****E****) Some aspects of this similarity must derive from perceptual discrimination failures (e.g., there are not really 360 independent colors on the color wheel). To estimate this underlying perceptual noise, we use a continuous report task where participants must match a visible color using the same color wheel. (****F****). We can plot the global psychophysical function -- averaged over all target colors -- using the triad task or the Likert task. Both are very similar and show the same underlying shape. Consistent with previous work, we find this similarity function is exponential once perceptual noise is taken into account (e.g., an exponential convolved with the measured perceptual noise function provides an excellent fit to this data).*

A key implication of these scaling results is that the axis of error along the response wheel previously used to analyze working memory capacity does not capture the psychological representation of the stimuli. Since participants are essentially incapable in a perceptual task of discerning whether an item 120° or 180° from the target in color space is more similar to the target, it is not surprising that they confuse these colors 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, a necessary precursor to developing models of memory. In fact, by taking this scaling -- a seemingly fixed perceptual property of the stimulus space -- into account, we find working memory is accurately described by an extremely straightforward one-parameter signal detection model that treats the average memory-match signal generated by each color on the wheel as arising from its psychophysical similarity to the target.

The model we propose here is, fundamentally, the same longstanding signal detection model often used across decades of research on long-term memory and perception[17-19], modified to take into account psychological similarity. When an item is cued at test, each test option (in this case, each of the 360 colors presented on the wheel) generates a noisy, cue-dependent familiarity signal, and the color that generates the maximum familiarity signal is selected (Fig. 2). The stronger the maximum signal is, the higher the confidence in the selected color.

In a standard signal detection model of an *n*-alternative forced-choice task, it is generally assumed that exactly one item has been previously seen, so its familiarity is centered on *d′*, whereas the other *n* - 1 items are equally unfamiliar and therefore centered on zero[18]. However, when memory is tested using a continuous stimulus space, it would be implausible to assume that a color 1° away in color 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 color x on the color wheel, denoted d_x, is based on that color's measured similarity to the target, i.e., $d_x = d' f(x)$, where d' is the model's only free parameter (memory strength) and $f(x)$ is the empirically determined psychophysical similarity function. The noise added to each color is correlated between nearby colors according to the empirically measured proportion of how often colors at that distance are confused in a perceptual matching task (Fig 1E, S2).

bioRxiv preprint doi: https://doi.org/10.1101/325472; this version posted August 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Because of the nonlinear similarity function, colors in the >~90° physical distance range do not cover a great psychological expanse but instead 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 color -- say, 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 color 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 report, people theoretically report the color with maximum familiarity.

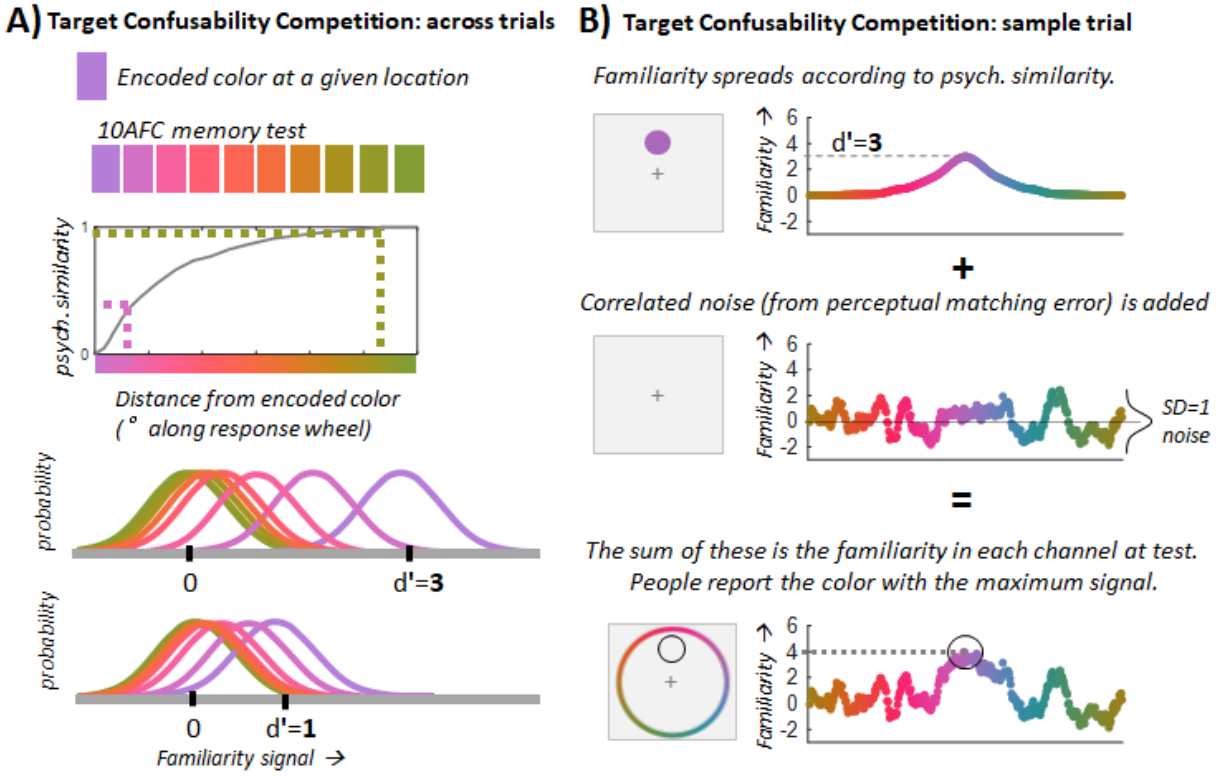

Figure 2. *(**A**) Our TCC model applied to a hypothetical 10-alternative forced-choice memory test. In standard 2-alternative long-term recognition memory experiments, "new" items sometimes feel more familiar or less familiar, and the memory-match signals generated by new items are therefore modeled as a normal distribution with a mean of 0 and standard deviation of 1. By contrast, on average, "old" items (encoded previously) elicit higher familiarity on average, modeled (in the simplest case) as a normal distribution with a mean of d' and standard deviation of 1, so that d' indicates how many standard deviations of memory strength is added to old items. In the case of a 10 alternative forced-choice, when asked to choose which of these alternatives corresponds to the old item, the simplest strategy is to sample a familiarity value from each distribution and then choose the alternative that generates the highest memory signal. To specify this model, we thus only need to specify the memory-match strength of every lure. Usually, all 9 lures are assumed to be centered on 0 when modeling such tasks[18]. 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 the target. In other words, if the target is purple, people will choose a slightly different purple lure much more often than an entirely unrelated lure (such as green in this example). The slightly different purple lure generates a much stronger familiarity signal than the green lure because of its greater similarity to the target. Examples of d'=3 and d'=1 illustrate the idea that as memory signal strength for the target decreases, more of the lure distributions cluster near the target -- and at d'=1, all of the far away colors are in a position to sometimes 'win the competition', but will do so on average equally often. 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*

bioRxiv preprint doi: https://doi.org/10.1101/325472; this version posted August 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

*colors also have correlated noise (measured using the perceptual matching function); i.e., there are not 360 independent colors on the color 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 color, with memory strength d'=3, the familiarity of purple as well as similar colors is increased (according to the measured psychophysical similarity function). Then, we add SD=1 noise -- in this case, the noise is not totally independent across color channels, but locally correlated (with this correlation empirically determined from the perceptual matching data). The resulting familiarity values, after being corrupted by noise, guide participants decisions. In a continuous report task, people simply report the color that generates the maximum familiarity value.*

Remarkably, this Target Confusability Competition (TCC) model, a straightforward signal detection model combined with measured psychophysical similarity, can explain all 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, S4). Previous cognitive models of visual working memory allow for many ways in which memory can vary (e.g., guess rate, precision, variation in precision[7,8,20]). By contrast, TCC holds that all manipulations affect only a single fundamental underlying parameter (the memory strength parameter, *d'*). 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 favor 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[7-8,20]).

While the main evidence in favor of TCC is its ability to parsimoniously characterize the effects of qualitatively different experimental manipulations (Fig. 3) and to make precise predictions across tasks and stimuli (see below), we also compared the fit provided by TCC to 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[20]. 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 subject when using metrics preferring simpler models (Table S2). This was true even though TCC fits are based on aggregated similarity functions from a different group of participants, suggesting the global structure of the psychophysical similarity function is largely a fixed aspect of a given stimulus space.

bioRxiv preprint doi: https://doi.org/10.1101/325472; this version posted August 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

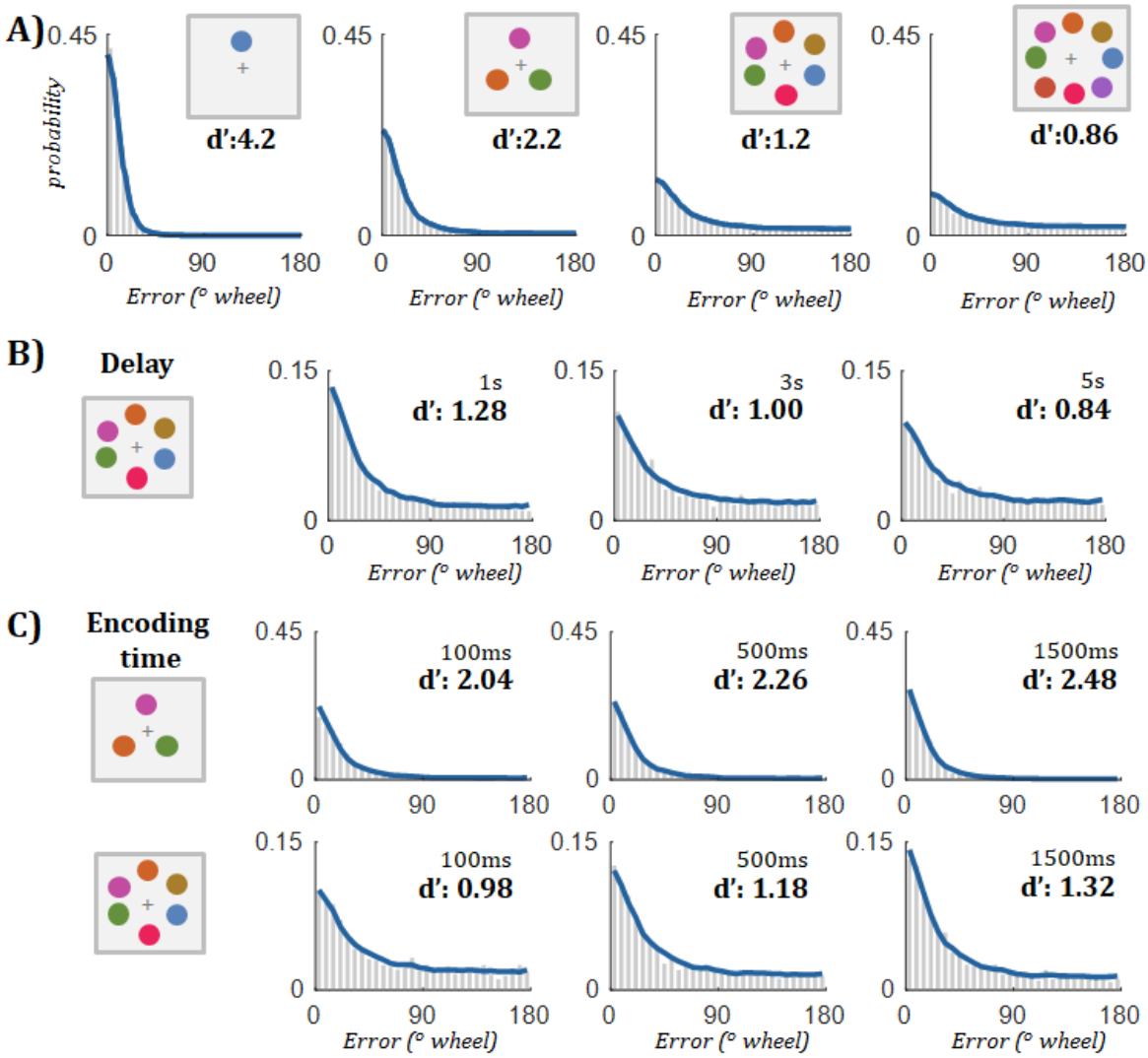

Figure 3. *(**A**) TCC fits to data at set size 1, 3, 6, and 8. Even at high set sizes, TCC -- with no concept of unrepresented items or guessing or poorly encoded items, and treating all items as encoded equally (i.e., with the same d') -- fits the data accurately because of the noisy nature of the signal detection process combined with the non-linear psychophysical similarity function. (**B**) TCC fits to data with varying delay (only set size 6 shown; remainder of data in Fig. S14). (**C**) TCC fits to data across different encoding times (only two set sizes shown; see Fig. S14). Across several key manipulations of visual working memory (set size, delay, and encoding time), which drastically alter the response distributions collected, TCC accurately captures (with only a single parameter d') the response distribution typically attributed to multiple parameters / 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 (A; set size 1: R^2=0.996, set size 3: R^2=0.986, set size 6: R^2=0.976, set size 8: R^2=0.959, , all p<0.001), delay (B; set size 1, 1s, 3s, 5s: R^2=0.997, R^2=0.995, R^2=0.987, set size 3: R^2=0.987, R^2=0.984, R^2=0.989; set size 6: R^2=0.977, R^2=0.950, R^2=0.970; all p<0.001); and encoding time (C; set size 1, 100ms, 500ms, 1.5s: R^2=0.983, R^2=0.994, R^2=0.996, set size 3: R^2=0.942, R^2=0.982, R^2=0.990; set size 6: R^2=0.950, R^2=0.987, R^2=0.980; all p<0.001).*

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 *d'* -- theoretically an interval-scale measure of memory strength[18,22] -- decreases according to a power law as set

bioRxiv preprint doi: https://doi.org/10.1101/325472; this version posted August 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

size changes (Fig S3), broadly consistent with fixed resource theories of memory[22,23]. Critically, memory strength decreases most at low set sizes (e.g., 1 to 3), suggesting limits of working memory may be best studied across lower set sizes, contrary to the majority of the field which seeks to pressure "capacity" via high set sizes to understand the nature of working memory.

TCC generates novel predictions about the connection between working memory paradigms. Ultimately, evaluating theories based on model comparisons of fit -- when all models fit the data well, as here -- is not as useful as investigating what they accurately predict[24]. TCC makes a precise and unique prediction that since all responses are generated from the same underlying process, measuring *d'* in any way -- even using only maximally different changes -- is sufficient to fully characterize memory performance. This is in stark contrast to the prediction of mixture models or variable precision models, which propose that fundamentally distinct memory states explain close-to-target responses vs. those further from the target.

