***This is the pre-reviewed version of the manuscript. Please contact the authors (mschurgin@ucsd.edu) for an updated version.**

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

Abstract

Models of visual memory have operated on the assumption that memory representations can be explained in physical stimulus space, leading to proposals that we have fixed limits on visual working memory capacity and there is a strong distinction between working memory and long-term memory representations. Here we show that perception and memory are not based on physical stimulus space, and instead rely on a transformed representation that is non-linearly related to the physical stimulus space. This result calls into question the fundamental basis of nearly all extant models. We show that when psychophysical distance rather than physical distance is considered, aspects of memory that have required fundamentally different models -- across different stimuli, tasks, and even the distinction between working and long-term memory -- can be explained with a unitary signal detection framework. These results lead to a substantial reinterpretation of the relationship between perception, working memory and long-term memory.

Working memory is typically conceptualized as a fixed capacity system, with a discrete number of items, each 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 representations, as they allow the precise quantification of information stored in memory. In one prominent paradigm, researchers present participants with a set of colored circles to remember and then probe one of them after a delay, asking people to report what color it was by clicking on a color wheel (Figure 1A). The typical analysis of such data focuses on the distribution of errors people make measured in degrees along the response wheel, *x*, where correct responses have 0° error, and errors range up to +/- 180°, reflecting the incorrect choice of the most distant item from the target on the response wheel (Figure 1B). This data is then modeled to understand and quantify memory performance[7,8]. An implicit assumption in this research is that because the perceptual space is approximately locally uniform[9], the psychophysical distance between items relevant for memory performance (e.g., 0° to +/- 180°) is relatively linear as well. However, research on perception shows that the relationship between a physical stimulus and the psychological representation of that stimulus is rarely linear[10,11]. Nevertheless, theories and models of human visual working memory rely on measurements conducted in physical stimulus space without taking this relationship into account.

Here we show that, contrary to the assumptions guiding extant models of working memory, the relationship between the physical stimulus and the underlying psychological representation is not linear. Moreover, once this non-linearity is taken into account, a new parsimonious and

powerful model of working memory becomes possible. This model provides a unified framework for understanding not only working memory but also visual long-term memory, and permits parameter-free generalization across different stimulus spaces (e.g., different color wheels) and across different tasks (e.g., continuous report vs. change detection). We begin with working memory for color as our main case study and then expand the psychophysical scaling framework and resulting model to account for working memory for orientation, face identity, and for the case of visual long-term memory.

Results

Measuring psychophysical distance. To determine the relationship between physical distance along the color wheel and psychophysical distance, we tested how accurately participants could determine which of two colors was closest to a target color using a triad task[12,13]. This triad task 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 many other colors to that target (i.e., all the colors on the color wheel; see Figure 1). We find that even with a fixed 30° distance between two color choices, participants were much more accurate at determining which color was closer in color space to the target when the two colors were near the target in color space compared to when they were far from the target (Figure 1C shows a subset of data; ANOVA $F(12,384) = 71.8$, $p<0.00001$, $\eta^2=0.69$; full data in Supplemental Material, section 7). In other words, in a purely perceptual task, participants largely could not tell whether a color 120° from the target or a color 150° from the target is closer to the target, whereas this task is trivial if the colors are 5° and 35° from the target. To compute a full psychophysical scaling function for color perception, we utilized two tasks: the triad task because it involves a comparison to a target, and the quad task because it minimizes the chances that participants' encode the choices relative to each other rather than the target. 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. For both tasks, the estimated psychophysical scaling function is similar and falls off in a nonlinear, exponential-like fashion with respect to physical distance (Figure 1D). It is also well-matched by a smoother measure that requires substantially less data, namely, the pairwise similarity of colors at different distances along the color wheel measured using a Likert scale similarity task.