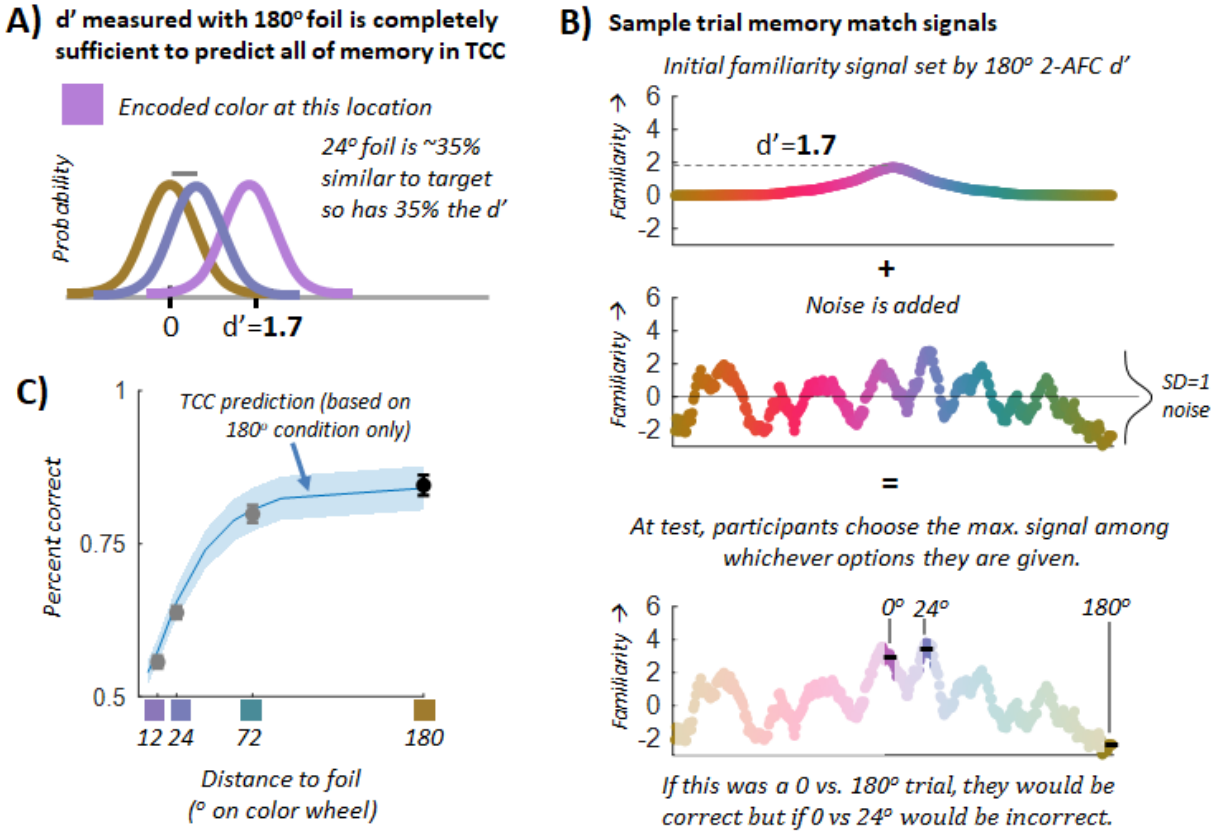

Figure 4. *(**A**) Since TCC states that visual working memory performance is determined by simply d' (memory signal strength) once psychological similarity is known, it makes novel predictions no other theory of working memory can make. In particular, it predicts that d' measured with a 180 degree, maximally distinct foil should be completely sufficient to predict all of memory performance, unlike models where errors to maximally distinct foils arise from different processes than errors to similar foils, e.g., where errors to maximally distinct foils solely from 'guessing' or extremely poorly encoded items. For example, if we've measured d' to a maximally distinct foil, we can predict that since a 24 degree foil is ~35% similar to the target, people will discriminate it with 35% of the d' relative to the 180 degree foil. (**B**) On a single trial, this prediction can be visualized in a straightforward way: If we know the target was encoded with d'=1.7, then TCC makes a strong prediction about how this familiarity spreads to other colors and how it is corrupted by noise. In continuous report, the decision rule is to report the maximum of*

bioRxiv preprint doi: https://doi.org/10.1101/325472; this version posted August 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

*the resulting color channel familiarity responses; in 2AFC, the decision rule -- based on the exact same underlying color response -- is to choose the highest familiarity signal of your response options. Thus, on this example trial, the participant would choose 0 over 180, but 24 over 0. Because TCC specifies this entire generative process, it thus makes precise predictions about how often people will make errors to different distance foils. (**C**) Here we plot the predicted percent correct of different distances of colors from the target (blue), a prediction based only on performance from the 180 degree condition (black) with no free parameters. When comparing subject's performance at different foil distances (gray) we demonstrate TCC accurately predicts performance across different foil distances.*

We evaluated this prediction in two experiments. In both experiments we had participants perform a memory task involving a 2-AFC test with maximally distinct colors (two options: 0° away from the target color vs. 180° away from the target color). We used the data from this 2-AFC task to compute *d′* in the standard way and then used TCC -- with this exact *d′* -- to compute parameter-free predictions for a variety of other conditions.

In one 2-AFC experiment (Fig. 4), we looked at how well participants could discriminate the target from more similar foils (e.g., the color they saw vs. a color only 12° away). We found that with no information other than the *d′* from maximally distinct 2-AFC comparisons, TCC correctly predicts memory performance with intermediate foil similarities (Fig 4C). More specifically, with no free parameters, TCC accurately predicts performance in intermediate conditions given 180° 2-AFC performance because it precisely quantifies how similarity impacts performance (see also Kahana & Sekuler[25]; Nosofsky[16]). By contrast, mixture models, based on the distinct concepts of guessing and precision, can use 180° 2-AFC performance only to measure 'guess rate,' leaving them largely unconstrained and able to predict a wide range of possible outcomes on other tasks, depending on the unknown factor of 'memory precision.' As a result, TCC is strongly preferred to mixture models by a Bayes factor model comparison (group Bayes factor preference for TCC > 200:1, individual subjs: $t(54)=11.19$, $p<0.001$; Fig. S9).

In a second experiment, we went further, showing that a single measure *d′* -- again measured with 2-AFC with maximally distinct 180° foils -- can predict how participants perform when there are more response options, up to and including continuous report, again with no free parameters (Figure 5). In this experiment we 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 BIC preference for TCC > 650:1, individual subjects: $t(51)=7.64$, $p<0.001$), and found that 2-AFC *d′* measured in the standard way maps directly to TCC's *d'*, which explains the full continuous report distribution. 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 it does not measure the 'precision' of memory.

bioRxiv preprint doi: https://doi.org/10.1101/325472; this version posted August 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

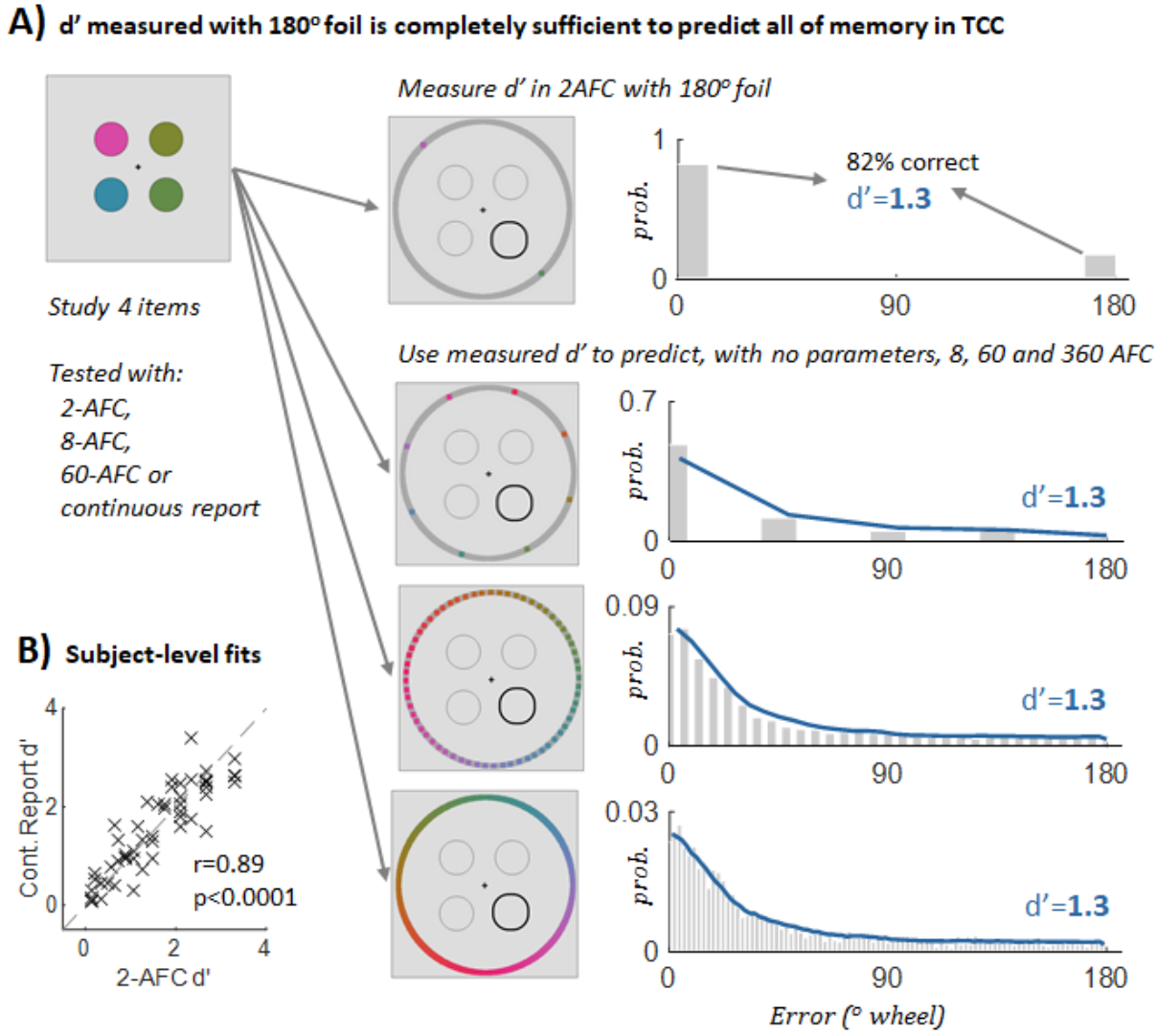

Figure 5. *(**A**) According to TCC, the d' in a 2AFC task is fundamentally the same d' in continuous report tasks (or any other AFC task). Thus, unlike other models, TCC makes a strong prediction that d' as measured with a 180 degree foil is completely sufficient to predict all of memory across any number of options presented at test, including 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 assess this, participants studied 4 items and were then tested with a 2-AFC, 8-AFC, 60-AFC or continuous report (360-AFC). During 2AFC trials, the foil was always 180 degrees away, which we used to calculate d'. We then used this measured d' to predict, with no parameters, 8-, 60-, and 360-AFC performance via TCC. This provides further evidence 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**) Taking only the 2-AFC data and the continuous report data as an example, we can also independently estimate d' from the data from these two tasks. We find a strong subject-level correspondence between TCC's continuous-report estimate of d' and d' estimated from the 2-AFC task in the traditional way, in line with what is expected simply from the noise ceiling of these measurements.*

bioRxiv preprint doi: https://doi.org/10.1101/325472; this version posted August 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Thus, with TCC, measuring only how well participants can distinguish between far apart test items (0° vs.180°) using a 2-AFC task is sufficient to completely 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 ceiling or floor. Together, these experiments provide strong evidence against previous models of visual working memory where the tails of the continuous-report distribution (the only aspect of performance measured with 180° foils in 2-AFC) are fundamentally distinct from the center of the distribution (e.g., in the competing models responses in the tail result from 'guesses' or 'low precision', whereas the central responses result from high precision memories). In addition, this discovery allows the reintegration of a huge literature on change detection with very distinct foils, with important theoretical and clinical implications[26], as it shows that measuring *d′* with maximally distinct foils is sufficient to fully understand memory response distributions.

Generalization across different stimulus spaces. So far we have focused largely on color space, which is the dominant way visual working memory is studied[7]. However, TCC is not limited to color and can be applied to any stimulus space. To show this 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[27], we collected memory data for set size 1 and 3. We also measured the psychophysical similarity function and measured the accuracy of perceptual matching on this face space (Fig. 6). Again, we found the TCC fit observed memory data remarkably well across both set sizes 1 and 3 (see Fig. 6) and fit reliably better than existing mixture models (Table S3).

Thus, TCC accounts for data across a variety of stimulus spaces: as long as the perceptual similarity space of the stimuli is measured using psychophysical scaling, 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 previously collected visual long-term memory continuous report task with colors[28]. Participants performed blocks where they sequentially saw 40 real-world objects' that were randomly colored, and then after a delay, were shown each object one at a time in grayscale and reported both whether the object was old or new and the color of the object using a color wheel (as in Brady et al.[29]). Some items were seen only once, and some repeated twice in the same color within a block (Fig. 6C). We fit color reports for previously seen items that were reported as having been seen (hits; see Methods). Again, we found that TCC fit the observed memory data remarkably well across both the unrepeated and repeated items (Fig. 6D). Thus, unlike working memory modeling frameworks which propose system-specific mechanisms (e.g., population coding combined with divisive normalization[22]), 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.

bioRxiv preprint doi: https://doi.org/10.1101/325472; this version posted August 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

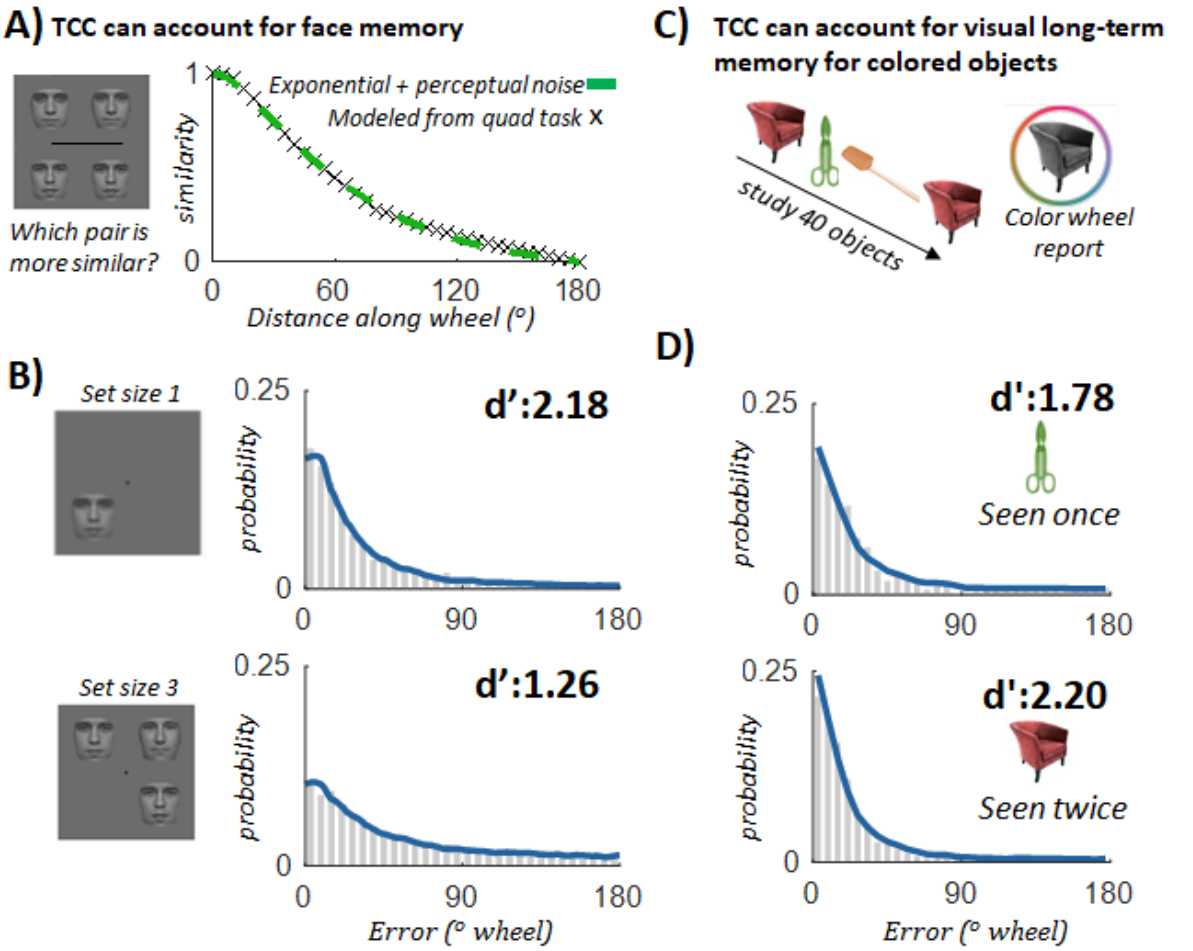

Figure 6. *(**A**) Using a 'quad' similarity task to reduce relational encoding, and the same MLDS method and perceptual matching task as with color, we collected a psychophysical distance function for face identity. (**B**) TCC fits to working memory data using face identity at set size 1 and 3 (R^2=0.993, p<0.001; R^2=0.970; p<0.001). TCC accurately captures face identity data, demonstrating its generalizability across diverse stimulus spaces. **(C)** To show generalization to other memory systems, we fit data on a visual long-term memory continuous report task with colors[28]. Participants performed blocks of memorizing 40 items, and then after a delay, reported the colors of the items using a color wheel. Some items were seen only once, and some repeated twice in the same color within a block. **(D)** TCC fits to visual long-term memory data for 'old' items reported as 'old', for items seen only once and for items repeated twice (R^2=0.96, p<0.001; R^2=0.98; p<0.001). TCC accurately captures visual long-term memory data, suggesting the psychological similarity function is a constraint on both working and long-term memory systems. Note that long-term memory performance in this task likely depends on a two-part decision -- item memory and source memory (e.g., the object itself, and then its color). This two-part decision is related to the processes of recollection and familiarity, but we do not address this here because we focus on understanding color reports only to items judged as 'old' (i.e., conceivably where item memory strength was strong).*

Implications of TCC: no objective guessing. One particularly important implication of TCC's fit to the data is that there is little-to-no objective "guessing" in working memory, and so its strong fit to the data provides evidence against a fixed capacity limit where participants only remember ~3-4 items[1,2] and consistent with more continuous conceptions of working memory[4]. That is, while colors far from the target in color space sometimes 'win the competition', this is not because the target was fundamentally unrepresented or varied in encoded memory strength

bioRxiv preprint doi: https://doi.org/10.1101/325472; this version posted August 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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, the target is always represented -- that is, people's familiarity signals are never unaffected by what they just saw 1 second ago (as in d'=0).