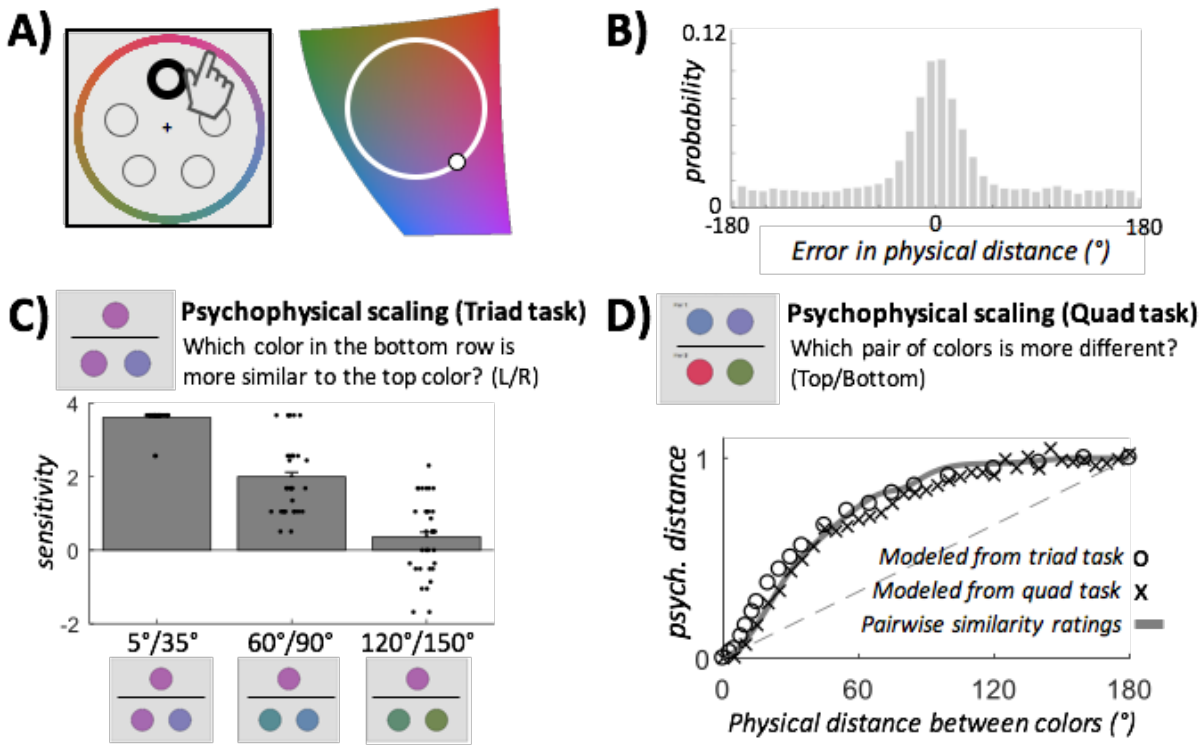

Figure 1. *(**A**) A widely used method in working memory is to select a color circle from a slice of color space, show working memory items drawn from this circle, and then at test probe the contents of a memory item by presenting the entire continuous circle to participants to make a response. Similar response wheels are used for other features, like orientation and face identity. (**B**) A histogram of results generally observed for such tasks, traditionally plotted as a function of physical distance, with the unit of measure being degrees of error along the response wheel. (**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 decreased as the choices became more distant from the target. Error bars represent within-subject SEM. **(D)** We can use the data from the triad task, and from another psychophysical scaling task, the quad task, to infer the normalized psychophysical distance of colors at different physical distances along the color wheel (black o's; black x's). This function is decidedly non-linear, and thus these data contradict the linear scaling assumption of existing working memory models. Rather than using the quad or triad task and fitting a model, we can also ask participants to rate the similarity of two colors on a Likert scale (1-7) and normalize their ratings to range from 0 to 1, with 0 being most similar and 1 being least similar. This similarity rating task provides a smooth measure of psychophysical distance that is in close agreement with the results of the triad and quad task.*

A key implication of these scaling results is that the physical axis of error along the response wheel previously used to analyze working memory capacity is uninterpretable. What appears to be a large distance in physical space (e.g., the distance between 120° and 180° on the x-axis of Figure 1B) is a very short distance in psychophysical space (Figure 1D). In particular, since participants are essentially incapable in a perceptual task of discerning whether an item 120° or 180° from the target in color space is closer to the target, it is not surprising that they confuse these colors equally often with the target in memory. This is true even though the color space itself is approximately uniform[9] with some local inhomogeneity[14,15]. As an analogy, consider a person standing on a road lined with equally-spaced telephone poles. As the person looks down the road in either direction, telephone poles recede into the distance and become less and less distinct from their neighbors. Although the local separation between the poles remains constant, the *perceived* (psychophysical) separation between them falls off with distance[16]. Analogously, adjacent points on the color wheel become increasingly indistinguishable from each other as the distance between the target and those adjacent points increases (Figure 1D).

Incorporating psychophysical distance into a standard signal detection model. The outcome of our psychophysical scaling experiments reveals that a foundational assumption guiding previous working memory models is incorrect. By measuring working memory performance in physical distance instead of psychophysical distance, the existing data has been erroneously interpreted in terms of (at least) two distinct parameters: ‘capacity’ or number of items remembered, and the ‘precision’ with which items are remembered[7]. However, once psychophysical distance is taken into account, this distinction is revealed to be illusory. Indeed, we find that the key data from visual working memory studies can be accurately fit by an extremely straightforward one-parameter signal detection model that treats the memory-match

signal generated by each color on the wheel as arising from its psychophysical distance to the target with Gaussian error (Figure 2).