While these conclusions follow from the excellent fits of the straightforward 1-parameter model to a wide variety of data and the generalization from maximally-distinct 2-AFC to other conditions, to evaluate this claim further, we assessed a 2-parameter hybrid model based on TCC but mixed with objective 'guessing'. This hybrid model assumes only a subset of items are represented and that the remainder have d'=0. Focusing on the highest set sizes (6 and 8), we found such a model was dispreferred in model comparisons in 100% of subjects sizes compared to TCC (mean BIC: -6.3, SEM: 0.20; p<0.0001). Furthermore, while this hybrid model accurately recovered its own parameters from simulated hybrid data, showing it detects objective 'guessing' if it is present (Fig. 7C), when fit it to empirical data it estimates 'guessing' rates near 0 in every set size in group data (Fig. 7B), and a guess rate <5% in the majority of individual subjects 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 4-5 of the 8 items are always entirely unrepresented[2]. This demonstrates it is possible to detect "random guesses" if present in the data, but TCC finds no evidence for such objective 'guessing'. Critically, like any standard signal detection model, TCC naturally accounts for the subjective feeling of guessing/low confidence[18] (Fig. S10, S11).

bioRxiv preprint doi: https://doi.org/10.1101/325472; this version posted August 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

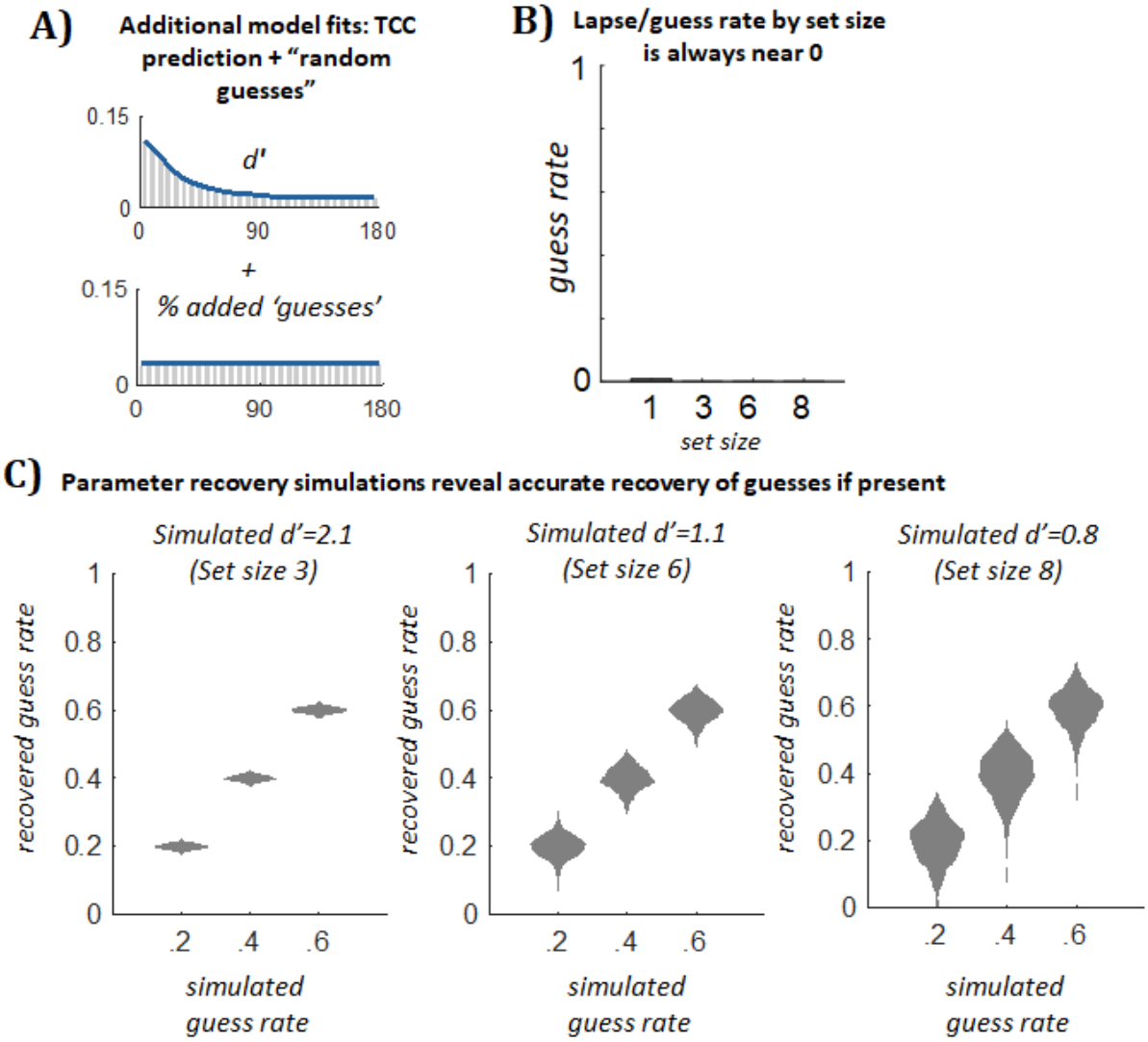

Figure 7. *(**A**) To validate whether TCC could detect objective guessing (i.e. a separate psychological state with no information) if present in the data, we considered a mixture of responses from TCC plus objective guessing, creating a mixture model of TCC and a uniform distribution. **(B**) Although model comparison strongly preferred TCC with no guessing, we nevertheless fit a hybrid TCC+guessing model (2 parameters) to real participant data, and 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 20% 'guesses' added to set size 8 d' levels is accurately recovered and never estimated as 0 (data are violin plots, showing entire distribution of recovered parameters). Furthermore, model comparison metrics -- even those, like BIC, designed to prefer simpler models -- prefer the hybrid model with the guessing parameter in every simulation with guessing added (all BIC>30:1 in favor of hybrid model). This provides strong evidence there is not objective guessing in visual working memory data, and that our modeling with TCC would be able to detect this guessing if it was present.*

Another critical distinction between signal detection models and those that predict a number of items which are lost completely or nearly completely[7] is that TCC predicts a curvilinear receiver operating characteristic (ROCs; as all signal detection models do[18]), whereas models where some items are unrepresented, as previous mixture models claim, predict a linear relationship between hit and false alarm rates when foils are very distinct[30] (Fig. 8). This is because

bioRxiv preprint doi: https://doi.org/10.1101/325472; this version posted August 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

threshold models predict that represented items contribute only to hit rates, as no represented item is ever so-imprecise as to cause people to pick a 180° foil over a target. In contrast, TCC makes a clear prediction that ROCs should be curvilinear because all items are stored with noise (Fig. 8B).

To generate ROC data, the rating method (based on confidence) is preferable to alternative methods that seek to manipulate response bias across conditions (e.g., by varying payoffs or signal-presentation probability), which often yield a limited range of data and may also change the nature of the underlying internal signal across points[31]. Thus, we asked participants to detect whether a given color is old or new (change detection) and give their confidence, again with very distinct foils (180 degrees).

We find clear evidence in favor of a signal detection process underlying working memory decisions: ROCs are reliably curvilinear (Fig. 8C) and z-ROCs at the highest set size do not show the positive quadratic component predicted by high-threshold models[32] (set size 6: mean z-ROC quadratic component: -0.13, SEM 0.113,$t(52)=1.28$, $p=0.21$). In addition, the ROCs are nearly symmetric, consistent with the idea that all items are represented with approximately the same underlying memory strength. Unequal memory strength between items results in z-ROC slopes below 1.0, but we find these slopes are very close to 1.0 even at set size 6 (z-ROC slopes: set size 3, 1.06, SEM 0.18, set size 6: 0.96, SEM 0.04). Together, the ROC analyses provide further evidence that all items are represented with noise and with approximately the same memory strength. This provides evidence for TCC's claim that most of the variation in response precision arises because of the stochastic noise characteristic of a signal detection process, not because some items are unrepresented[7] or because items vary widely in encoding strength[8].

bioRxiv preprint doi: https://doi.org/10.1101/325472; this version posted August 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

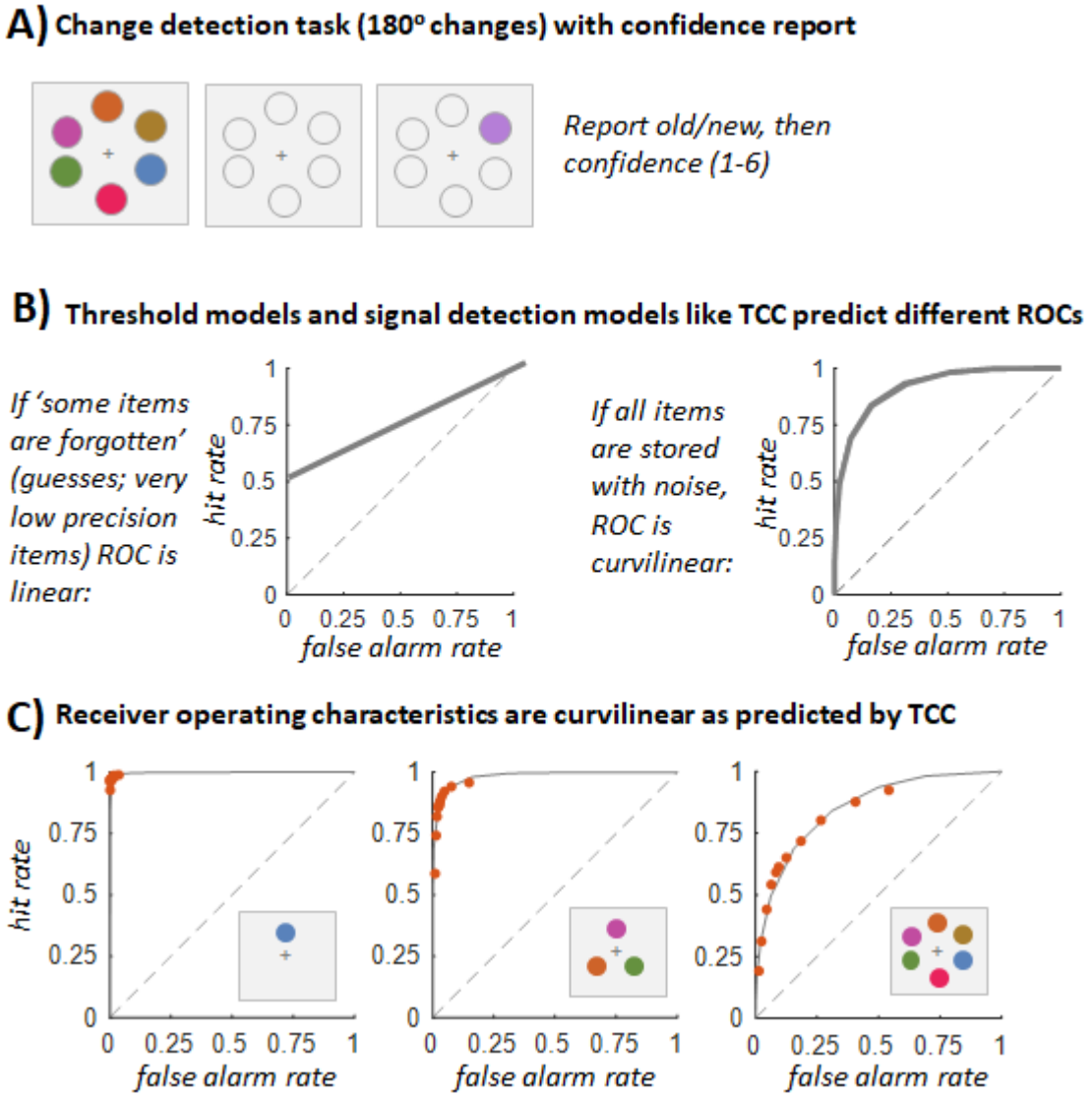

Figure 8. *(**A**) Participants completed a change detection task with 180 degree changes. After reporting whether the test item was old or new, participants then reported the confidence of their decisions on a 1-6 scale (1 = no confidence, 6 = extremely confident). (**B**) We can use the confidence data to create receiver operating characteristics (ROC curves), in which signal detection theory makes a clear prediction contrary to several existing theories of working memory. For example, a threshold (slot or continuous resource) model of working memory predicts ROC curves should be linear, as if some items are forgotten (via guesses or having very low precision), only remembered items can add to hits during 180 degree changes. In contrast, signal detection theory dictates all items are stored with some noise, and thus the ROC should be curvilinear – and if all items are represented with approximately equal d', the curves should also be symmetric. (**C**) ROC curves are clearly curvilinear across all set sizes (including set size 1), as predicted by TCC. These ROC curves are approximately symmetric, providing further support for the idea that all items are represented with similar d' as opposed to items varying significantly in d' from each other (which creates an unequal variance ROC).*

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

bioRxiv preprint doi: https://doi.org/10.1101/325472; this version posted August 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

models has led to what seem to be illusory independent estimates of '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[36-38], and even about the nature of consciousness[33] have been made based on dissociations between model-based estimates of 'precision' and 'capacity'. However, we have shown these apparent dissociations are an illusion of modeling the data without taking into account the structure of the stimulus space (Fig. S6-8). When the global similarity structure of stimuli 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[33-38].

The role for variability between items in the TCC.

While the TCC rejects the idea that the distribution of responses collected from continuous report is explained primarily by remembered and not-remembered items (or items that are encoded with extremely different precisions[8]), this does not mean some variability between items is not present in working memory tasks. Psychophysical scaling naturally accounts for many stimulus-specific variability effects (Fig. S5). 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, such as the effect of varying confidence on continuous report data or allowing participants to choose their best item for report (Fig.S10, S11). Thus, it remains an open question to what extent d' varies between items and trials, with the relatively-equal variance ROC data at set size 6 suggesting that there is likely surprisingly little variation in memory strength between items even at set size 6. Critically, 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 likely does not include an appreciable role for items with d' = 0.

Potential mechanisms

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[39] and single-unit recording[40].

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[15] or population coding[22]. 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 color space (e.g., 15° on the color wheel), there would be extremely limited overlap in the population of neurons that code for any two colors even a medium distance apart on the wheel.

Thus, the current model is in many ways related to existing models of working memory based on population codes[22,41]. Indeed, the similarities between the framework of population coding and the cognitive model proposed here offers significant promise for bridging across levels of understanding in neuroscience, with a population coding implementations of TCC possible[42,43]. However, as compared to existing population-based models[22], the cognitive basis of the current model -- with the measured scaling function following the well-known cognitive laws of

bioRxiv preprint doi: https://doi.org/10.1101/325472; this version posted August 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

similarity[14,16] -- allows us to fit data with an extremely simple 1-parameter model that 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 to population coding models where the decision process is based on variability in spikes[22], which are hard to reconcile with data from high-level stimuli like faces and data from long-term memory.

Previous work has shown psychophysical similarity metrics are likely distinct for different stimuli in the same stimulus space (e.g., memory varies across colors[9-10]; see Fig S5). The underlying space upon which the exponential similarity function is imposed may be designed to take advantage of efficient coding of environmental regularities[44], such that the more frequent the stimuli, the more neural resources we devote, giving improved discriminability and predictable memory biases[45]. 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, S5). In addition, psychophysical similarity may not be a fixed property but may be dependent on how the current environment affects discriminability[46,47]. For example, memory biases are altered when discriminability is affected by adaptation or contextual effects[45].

Signal detection, guessing, and connections to long-term memory.
A key component of previous mixture model frameworks has been that observers are either in a "memory state," with relatively high-precision information available, or they are in a "guess state", with either no information or (extremely) low information available[33-38]. Under signal detection theory, there is no distinction between "memory" and "guess" states as previously theorized. Instead, both high and low confidence and high and low error responses arise naturally through the same underlying process. Subjectively, guessing still occurs, but it occurs when the information upon which the decision is made is non-diagnostic, not when task-relevant information is absent (Fig S11).

The distinction is subtle but critical. In TCC, when d' is low, all of the stimuli on the response wheel may yield similarly low memory-match signals. Because the signals are largely non-diagnostic, every stimulus is almost equally likely to yield the strongest signal upon which the decision is based. Thus, responding will be essentially random, and decisions will feel like guessing. Nevertheless, such decisions are still based on information (namely, the strongest memory signal, weak as it may be) and are therefore not random guesses akin to a coin flip in the absence of information. In addition, TCC (like all signal detection models) is based fundamentally on the idea of sampling (and potentially re-sampling[48]) as the basis of memory judgments, providing an alternative explanation for many phenomena in memory that have been taken as evidence of a high-threshold view where people are assumed to make many all-or-none guesses.

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, like the relationship between confidence and accuracy[49,50] (Fig. S10) 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'(Fougnie et al. in revision). 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

bioRxiv preprint doi: https://doi.org/10.1101/325472; this version posted August 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

both domains by emphasizing that competition and perceptual confusability between items is a major limiting factor across both working memory and long-term memory.

bioRxiv preprint doi: https://doi.org/10.1101/325472; this version posted August 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Methods

All 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 color experiments used a circle in CIE *L*a*b** color space, centered in the color space at (*L* = 54, *a* = 21.5, *b* = 11.5) with a radius of 49. Some experiments were run in the lab, while others were conducted using Amazon Mechanical Turk. Mechanical Turk users form a representative subset of adults in the United States[51], and data from Turk are known to closely match data from the lab on visual cognition tasks[52,53], including providing extremely reliable and high-agreement on color report data[54]. Any systematic differences between the lab studies – where we collect most memory data – and the Turk studies – where we collect most similarity data – would decrease the appropriateness of the similarity function for fitting the memory data, hurting the fit of TCC.

1. Fixed distance triad experiment. N=40 participants on Amazon Mechanical Turk judged which of two colors presented were more similar to a target color. The target color was chosen randomly from 360 color values that were evenly distributed along a circle in the CIE *L*a*b** color space, as described above. The pairs of colors were chosen to be 30 degrees apart from one another, with the offset of the closest color to the target being chosen with an offset (in deg) of either 0, 5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 120, 150 (e.g., in the 150 degree offset condition, the two choice colors were 150 and 180 degrees away from the target color; in the 0 deg offset condition, one choice exactly matched the target and the other was 30 deg away).

Participants were asked to make their judgments 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 10 repeats of each of the 13 offset conditions, each with a different distance to the closest choice color to the target, and trials were conducted in a random order. The trials were not speeded, and the colors 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 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. The data from a subset of offset conditions is plotted in Figure 1C, and the full dataset is plot in Figure S1.

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

Participants were asked to make their judgments 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 3 repeats of each of the possible pairs of offset conditions, and trials were conducted in a random order. The trials were not speeded, and the colors remained visible until participants chose an option. Following our a

bioRxiv preprint doi: https://doi.org/10.1101/325472; this version posted August 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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.

To compute psychophysical similarity from these data, we used the model proposed by Maloney and Yang[13], the Maximum Likelihood Difference Scaling method. This method can be adapted to measure the appropriate psychophysical scaling of similarity between colors as a function of the distance between colors along the wheel. In particular, any given trial has a target color, S_i, and two options for which is more similar, S_j and $S_{k,}$. Let $I_{ij} = S_j - S_i$, the length of the interval between S_i and S_j on the color wheel, which is always in the set [0,3,5,...180], and ψ_{ij}, the psychophysical similarity to which this distance corresponds. If people made decisions without noise then they should pick item j if and only if $\psi_{ij} > \psi_{ik}$. We add noise by assuming participants decisions are affected by Gaussian error, such that they pick 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 has 20 free parameters, corresponding to the similarity scaling values for each possible distance length (e.g., how similar a distance of 5 or 10 on the color wheel really is 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 predict observers' similarity judgments, in that equal intervals in the scaled space are discriminated with equal performance. Once the scaling is estimated, we normalize the psychophysical scaling parameters so that psychophysical similarity ranges from 0 to 1.

3. Likert color similarity experiment. N=50 participants on Mechanical Turk judged the similarity of two colors presented simultaneously on a Likert scale, ranging from 1 (least similar) to 7 (most similar). The colors were chosen from a wheel consisting of 360 color values that were evenly distributed along the response circle in the CIE *L*a*b** color space. The pairs of colors were chosen by first generating a random start color from the wheel and then choosing an offset (in degrees) to the second color, from the set 0, 5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 120, 150, 180. Participants were given instructions by showing them two examples: (1) in example 1, the two colors were identical (0 deg apart on the color wheel) and participants were told they should give trials like this a 7; (2) in example 2, the two colors were maximally dissimilar (180 deg apart on the color wheel) and participants were told they should give this trial a 1. No information was given about how to treat intermediate trials. Participants were asked to make their judgments 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 did 140 trials, including 10 repeats of each of the 14 offset conditions, each with a different starting color, and trials were conducted in a random order. The trials were not speeded, and the colors remained visible until participants chose an option. 2 participants were excluded for failing a manipulation check (requiring similarity >6 for trials where the colors were identical). Based on an a priori exclusion rule, we excluded trials with reaction times <200ms or >5000ms, which accounted for 3.0% (SEM:0.4%) of trials. Similarity between two colors separated by $x°$ was measured using a 7-point Likert scale, where $S_{min} = 1$ and $S_{max} = 7$. To generate the psychophysical similarity function, we simply normalize this data to range from 0 to 1, giving a psychophysical similarity metric, such that $f(x) = ((S_x - S_{min}) / (S_{max} - S_{min}))$.

4. Perceptual matching experiment. N=40 participants on Mechanical Turk were shown a color and had to match this color, either using a continuous report color wheel (100 trials) or choosing among 60 options (100 trials; spaced 6 degrees apart on the color wheel, always including the target color). The 60-AFC version was designed to limit the contribution of motor noise, since the colors in this condition were spaced apart and presented as discrete boxes that

bioRxiv preprint doi: https://doi.org/10.1101/325472; this version posted August 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

could not easily be 'misclicked'. Colors were generated using the same color wheel as other experiments, and participants had unlimited time had to choose the matching color. The color and color wheel/response options remained continuously visible until participants clicked to lock in their answer. The color was presented at one of 4 locations centered around fixation (randomly), approximately matching the distance to the color wheel and variation in position used in the continuous report memory experiments. 1 participant's data was lost due to experimenter error and 2 participants were excluded for an average error rate greater than 2 standard deviations away from the mean.

To convert this data into a perceptual correlation matrix -- asking how likely participants are to confuse a color *x* degrees away in a perception experiment -- we relied upon the 60-AFC data alone, since this data has no contribution from motor noise and so is solely a measure of perceptual noise. However, using the continuous report data instead result 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 degree errors (-60, -54… 0, … 54, 60), and then linearly interpolated between these to get a value of the confusability for each degree of distance. We then normalized this to range from 0 to 1.

5. Modeling Data Using the Target Confusability Competition (TCC) Model. 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 alternative forced-choice 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=±180°, 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 color, (1) each of the colors on the color 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 color x has the maximum m_x, (3) the mean of the memory-match signal for each color, d_x, is determined by its psychophysical similarity to the target according to the measured function ($f(x)$), such that $d_x = d' f(x)$ (Figure 2) and the noise is correlated across nearby colors according to confusability in a perceptual matching task. As $f(x)$, the psychophysical similarity function, we use the smooth function estimated from the Likert similarity experiment although the triad task modeled similarity function predicts fundamentally the same results (Fig. S4).

According to the model, the mean memory-match signal for a given color 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 color on the wheel generates a memory-match signal, m_x, conceptualized as a random draw from that color's distribution centered on d_x. That is, if the noise was uncorrelated between nearby colors, $m_x \sim N(d_x, 1)$. The response (r) on a given trial is made to the color on the wheel that generates the maximum memory-match signal, $r = argmax(m)$.

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-is (see Fig. S2), but as specified so far, this model assumes that 360 independent color probes elicit independently noisy memory-match signals. The shape of

bioRxiv preprint doi: https://doi.org/10.1101/325472; this version posted August 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

the distributions the model predicts are effectively independent of how many color 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 color channels used, and so to make the d' value in TCC comparable to real signal detection *d′* values, it is important to consider "how many" color channels people are accessing.

Rather than make this a discrete decision (e.g., people consider 30 channels), we instead estimated the covariance between nearby channels in a continuous manner. The familiarity value of color 181 and 182 on the wheel cannot possibly be fully independent, since these two colors are perceptually indistinguishable. Following this intuition, we make a simple assumption: the amount of shared variance in the noise between any two color channels is simply how often colors at that distance are confused in a perceptual matching task. Thus, *p(x)*, the correlation in the noise between any two colors *x* apart on the color wheel, is given by C_x / C_0, where C_x is how often colors *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 4). On average, colors 1 degree away are chosen about 96% as often as the correct color in the matching task, so the noise between any two channels 1 degree apart is assumed to share 96% of its variance; 82% at 5 degrees; etc. Thus, having measured both the similarity function and the perceptual matching matrix, to sample from the full (correlated-noise) TCC model, we can use MATLAB code that is nearly as straightforward as 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 color, 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*'s [-179,180]. The covariance between colors (*R*) is given by the perceptual confusability of colors at that distance, *p(x),* which is also taken to be symmetric:

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

To use the perceptual correlation data as the covariance in the correlated model, because it might not always be a perfect correlation matrix (e.g., not perfectly symmetric, 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 symmetric until it was a valid correlation matrix. This resulted in only minimal changes compared to 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;* i.e., subjective confidence) and reported color value, *r,* are then the max and argmax, respectively, of this vector:

bioRxiv preprint doi: https://doi.org/10.1101/325472; this version posted August 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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

And the error, *e*, is the circular distance from *r* to *0.* The distribution of *m* is in theory directly computable[55], 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.[56] 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, e.g., the response on a given trial, rather than being $\text{argmax}(X_{-179}, \ldots, X_{180})$, likely includes motor noise of some small amount, for example, 2°:

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

In the model fitting reported in the present paper, we include a fixed normally distributed motor error with SD=2°, although we found the results are not importantly different if we do 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[57] making use of maximum likelihood. For fits of the correlated model -- which 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 get 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 Supplement S2). Thus, it is also possible to fit the uncorrelated version through maximum likelihood with the appropriate adjustment to *d'*, and doing so results in the same fits.

6. Continuous color report data (set size 1, 3 and 6, 8). The continuous color report data used for fitting the model was collected in the lab to allow a larger number of trials per participant. N = 20 participants performed 100 trials of a memory experiment at each of set size 1, 3, 6 and 8, for a total of 400 trials (plus 4 practice trials). The display consisted of 8 placeholder circles. Colors were then presented for 1000ms, followed by an 800ms ISI. For set sizes below 8, the colors appeared at random locations with placeholders in place for any remaining locations (e.g. at set size 3, the colors appeared at 3 random locations with placeholders remaining in the other 5 locations). Colors were constrained to be at least 15° apart in color space along the response wheel. After the ISI, a target item was probed by marking a placeholder circle was marked with a thicker outline, and participants were asked to respond on a continuous color wheel to indicate what color had been presented at that location. Error was calculated as the number of degrees on the color wheel between the probed item and the response.

7. Continuous report memory as a function of delay (set size 1, 3, 6). N = 20 participants in-lab completed a color working memory task similar to the previous high set size experiment, but with the following exceptions. Participants performed 12 blocks of 75 trials (900 trials total). Each block contained an equal number of trials at set size 1, 3, and 6. The display consisted of

bioRxiv preprint doi: https://doi.org/10.1101/325472; this version posted August 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

6 placeholder circles. Colors were presented for 500ms, and followed by a delay of either 1000ms, 3000ms, or 5000ms. Delay time was blocked, and participants were informed at the beginning of each block the delay time for that block. Each combination of the 3 set sizes and the 3 delays was used in 100 trials.

8. Continuous report memory as a function of encoding time (set size 1, 3, 6). N = 20 participants in-lab completed a color 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 size 1, 3, and 6. Colors were presented for either 100ms, 500ms, or 1500ms. Encoding time was blocked, and participants were informed at the beginning of each block the encoding time for that block. Following encoding, there was a 1000ms delay before a target item was probed. Each combination of the 3 set sizes and the 3 encoding times was used in 100 trials.

9. Model comparisons to mixture models. For all model comparisons in the paper, we created new versions of mixture models designed to be directly comparable to TCC. In particular, to make predictions derived from mixture models comparable to those derived from TCC (which specifies a probability of response discretely for each 1 degree of the wheel, not over a continuous distribution), we use discrete versions of the 2-parameter and 3-parameter mixture models in which the probabilities of the data are normalized over each of 360 possible integer error values (not over the continuous space of errors).

We performed two types of model comparisons: 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[58]. Specifically, we fit the TCC and the 2- parameter and variable precision mixture models to data from each set of N-1 trials separately for each subject 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 subjects separately for each set size. TCC and mixture models provided relatively comparable fits (see Table S2), which is to be expected given the mixture model can almost perfectly accurately mimic TCC (see Fig. S6) and given that the amount of data used to fit the models is much greater than the number of parameters in either model (which ranges from 1-3), so cross-validation provides effectively no penalty for complexity.

We then compared how well the competing models (TCC; 2-parameter mixture model; 3-parameter variable precision mixture model) fit data from individual participants for the color 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 (Table S2). Note that this was true even though TCC fits are based on aggregated similarity functions from a different group of participants, suggesting the global structure of the psychophysical similarity function is largely a fixed aspect of a given stimulus space. Ideally, TCC would be fit with a similarity function specific to each individual target color (which can be done and predicts the appropriate deviations; see Fig. S5), 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 rely 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.

10. 2-AFC at different foil similarities. N=60 participants on Mechanical Turk completed 200 trials of a 4-item working memory task. On each trial, they saw 4 colors randomly chosen from

bioRxiv preprint doi: https://doi.org/10.1101/325472; this version posted August 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

the color wheel (subject to the constraint that no two colors were within 15 deg. of each other). The colors were presented for 1000ms and then after an 800ms delay, had to answer a 2-AFC memory probe about one of the colors. The foil color in the 2-AFC could be offset from the target 180, 72, 24, or 12 degrees (50 trials/condition). These conditions were interleaved so that participants needed to maintain detailed memories of the color on every trial, since conceivably if only 180 degree 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 color wheel -- e.g., the 180 degree foils were presented 180 deg. apart on the screen, and the 12 deg. foils were presented 12 deg. apart on the screen, to visually indicate the distance between the target and foil in color space.

Performance was scored as the number correct out of 50 at each offset of the memory foil. 5 participants were excluded for below chance performance in the maximally easy 180 deg. offset condition, leaving N=55 participants.

In order to assess the predictions of TCC for this data in a way amenable to the use of Bayes factors, we took the number correct out of 50 in the 180 deg. foil condition and used this to calculate a distribution over d' values (e.g., 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, e.g., for 24 deg. foils, in MATLAB notation:

memoryMatchStrengths = mvnrnd(similarityFunction * $dprime_{180}$, percepCorrMatrix, 50);
$isCorrect=memoryMatchStrengths_{0deg}>memoryMatchStrengths_{24deg}$

In other words, to assess performance in the 24 deg. offset condition, we assumed responses were generated according to the argmax of only these 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 deg. 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 subjects' performance in the other conditions given their 180 deg. foil performance in TCC.

To assess the likelihood of performance at different offsets in the mixture model framework of Zhang and Luck[7], we use performance at the 180 deg. foil conditions to assess the "guess rate" of participants (guess rate = 1 - ($2*percentCorrect_{180}$-1)) in the standard way (e.g., Brady et al. 2008). However, in this framework, 180 deg. foils leave an unknown free parameter: memory precision cannot be assessed using such foils, and thus is free to vary. Thus, to predict a 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[57]). 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

bioRxiv preprint doi: https://doi.org/10.1101/325472; this version posted August 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

of 0.01 in d', and we assessed the summed log likelihood of each of the 3 other offsets (e.g., not including the 180 deg. condition) as our final data likelihood.

11. 2-AFC generalization to n-AFC and continuous report. N=60 participants on Mechanical Turk completed 200 trials of a 4-item working memory task. On each trial, they saw 4 colors randomly chosen from the color wheel (subject to the constraint that no two colors were within 15 deg. of each other). The colors were presented for 1000ms and then after an 800ms delay, had to answer a probe about one of the colors. This probe could be a 2-AFC (with 180 deg. different foil), an 8-AFC (with the choices equally spaced around the color 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 color on every trial, since conceivably if only 180 degree 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 color wheel -- e.g., the 2-AFC 180 degree foils were presented 180 deg. apart on the screen, and the 60-AFC foils deg. foils were presented 6 deg. apart on the screen, to visually indicate the distance between the target and foils in color space.

Performance was scored as the number correct out of 50 at each offset of the memory foil. One participant's data was lost, and 7 participants were excluded for below chance performance in the maximally easy 2-AFC, 180 deg. offset condition, leaving N=52 participants.

The simplest metric is simply to compare the d' computed from 2-AFC performance (e.g., (norminv(hit)-norminv(fa))/sqrt(2)) to the d' from fitting TCC to the continuous report data. These are identical up to the level of the noise in the fits (Fig. 5B).

In order to assess the predictions of TCC for this 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 deg. foil condition and used this to calculate a distribution over d' values (e.g., 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 350-AFC) straightforwardly, simply first choosing the maximum out of the relevant foil options that are present, e.g., 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 deg. 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 allows us to understand the likelihood of each subjects' performance in the other conditions given their 180 deg. 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 use performance at the 180 deg. foil conditions to assess the "guess rate" of participants (guess rate = 1 - (2*percentCorrect$_{180}$-1)) in the standard way (e.g., Brady et al.[59]). However, in this framework, 180 deg. foils leave an unknown free parameter: memory precision cannot be assessed using such foils, and thus is free to vary. Thus, to predict a 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',

bioRxiv preprint doi: https://doi.org/10.1101/325472; this version posted August 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

we then calculated the likelihood of subject's continuous report performance under these parameters. 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'. 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 deg. condition.

12. Old/new and confidence experiment to compute ROCs. N = 70 in-lab participants performed 300 trials of a change detection task, 100 at set size 1, 100 at set size 3, and 100 at set size 6. The display consisted of 6 placeholder circles. Colors were then presented for 500ms, followed by an 1000ms ISI. For set sizes below 6, the colors appeared at random locations with placeholders in place for any remaining locations (e.g. at set size 3, the colors appeared at 3 random locations with placeholders remaining in the other 3 locations). Colors were constrained to be at least 15° apart along the response wheel. After the ISI, a single color reappeared at one of the positions where an item had been presented. On 50% of trials each set size, this was the same color that had previously appeared at that position. On 50% of trials, it was a color from the exact opposite side of the color wheel, 180° along the color wheel from the shown color at that position. Participants' task was to indicate whether the color that reappeared was the same or different than the color that had initially been presented at that location. After indicating whether the color was the same or different from the target in the previous array, participants then reported their confidence. Participants were presented an interval from 1-6 and had been instructed that 1 meant "very unsure" and 6 meant "very sure" and to report their confidence using the entire scale. 3 participants were excluded for an overall percent correct >2 standard deviations below the mean, leaving a final sample of N=67.

13. Face identity continuous report data (set size 1 and 3). 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 in Haberman, Brady and Alvarez[27]. We used set sizes 1 and 3, and the encoding display was shown for 1.5 seconds 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 and not included in the analysis. 14 participants were excluded for having near-chance performance levels (d'<0.50) at set size 3, although including all subjects with d'>=0 does not affect our conclusions or the fit of TCC.

14. Face identity similarity 'quad' task. N=102 participants on Mechanical Turk judged which of two pairs of faces presented were more distinct (e.g., which pair had constituent items that were more different from each other). On each trial, we chose 2 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.

Participants were asked to make their judgments 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 2 standard deviations below the mean, resulting in a final sample of N=85.

To compute psychophysical distance from these data, we used the model proposed by Maloney and Yang[13], the Maximum Likelihood Difference Scaling method. In particular, any given trial has two pairs of faces, where their face-wheel values are, 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 ψ_{ij}, the psychophysical similarity to which this distance corresponds. If people made decisions without noise then they should pick pair i,j if and only if $\psi_{ij} > \psi_{kl}$. We add noise by assuming participants decisions are affected by Gaussian error, such that they pick pair i,j if $\psi_{ij} + \varepsilon > \psi_{kl}$.

bioRxiv preprint doi: https://doi.org/10.1101/325472; this version posted August 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

We set the standard deviation of the Gaussian ε noise to 1, so that the model has 9 free parameters, corresponding to the psychophysical scaling values for each possible interval length (e.g., how similar a distance of 5 or 10 'really' is 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 predict observers' judgments, in that equal intervals in the scaled space are discriminated with equal performance. Once the scaling is estimated, we normalize the psychophysical scaling parameters so that psychophysical similarity ranges from 0 to 1. Using Likert ratings for faces gave qualitatively very similar results (not included here).

15. Face identity perceptual matching. N=40 participants on Mechanical Turk were shown a face and had to match this face using a continuous report wheel (100 trials). Because the contribution of motor noise appeared to be minimal in the color matching task (relative to perceptual error), we used only a continuous report wheel (no 60-AFC). Faces were generated from the same continuous face space used in other experiments, and participants had unlimited time had 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 4 locations centered around fixation (randomly), approximately matching the distance to the face wheel and variation in position used in the continuous report memory experiments.7 participants were excluded for below chance error rates.

To convert this data into a perceptual correlation matrix -- asking how likely participants are 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 deg.: -180, -175, … 180) and then linearly interpolated between these to get a value of the confusability for each degree of distance. We then normalized this to range from 0 to 1.

16. Visual long-term memory color report task. Long-term memory data was taken from Miner, Schurgin, Brady[28]. N=30 participants in the lab at UC San Diego performed 5 blocks of a long-term memory experiment. In each block they memorized real-world objects' colors, and then after a brief delay, were shown a sequence of memory tests. Each block's study session consisted of 20 items of distinct categories seen only once and 10 items also of distinct categories seen twice, for a total of 40 presentations of colored objects. Each presentation lasted 3 seconds followed by a 1 second inter-stimulus interval. At test, 20 old objects were presented (10 seen once, 10 seen twice) and 20 new objects of distinct categories were presented. Participants saw each object in grayscale and made an old/new judgment, and then, if they reported the item was old, they reported its color using a continuous color wheel. As described in Miner et al.[28], 7 participants were excluded per preregistered criterion.

Note that long-term memory performance in this task likely depends on a two part decision -- item memory and source memory (e.g., the object itself, and then its color). This two-part decision is related to the processes of recollection and familiarity that can be modeled in various ways[60], but we do not address this here because we focus on understanding color reports only to items judged as 'old' (i.e., conceivably where item memory strength was strong). Participants were highly accurate at judging old/new (*d'*: 3.13, SEM: 0.19), so this analysis includes most of the items participants saw and all of the old items that were reported old (e.g., the only items they saw and made a color report for). Future research should clarify how TCC connects to distinctions between recollection and familiarity.

17. 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 color working memory research. In particular, we

bioRxiv preprint doi: https://doi.org/10.1101/325472; this version posted August 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

examined the two parameters most commonly reported by those fitting mixture models to their data, precision (in terms of SD) and guessing.

We searched for papers that used these mixture model techniques by finding papers that cited the most prominent mixture modeling toolboxes, Suchow, Brady, Fougnie & Alvarez[57] and Bays et al.[61]. We used a liberal inclusion criteria in order to obtain as many data points as possible. Our inclusion criteria were papers that cite either of these toolboxes and report data where: 1) There was some delay between the working memory study array and test; 2) Instructions were to remember all the items; 3) SD/guess values were reported or graph axes were clearly labeled; 4) Participants were normal, healthy, and between ages 18-35. 5). Colors used were widely spaced, discriminable colors from the CIE *L*a*b** color space. Note that even slight changes in the color wheel used between papers (or the addition of noise to stimuli[7]) changes the perceptual confusability of the stimuli and therefore ideally calls for a different similarity function to be measured and therefore a different prediction from TCC about the relationship between 'guess rate' and 'SD'. However, in the current literature analysis we simply assumed these were the same for all papers. For papers that did not report SD/guess values in the text or tables, these values were obtained by digitizing figures with clear axis labels[62,63].

These inclusion criteria resulted in a diverse set of data points, including studies using sequential or simultaneous presentation, feedback vs no feedback, cues vs no-cues, varying encoding time (100-2000 ms), and variable delay (1-10 sec). A total of 14 papers and 56 data points were included.

bioRxiv preprint doi: https://doi.org/10.1101/325472; this version posted August 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

References

1. Cowan, N. (2001). Metatheory of storage capacity limits. Behavioral and Brain Sciences, 24(1), 154-176.

2. Luck, S. J., & Vogel, E. K. (2013). Visual working memory capacity: from psychophysics and neurobiology to individual differences. Trends in Cognitive Sciences, 17(8), 391-400.

3. Baddeley, A. (2003). Working memory: looking back and looking forward. *Nature Reviews Neuroscience*, *4*(10), 829.

4. Ma, W. J., Husain, M., & Bays, P. M. (2014). Changing concepts of working memory. *Nature Neuroscience, 17*(3), 347.

5. Fukuda, K., Vogel, E., Mayr, U., & Awh, E. (2010). Quantity, not quality: The relationship between fluid intelligence and working memory capacity. *Psychonomic Bulletin & Review*, *17*(5), 673-679.

6. Alloway, T. P., & Alloway, R. G. (2010). Investigating the predictive roles of working memory and IQ in academic attainment. *Journal of Experimental Child Psychology*, *106*(1), 20-29.

7. Zhang, W., & Luck, S. J. (2008). Discrete fixed-resolution representations in visual working memory. *Nature*, *453*(7192), 233.

8. van den Berg, R., Shin, H., Chou, W. C., George, R., & Ma, W. J. (2012). Variability in encoding precision accounts for visual short-term memory limitations. *Proceedings of the National Academy of Sciences*, *109*(22), 8780-8785.

9. Bae, G. Y., Olkkonen, M., Allred, S. R., Wilson, C., & Flombaum, J. I. (2014). Stimulus-specific variability in color working memory with delayed estimation. *Journal of Vision, 14*(4), 7-7.

10. Allred, S. R., & Flombaum, J. I. (2014). Relating color working memory and color perception. *Trends in Cognitive Sciences*, *18*(11), 562-565.

11. Pratte, M. S., Park, Y. E., Rademaker, R. L., & Tong, F. (2017). Accounting for stimulus-specific variation in precision reveals a discrete capacity limit in visual working memory. Journal of Experimental Psychology: Human Perception and Performance, 43(1), 6.

12. Torgerson, W. S. (1958). Theory and methods of scaling. New York: Wiley.

13. Maloney, L. T., & Yang, J. N. (2003). Maximum likelihood difference scaling. *Journal of Vision*, *3*(8), 5-5.

bioRxiv preprint doi: https://doi.org/10.1101/325472; this version posted August 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

14. Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*(4820), 1317-1323.

15. Sims, C. R. (2018). Efficient coding explains the universal law of generalization in human perception. *Science*, *360*(6389), 652-656.

16. Nosofsky, R. M. (1992). Similarity scaling and cognitive process models. Annual review of Psychology, 43(1), 25-53.

17. Tanner Jr, W. P., & Swets, J. A. (1954). A decision-making theory of visual detection. *Psychological review*, *61*(6), 401.

18. Macmillan N. A. & Creelman, C. D. (2005). Detection theory: A user's guide (2nd ed.). Mahwah, NJ: Erlbaum.

19. Wilken, P., & Ma, W. J. (2004). A detection theory account of change detection. *Journal of vision*, *4*(12), 11-11.

20. Fougnie, D., Suchow, J. W., & Alvarez, G. A. (2012). Variability in the quality of visual working memory. *Nature communications*, *3*, 1229.

22. Loftus, G. R., & Bamber, D. (1990). Weak models, strong models, unidimensional models, and psychological time. Journal of Experimental Psychology: Learning, Memory, and Cognition, 16(9), 16-9.

22. Bays, P. M. (2015). Spikes not slots: noise in neural populations limits working memory. *Trends in cognitive sciences, 19*(8), 431-438.

23. Smith, P. L., Lilburn, S. D., Corbett, E. A., Sewell, D. K., & Kyllingsbæk, S. (2016). The attention-weighted sample-size model of visual short-term memory: Attention capture predicts resource allocation and memory load. Cognitive psychology, 89, 71-105.

24. Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological review*, *107*(2), 358.

25. Kahana, M. J., & Sekuler, R. (2002). Recognizing spatial patterns: A noisy exemplar approach. Vision research, 42(18), 2177-2192.

26. Gold, J. M., Wilk, C. M., McMahon, R. P., Buchanan, R. W., & Luck, S. J. (2003). Working memory for visual features and conjunctions in schizophrenia. Journal of Abnormal Psychology, 112(1), 61.

bioRxiv preprint doi: https://doi.org/10.1101/325472; this version posted August 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

27. Haberman, J., Brady, T. F., & Alvarez, G. A. (2015). Individual differences in ensemble perception reveal multiple, independent levels of ensemble representation. *Journal of Experimental Psychology: General, 144*(2), 432.

28. Miner, A. E., Schurgin, M. W., & Brady, T. F. (in prep). Extremely precise visual long-term memories for frequently encountered objects. *PsyArXiv*, qr6b9.

29. Brady, T. F., Konkle, T., Gill, J., Oliva, A., & Alvarez, G. A. (2013). Visual long-term memory has the same limit on fidelity as visual working memory. *Psychological Science, 24*(6), 981-990.

30. Rouder, J. N., Morey, R. D., Cowan, N., Zwilling, C. E., Morey, C. C., & Pratte, M. S. (2008). An assessment of fixed-capacity models of visual working memory. Proceedings of the National Academy of Sciences, 105(16), 5975-5979.

31. Wixted, J. T. (2019). The forgotten history of signal detection theory. Journal of experimental psychology: learning, memory, and cognition.

32. Glanzer, M., Kim, K., Hilford, A., & Adams, J. K. (1999). Slope of the receiver-operating characteristic in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(2), 500.

33. Asplund, C. L., Fougnie, D., Zughni, S., Martin, J. W., & Marois, R. (2014). The attentional blink reveals the probabilistic nature of discrete conscious perception. *Psychological science, 25(*3), 824-831.

34. Pratte, M. S. (2018). Iconic Memories Die a Sudden Death. *Psychological science, 29*(6), 877-887.

35. Richter, F. R., Cooper, R. A., Bays, P. M., & Simons, J. S. (2016). Distinct neural mechanisms underlie the success, precision, and vividness of episodic memory. *Elife, 5*.

36. Zokaei, N., McNeill, A., Proukakis, C., Beavan, M., Jarman, P., Korlipara, P., Hughes, D., Mehta, A., Hu, M. T. M., Schapira, A. H.V., & Husain, M. (2014). Visual short-term memory deficits associated with GBA mutation and Parkinson's disease. *Brain, 137*(8), 2303-2311.

37. Rolinski, M., Zokaei, N., Baig, F., Giehl, K., Quinnell, T., Zaiwalla, Mackay, C. E., Husain, M. & Hu, M. T. (2015). Visual short-term memory deficits in REM sleep behaviour disorder mirror those in Parkinson's disease. *Brain, 139*(1), 47-53.

38. Pertzov, Y., Miller, T. D., Gorgoraptis, N., Caine, D., Schott, J. M., Butler, C., & Husain, M. (2013). Binding deficits in memory following medial temporal lobe damage in patients with voltage-gated potassium channel complex antibody-associated limbic encephalitis. *Brain, 136*(8), 2474-2485

bioRxiv preprint doi: https://doi.org/10.1101/325472; this version posted August 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

39. Henson, R. N. A., Rugg, M. D., Shallice, T., & Dolan, R. J. (2000). Confidence in recognition memory for words: dissociating right prefrontal roles in episodic retrieval. *Journal of cognitive neuroscience*, *12*(6), 913-923.

40. Rutishauser, U., Ye, S., Koroma, M., Tudusciuc, O., Ross, I. B., Chung, J. M., & Mamelak, A. N. (2015). Representation of retrieval confidence by single neurons in the human medial temporal lobe. *Nature neuroscience*, *18*(7), 1041.

41. Bays, P. M. (2014). Noise in neural populations accounts for errors in working memory. *Journal of Neuroscience*, *34*(10), 3632-3645.

42. Marr, D. (1982). Vision: A computational investigation into the human representation and processing of visual information. Henry Holt and co. Inc., New York, NY.

43. Bays, P. M. (2019). Correspondence between population coding and psychophysical scaling models of working memory. BioRxiv, 699884.

44. Wei, X. X., & Stocker, A. A. (2015). A Bayesian observer model constrained by efficient coding can explain'anti-Bayesian'percepts. *Nature neuroscience*, *18*(10), 1509.

45. Wei, X. X., & Stocker, A. A. (2017). Lawful relation between perceptual bias and discriminability. *Proceedings of the National Academy of Sciences*, *114*(38), 10244-10249.

46. Krauskopf, J., & Gegenfurtner, K. R. (1992). Color discrimination and adaptation. *Vision Research, 32(11), 2165–2175.*

47. Giesel, M., Hansen, T., & Gegenfurtner, K. R. (2009). The discrimination of chromatic textures. *Journal of Vision, 9(9):11, 1–28.*

48. Fougnie, D., Alvarez, G. A., & Brady, T. F. (in revision). If at first you don't retrieve, try, try again: Second chances reveal more information in working memory.

49. Rademaker, R. L., Tredway, C. H., & Tong, F. (2012). Introspective judgments predict the precision and likelihood of successful maintenance of visual working memory. *Journal of Vision*, *12*(13), 21-21.

50. Wixted, J. T., & Wells, G. L. (2017). The relationship between eyewitness confidence and identification accuracy: A new synthesis. *Psychological Science in the Public Interest, 18*(1), 10-

51. Difallah, D., Filatova, E., & Ipeirotis, P. (2018, February). Demographics and dynamics of mechanical Turk workers. In Proceedings of the eleventh acm international conference on web search and data mining (pp. 135-143). ACM

bioRxiv preprint doi: https://doi.org/10.1101/325472; this version posted August 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

52. Brady, T. F., & Alvarez, G. A. (2011). Hierarchical encoding in visual working memory: Ensemble statistics bias memory for individual items. *Psychological Science*, *22*(3), 384-392.

53. Brady, T. F., & Tenenbaum, J. B. (2013). A probabilistic model of visual working memory: Incorporating higher order regularities into working memory capacity estimates. *Psychological review*, *120*(1), 85.

54. Brady, T. F. and Alvarez, G.A. (2015). Contextual effects in visual working memory reveal hierarchically structured memory representations. *Journal of Vision, 15*(15):6.

55. Nadarajah, S., Afuecheta, E., & Chan, S. (2019). On the distribution of maximum of multivariate normal random vectors. *Communications in Statistics-Theory and Methods*, *48*(10), 2425-2445.

56. Fougnie, D., Asplund, C. L., & Marois, R. (2010). What are the units of storage in visual working memory?. *Journal of Vision*, *10*(12), 27-27.

57. Suchow, J. W., Brady, T. F., Fougnie, D., & Alvarez, G. A. (2013). Modeling visual working memory with the MemToolbox. *Journal of vision*, *13*(10), 9-9.

58. Myung, J. I. & Pitt, M. A. (2018). Model comparison in psychology. In J. Wixted (Editor-in-Chief) & E.-J. Wagenmakers (Volume Editor), *Steven's Handbook of Experimental Psychology and Cognitive Neuroscience, Fourth Edition, Vol.5: Methodology*. New York, NY: John Wiley & Sons.

59. Brady, T. F., Konkle, T., Alvarez, G. A., & Oliva, A. (2008). Visual long- term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences, 105*(38), 14325–14329.

60. Wixted, J. T., & Mickes, L. (2010). A continuous dual-process model of remember/know judgments. *Psychological review*, *117*(4), 1025.

61. Bays, P. M., Catalao, R. F., & Husain, M. (2009). The precision of visual working memory is set by allocation of a shared resource. *Journal of vision*, *9*(10), 7-7.

62. Rohatgi, A. (2011). WebPlotDigitizer. *URL http://arohatgi. info/WebPlotDigitizer/app*.

63. Mukherjee, S., Seok, S. C., Vieland, V. J., & Das, J. (2013). Cell responses only partially shape cell-to-cell variations in protein abundances in Escherichia coli chemotaxis. *Proceedings of the National Academy of Sciences*, *110*(46), 18531-18536.

64. Thurstone, L. L. (1927). A law of comparative judgment. Psychological Review, 34, 273-286.

bioRxiv preprint doi: https://doi.org/10.1101/325472; this version posted August 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

65. Rademaker, R.L., Park, Y., Sack, A.T., & Tong, F. (2018). Evidence of gradual loss of precision for simple features and complex objects in visual working memory. Journal of Experimental Psychology: Human Perception and Performance. Advance online publication.

66. Rotello, C. M. (2017). Signal detection theories of recognition memory. *In J. H. Byrne (Ed.) & J. T. Wixted (Vol. Ed.), Learning and Memory: A Comprehensive Reference, Vol. 2: Cognitive Psychology of Memory (2nd ed.,pp. 201-226). Oxford: Elsevier.*

bioRxiv preprint doi: https://doi.org/10.1101/325472; this version posted August 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Acknowledgements
We thank Viola Störmer, Rosanne Rademaker, Jeremy Wolfe, Daryl Fougnie and Talia Konkle for comments on these ideas and on the manuscript, and Yong Hoon Chung and Brittany Hawkins for help with data collection. For funding, we would like to acknowledge NSF CAREER (BCS-1653457) to TFB.

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 code, ran the model, and analyzed output data; M.W.S., J.T.W., and T.F.B. wrote the manuscript.

Competing Interests
The authors declare no competing interests.

Materials and Correspondence
Material and correspondence requests should be addressed to Dr. Mark Schurgin (mschurgin@ucsd.edu)

bioRxiv preprint doi: https://doi.org/10.1101/325472; this version posted August 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Supplemental Figures

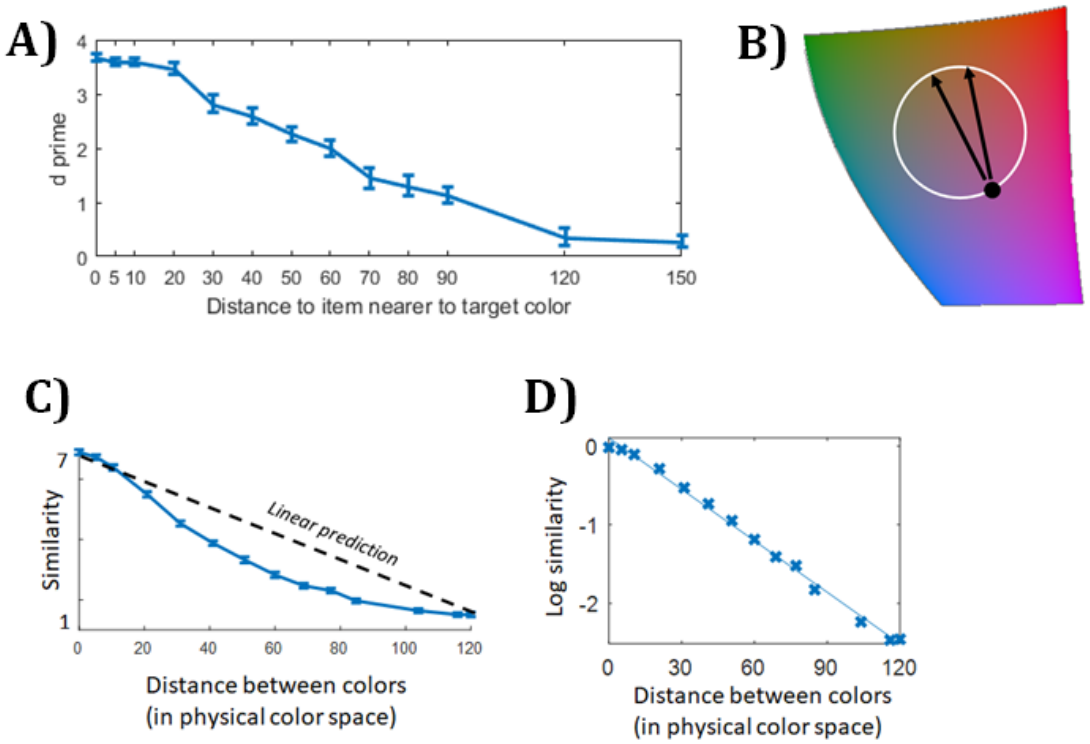

Figure S1. **(A)** Data from all distances in the fixed distance triad task (Figure 1C). On each trial, there was a target color, here always at 0°, and participants task was to choose which of two other colors was closer to the target color in color space. The two choice colors always differed by 30°. 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 Figure 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 Figure 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 higher 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., e.g., 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 Figure 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[16].

bioRxiv preprint doi: https://doi.org/10.1101/325472; this version posted August 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

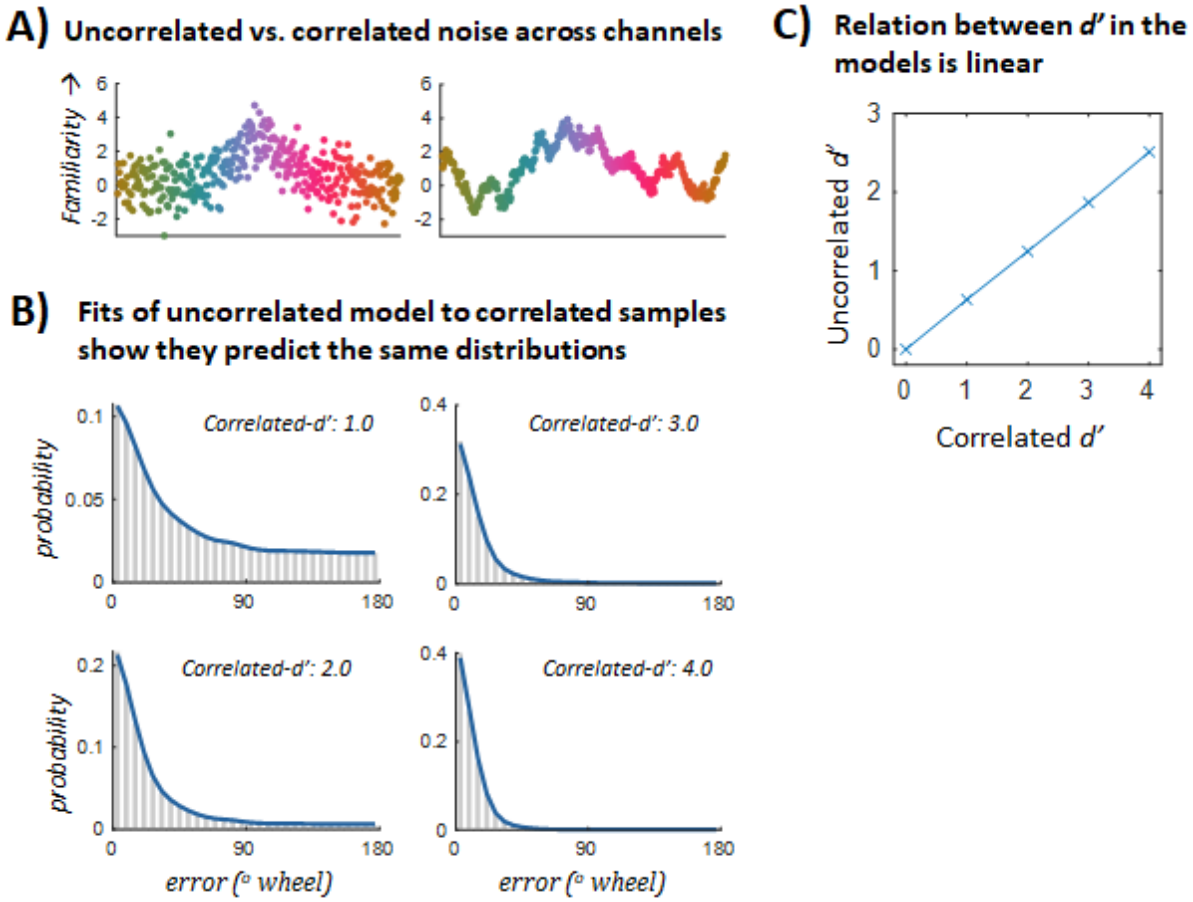

Figure S2. Uncorrelated vs. correlated noise versions of TCC. Only the correlated-noise TCC produces true *d'* values -- those that are interchangeable with *d'* you'd estimate from a same/diff task with the same stimuli. However, the simpler uncorrelated noise TCC predicts the exact same distributions of errors, 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.

bioRxiv preprint doi: https://doi.org/10.1101/325472; this version posted August 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

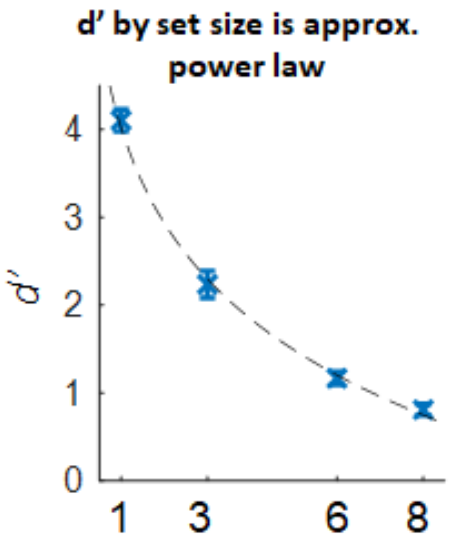

Figure S3. While memory strength varies according to a variety of different factors, many researchers have been particularly interested in the influence of set size. TCC shows that memory strength (d') decreases according to a power law as set size changes, broadly consistent with fixed resource theories of memory[22,23].

bioRxiv preprint doi: https://doi.org/10.1101/325472; this version posted August 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

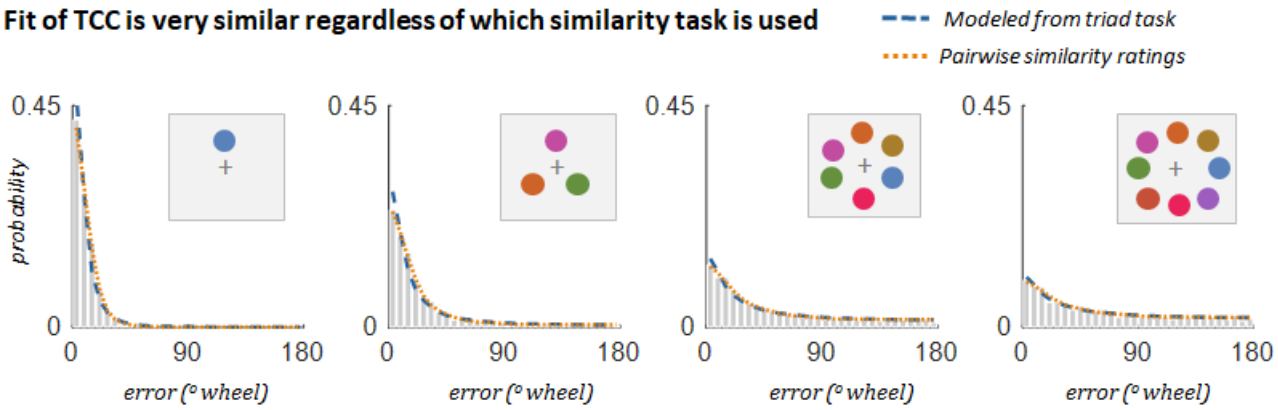

Figure S4. In the current data for color, both the model-based triad psychophysical scaling data and the Likert similarity rating produce extremely similar data (see Figure 1). Thus, they all produce similar fits to the memory data (shown here). It is important to note that depending on the number of trials, a large number of data points (i.e. 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, it is important to note that 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 that 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[64] and Torgerson[12].

bioRxiv preprint doi: https://doi.org/10.1101/325472; this version posted August 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

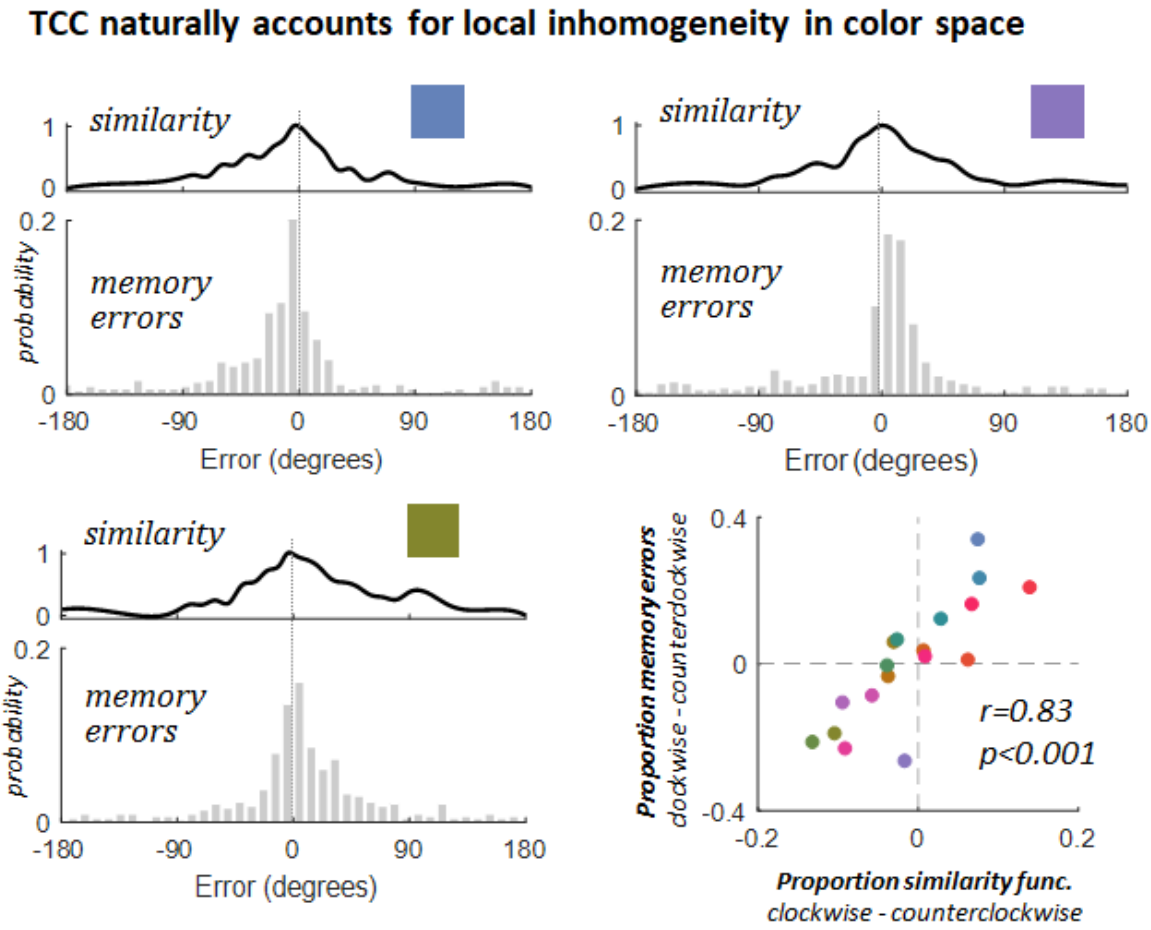

Figure S5. Many stimulus spaces contain non-uniformities, which may affect subsequent working memory performance. Indeed, Bae et al.[9] discovered non-uniformities in working memory for color, where responses for targets tend to be more precise for some colors than others and can be biased towards nearby categorical anchors (i.e. 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[10,11]. 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 (e.g., Figure 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 having predictable consequences for memory. 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.

bioRxiv preprint doi: https://doi.org/10.1101/325472; this version posted August 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

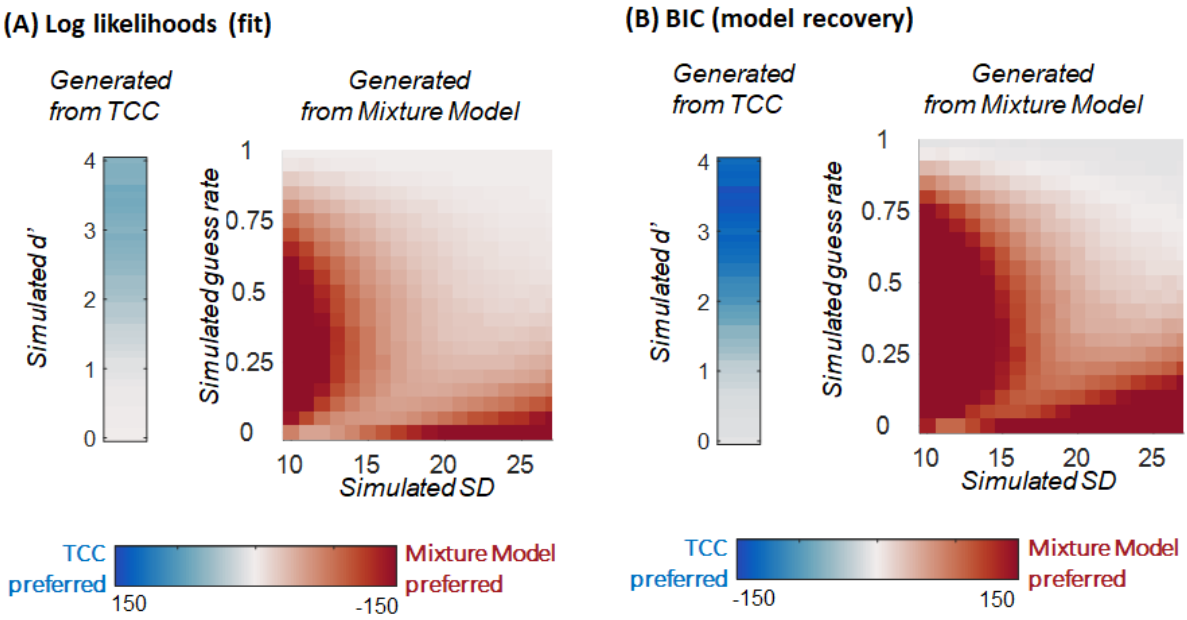

Figure S6. (**A**) We generated data from both TCC (*d'*) and the standard mixture model (precision [SD] and guessing), performing 50 simulations of 2000 trials worth of data each for each of the models (consistent with the amount of group data in the main experiments), and then fit both models to the generated data to see which yielded a higher log-likelihood. Even with no penalty for complexity -- simply using log likelihood -- for data generated by TCC, the standard mixture model fit all data with a $d' < 1$ better than TCC itself. Thus, for data generated by TCC, the standard mixture model, being considerably more flexible than TCC in the range of distributions it can fit, fits the data about as well -- and in some cases, better -- than TCC. When fitting data generated by the mixture model, TCC was dispreferred at all values in terms of fit, and strongly dispreferred for huge swaths of potential mixture model parameters. This is because the mixture model can generate a huge variety of distributions that TCC cannot mimic. The same is true, but even more so, for the 3-parameter variable precision model, which can fit an even much larger range of distributions than even the standard 2-parameter mixture model. Only a miniscule part of the distributions predicted by the 3-parameter variable precision model can even be approximated by TCC, and this model can perfectly mimic TCC. (**B**) Same data, with BIC instead of log-likelihood. Taking into account model complexity increases the preference for TCC in TCC-generated data and creates a very slight TCC preference in mixture model data with simulated "guess rates" very near 1.0, where the two models make identical predictions in terms of error (of equal responding to all options); though note the two models make differing predictions about confidence at these values, predicting different ROCs. In general, with this amount of data, BIC appears well-calibrated, accurately recovering the appropriate model in nearly all cases and with a stronger preference for the relevant models where they diverge from each other more.

bioRxiv preprint doi: https://doi.org/10.1101/325472; this version posted August 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

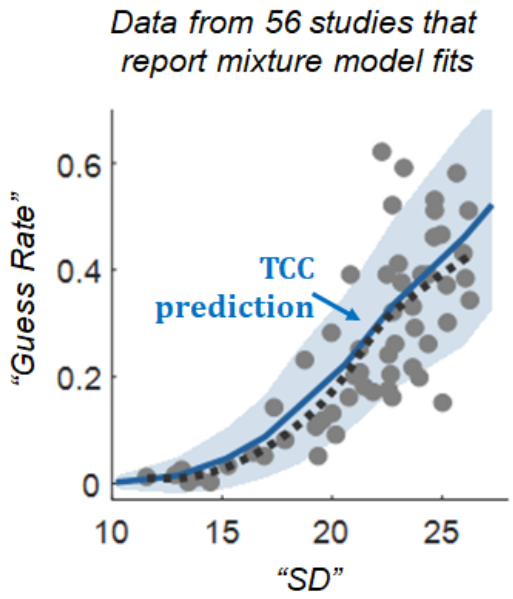

Figure S7. Analysis of previous literature measuring the most widely used model parameters currently used to analyze working memory performance. Gray dots are values reported in papers found in the literature; the dashed black curve is a LOESS (local regression) smoothed version of these points. The solid blue curve reflects the average "guess" and "SD" parameters when fitting the mixture model to data generated by TCC, as a function of the *d'* of TCC. The blue shading shows 2 standard deviations when each participant has 100 trials/condition. Despite claiming to independently model multiple parameters, this entire diverse set of data points falls near the trade-off between these parameters predicted when fitting data sampled from the TCC with the 2-parameter model -- in other words, one parameter is sufficient to capture much of the data observed in working memory tasks (data that has previously been thought to require at least two -- and often 3 parameters -- to explain). Note that the region in Fig. S7 that TCC predicts is also the only region of Fig. S6 where the TCC can fit data generated from the mixture model. In addition, note that some of these papers use different color wheels than the one we use to generate the similarity function, and thus some of the deviation from the TCC prediction line -- minor as it is -- is caused by using an "incorrect" TCC prediction (e.g., using a prediction from an incorrect stimulus space). In addition to fitting a two parameter model, some previous research has claimed to dissociate these parameters. If a one-parameter model can account for the data, how has previous research so often found dissociations between these parameters? The majority of these dissociations find that precision (SD) does not change when the 'guess rate' (or capacity) does change[7,29]. However, this dissociation is naturally explained by TCC because at low d' values, 'guess rate' can change by a huge amount with SD changing by only a few degrees. For example, over a wide range of guess rates, precision may only vary between SD=21 and SD=24, a difference that is visually indistinguishable and would require extremely high power to detect. As an example, sampling 20 subjects of 100 trials each of data from the TCC at *d'*=0.65 vs. *d'*=0.45 and fitting these data with the 2-parameter mixture model reveals that such an experiment would find $p<0.05$ for 'capacity' greater than 60% of the time but $p<0.05$ for 'precision' approximately 11% of the time, despite both parameters being necessarily linked in the data from TCC. In line with this interpretation, many researchers have now found that with high enough power, previous studies claiming only a change in 'guess rate' but not 'SD' actually find changes in both, with very small changes in SD present along with large changes in 'guess rate'[65]. Other dissociations have sometimes been found -- for example, Zhang and Luck[7] report a manipulation that causes a change in SD but not 'guess rate' -- but these dissociations inevitably rely on comparisons across different sets of stimuli with different psychophysical similarity functions (e.g., the Zhang and Luck manipulation adds color noise to the items, making them less distinct), which is perfectly consistent with TCC.

bioRxiv preprint doi: https://doi.org/10.1101/325472; this version posted August 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

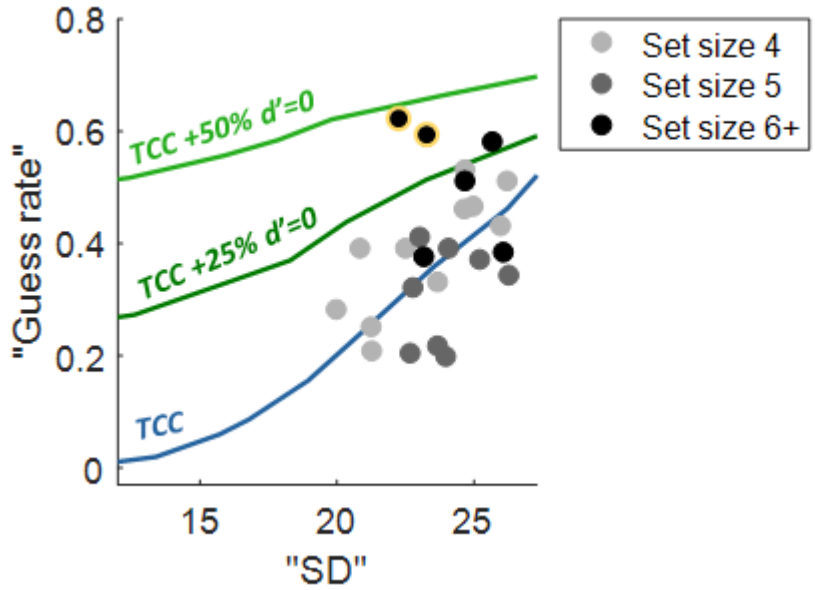

Figure S8. Existing working memory data from high set sizes (4+) is often claimed to provide evidence for 'slots' or for the existence of very low precision items, with these items that are unrepresented or poorly represented giving rise to the long tails of the distribution. By contrast, TCC predicts such long tails with no sense of unrepresented or poorly represented items. Here we show how TCC predicts that mixture model parameters from the standard two parameter mixture model should change as a function of *d'* in the TCC model. The blue line and all of the data points are the same as Figure S7, but with the data points now labeled by set size and only "high" set sizes (>=4) plotted, as these are the points where traditional models claim many items must be unrepresented or extremely poorly represented. Note that the vast majority of the points are better fit by the straightforward TCC model -- which simply assumes all items are equally well represented -- than by models that add some proportion of 'unrepresented' items to TCC (plotted in green; note that as expected, these models selectively change the predicted 'guess rate' parameter). For a slot model prediction with 3 items represented, nearly 50% of items should be unrepresented at set size 6, and this is clearly incompatible with the previous data as well as the data we report in the main manuscript. In general, the parameters found in the previous literature are perfectly consistent with the basic TCC prediction with no added assumptions about unrepresented items or poorly represented items. Note that the two set size 6 points outlined in yellow come from the original Zhang and Luck[7] paper that introduced mixture models to this literature and used them to argue for slots. The fact that they are an outlier on this plot may be the reason those authors proposed a model that argues that only 'guess rate' but not 'standard deviation' changes as a function of set size.

bioRxiv preprint doi: https://doi.org/10.1101/325472; this version posted August 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

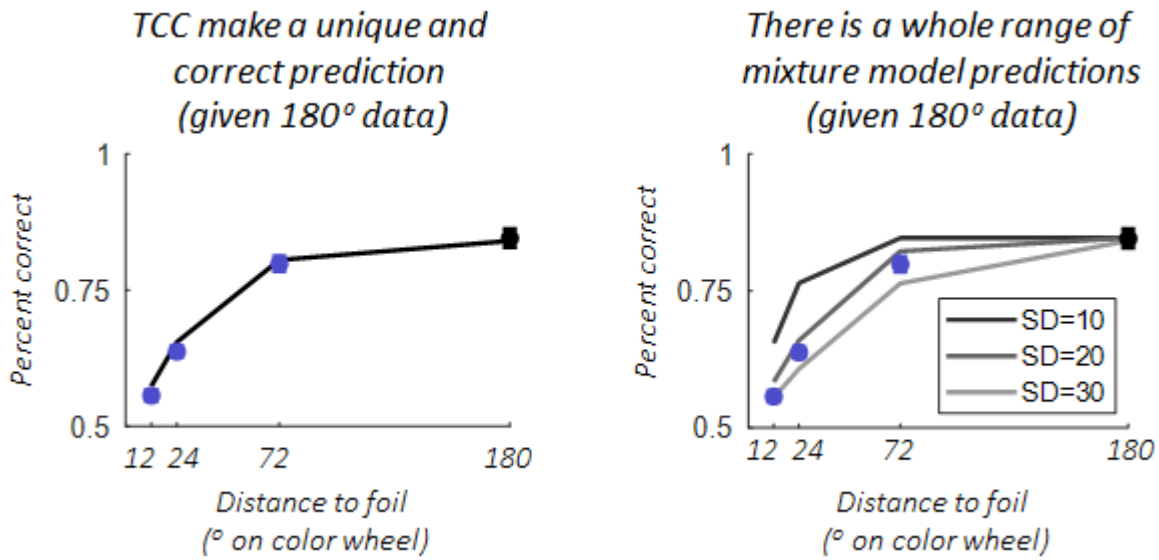

Figure S9. Given 2-AFC performance with maximally distinct 180 degree foils (black dot), TCC makes a unique prediction about exactly how well people should perform on other foils -- with no free parameters. By contrast, using the 180 degree foils to constrain the mixture model allows this model to set the 'guess rate', but it leaves the precision of memory unknown. Thus, mixture models, while capable of fitting the data the same as TCC for a certain precision parameter (since ultimately they can predict any distribution TCC can, as they are much more flexible), do not make a unique prediction. Making strong predictions is the most critical test of a model[24] and can be formalized using a Bayes factor, which provides strong evidence in favor of TCC in this case. Similar logic applies in the experiment taking 180 degree 2-AFC and generalizing to continuous report and other n-AFC conditions.

bioRxiv preprint doi: https://doi.org/10.1101/325472; this version posted August 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

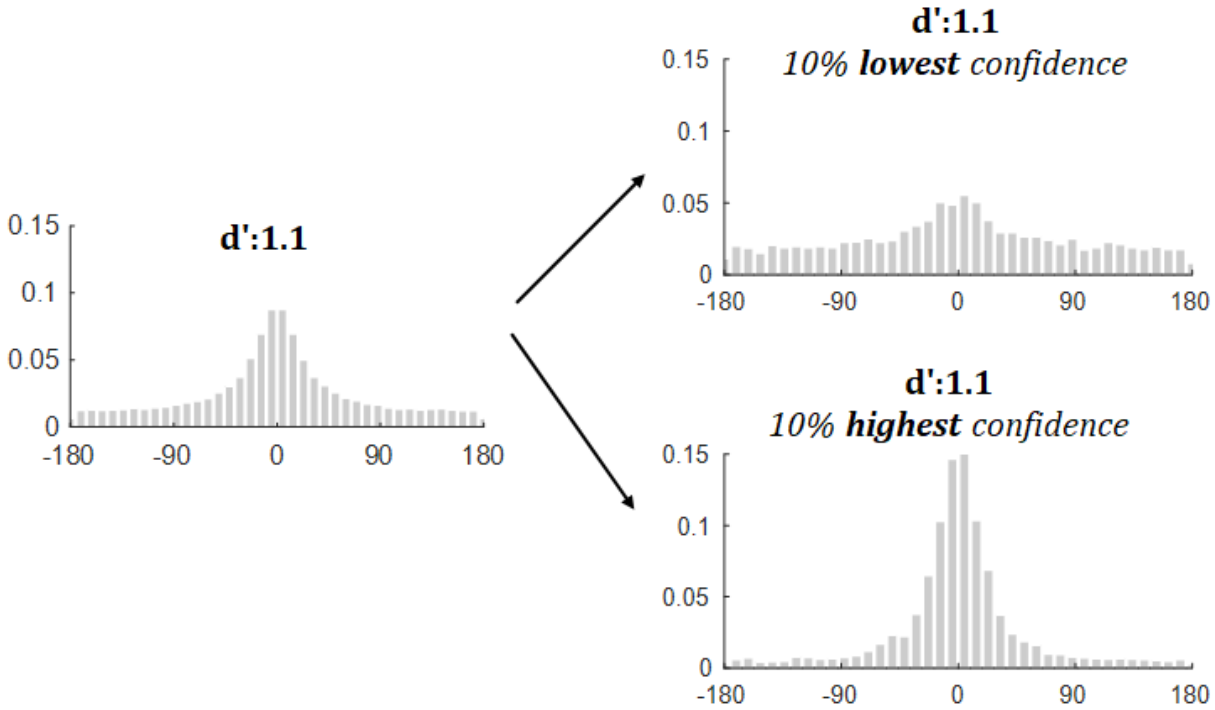

Figure S10. Some studies used to support variability of information across individual items or trials have done so by using a confidence metric[26]. While variability and confidence are distinct from one another, in a large amount of research they are inextricably linked. An interesting advantage and implication of signal detection-based models is that they naturally predict confidence data[66]. 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, the 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[20] 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 somewhat, although the nearly-equal variance ROC curves even at high set sizes suggest *d′* likely variance between items may be small.

bioRxiv preprint doi: https://doi.org/10.1101/325472; this version posted August 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

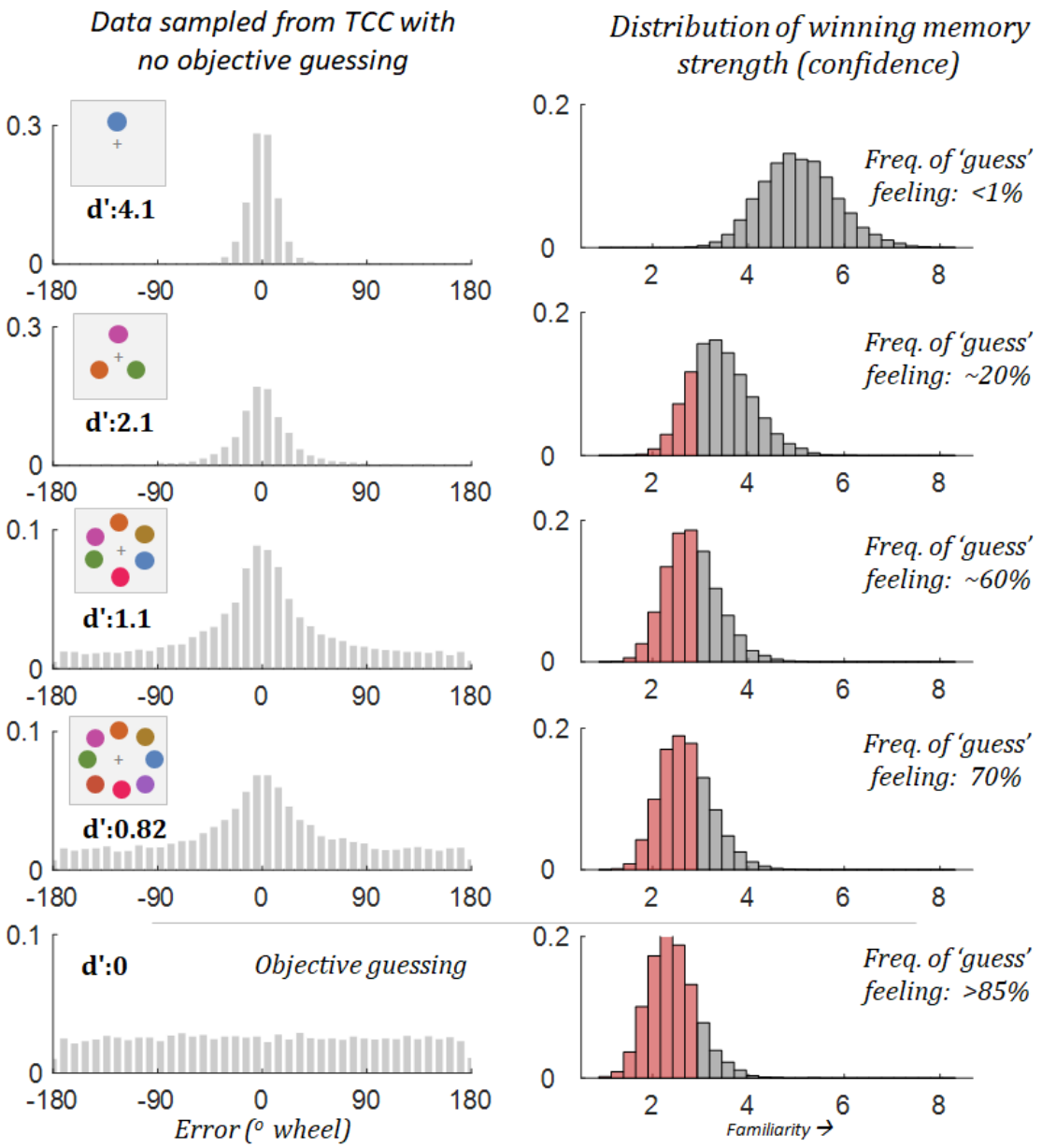

Figure S11. Participant in a set size 8 working memory experiment often feel like they do not remember an item and are "guessing", leading to a wide variety of models that predict people know nothing about many items at high set sizes and truly are objectively guessing. However, as noted in Figure S10, 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.

bioRxiv preprint doi: https://doi.org/10.1101/325472; this version posted August 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

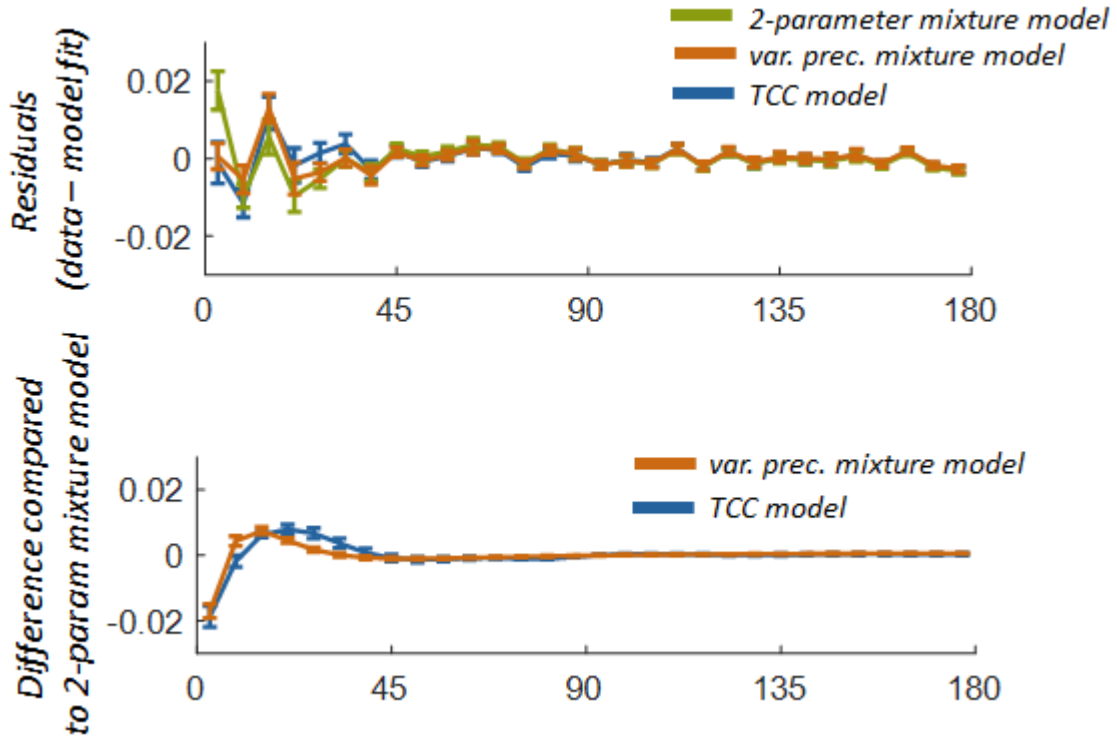

Figure S12. The fact that the data is "too peaky" for the standard mixture model to fit is often used as evidence of variability in precision. However, TCC, even with no variability in *d′*, does not show such residuals because the similarity function -- perceptual error convolved with an exponential -- is not a von Mises distribution. (**A**) Residuals of models in our main experimental data (calculated from model's fit separately for each participant and set size; error bars show SEM across participant). Notice that the 2-parameter mixture model has a significant residual near 0 error, as the data is peakier than this model, whereas the variable precision model and TCC model do not have this residual (All models do have some other residuals, though they are very similar between all the models). **(B)** Difference in residuals between the 2-parameter mixture model and the other two models. Both the variable precision model and TCC show significantly less residual than the two-parameter mixture model near 0 errors, suggesting they both properly account for the peakiness in the data. Note that this is true of TCC even though it has no variability in *d′*.

bioRxiv preprint doi: https://doi.org/10.1101/325472; this version posted August 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

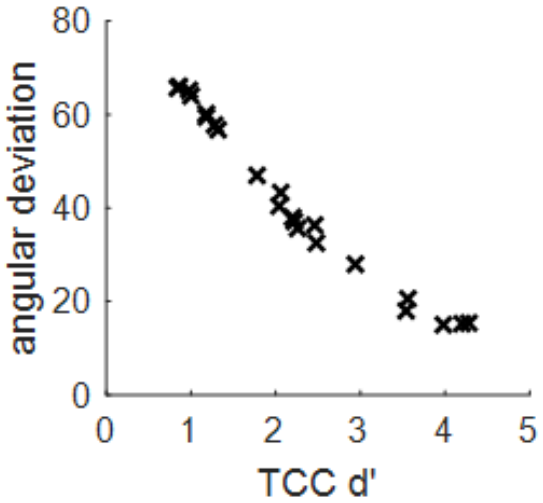

Figure S13. Plot of the best fit TCC d' vs. the angular deviation of the error data (a circular analog of the standard deviation; as computed with MATLAB's circ_std function) for all 22 datasets from Fig. 3. For data like the current data where there is nearly no location-based confusions ('swaps'), the simpler analysis of this descriptive statistic (angular deviation) is linearly related to d' for d' less than approximately 3.0, and thus, for data not near ceiling, may be an adequate substitute for fitting the full TCC. This is useful because the angular deviation is just a descriptive statistic of the data and thus does not require the collection of similarity data or perceptual confusability data. Note that just as with percent correct -- which is approximately linear with d' when far from ceiling, but becomes deeply non-linear near ceiling -- the d' curve begins to bend near ceiling. This is because improving from 95% correct to 99% correct requires a very large change in d', and similarly, improving your performance in continuous report when it is already very good requires a large change in memory strength. In theory the same should be true near floor, although these 22 datasets do not clearly demonstrate that because there is little data with d'<1.0. However, for data away from ceiling and floor and with little or no 'swaps', computing angular deviation may be sufficient to summarize data in a framework compatible with TCC.

bioRxiv preprint doi: https://doi.org/10.1101/325472; this version posted August 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

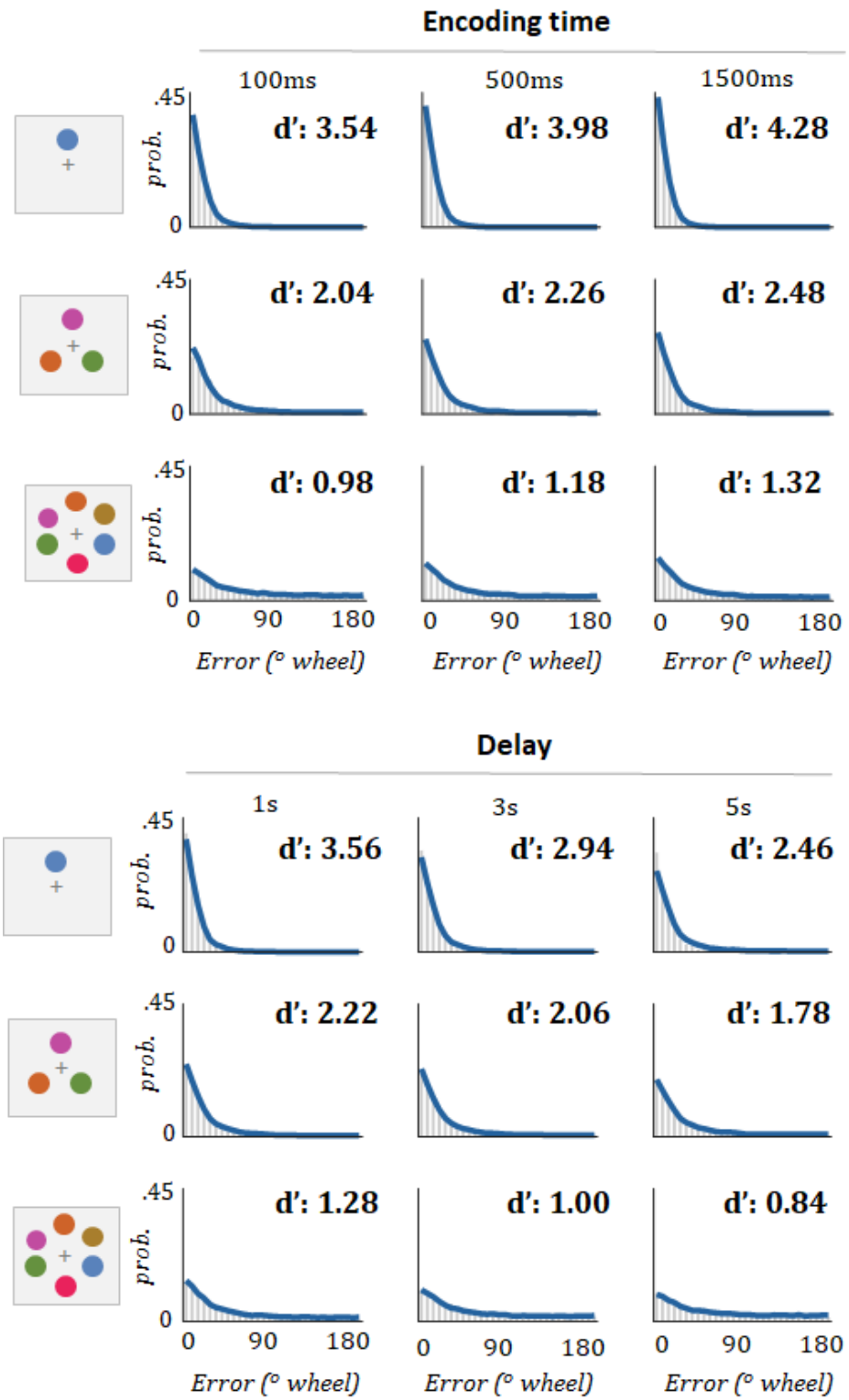

Figure S14. Fits of TCC to the all encoding and delay conditions, including those not plotted in Fig. 3. TCC provides a strong fit at all encoding and delays (see correlations and model comparisons in Fig. 3).

bioRxiv preprint doi: https://doi.org/10.1101/325472; this version posted August 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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

Table S1. Data points used in the literature review collected from a total of 14 papers.

bioRxiv preprint doi: https://doi.org/10.1101/325472; this version posted August 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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

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

Table S2. TCC's fit to color memory data is reliably preferred by model comparison metrics that emphasize simplicity (e.g., BIC) across all set sizes compared to mixture models and variable precision mixture models. It provides a similar fit to these models when using leave-one-out cross validation on log likelihood, as both TCC as well as the two mixture models predict effectively the same distribution of errors when fit with N-1 error points (as N=2000 error datapoints >> the number of parameters for all models). Fitting to the group data rather than individual subjects gives BIC values at set size 1,3,6 and 8 of -24, -56, -26, -25 for TCC vs. standard mixture model (all very strong evidence favoring TCC), and BIC values of -2, -23, -15, -19 for TCC vs. variable precision model (e.g., both models fit set size 1 data well -- the least distinct set size, since there are no long tails -- but all others are very strong evidence in favor of TCC). Note that, as shown in Fig. S6, model recovery using BIC is well calibrated using this number of trials.

bioRxiv preprint doi: https://doi.org/10.1101/325472; this version posted August 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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

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

Table S3. TCC applied to face memory. As with colors, TCC is reliably preferred by model comparison metrics that emphasize simplicity (e.g., BIC) across all set sizes compared to mixture models and variable precision mixture models. Also, as with color, it provides a similar fit to these models when using leave-one-out cross validation on log likelihood, as both TCC as well as the two mixture models predict effectively the same distribution of errors when fit with N-1 points (as N >> the number of parameters for all models). Fitting to the group data rather than individual subjects gives BIC values at set size 1 and 3 of -177 and -24 for TCC vs. standard mixture model (all very strong evidence favoring TCC), and BIC values of -53, -10 for TCC vs. variable precision model (all very strong evidence in favor of TCC). Note that, as shown in Fig. S6, model recovery using BIC is well calibrated using this number of trials.