This model is fundamentally the same signal detection model used to understand how people perform long-term memory and perceptual detection tasks. In particular, when an item is probed at test and participants respond by selecting a choice on a color wheel, each test option (in this case, the 360 colors on the wheel) generates a memory-match signal, and on each trial participants choose the color that generates the maximum memory-match signal (Figure 2A). The memory-match signals have variability -- they are generated as samples from a normal distribution -- and the mean memory-match signal for a given color x on the color wheel, denoted d_x, is given by that color's psychophysical distance to the target, i.e., $d_x = d'(1-f(x))$, where d' is the model's only free parameter and $f(x)$ is the empirically determined psychophysical scaling function (shown on the y-axis in Figure 1D). For $x = 0$, $f(x) = 0$, $1-f(x) = 1$ and $d_x = d'$; for $x = 180$, $f(x) = 1$, $1-f(x) = 0$ and $d_x = 0$. Because of the nonlinear psychophysical scaling, colors in the >~90° physical distance range do not cover a great expanse but instead all cluster near $f(x) \approx f(x)_{max}$ such that $d_x \approx 0$ for x = ~90° to 180°.

Remarkably, this Target Confusability Competition (TCC) Model, a straightforward signal detection model that takes into account psychophysical distance, can explain the key stimulus-driven features of working memory. Of most importance, it accurately characterizes memory performance for color report across different set sizes (Figure 2B), including the relative height of the "long tails" of the distribution when plotted in physical distance compared to the width of the central distribution (a critical distinction guiding previous models of working memory). This is demonstrated by the Pearson correlation between the binned data and model fits (set size 1: R^2=0.996, set size 3: R^2=0.980, set size 6: R^2=0.992, all p<0.001).

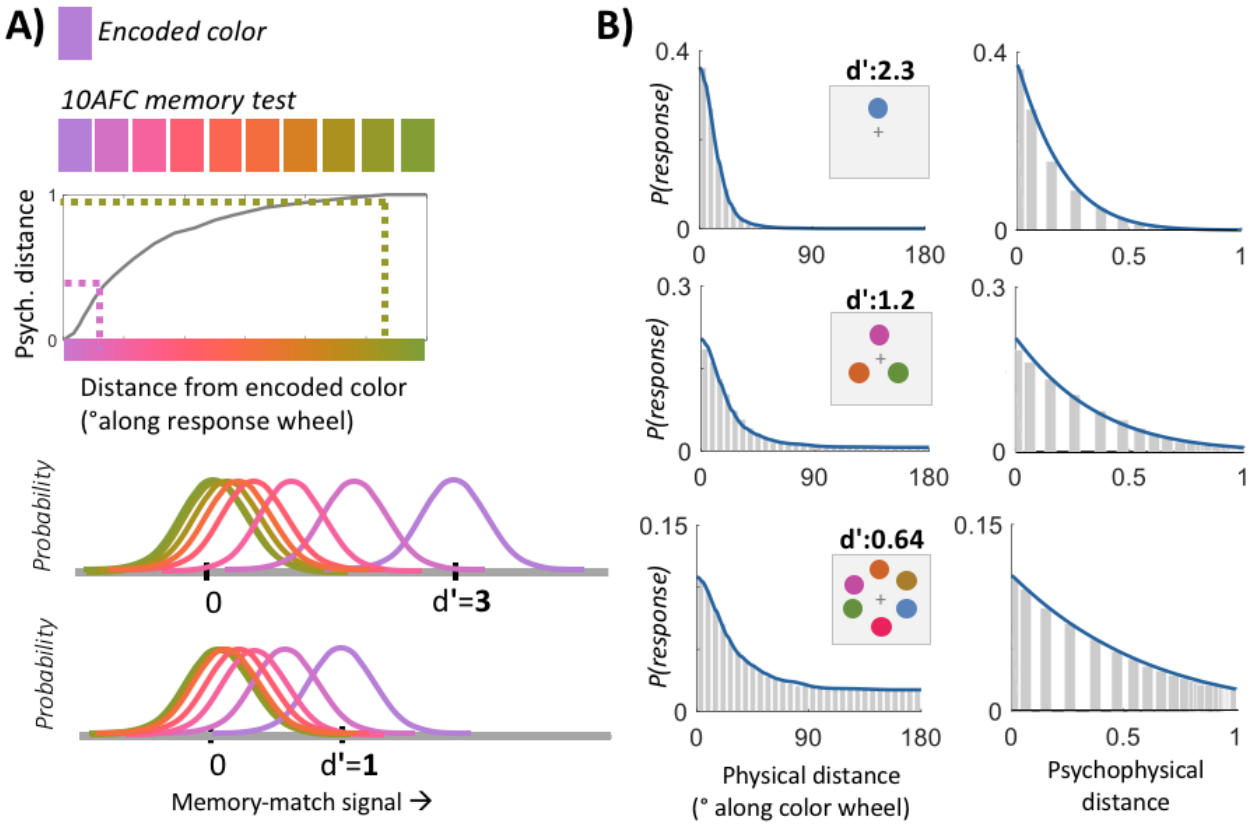

Figure 2. *(**A**) Signal detection model applied to memory in a hypothetical 10-alternative forced-choice test. In standard 2-alternative long-term recognition memory experiments, "new" items sometimes feel more familiar or less familiar. The memory-match signals generated by new items are therefore modeled as a normal distribution with a mean of 0 and standard deviation of 1. By contrast, "old" items (encoded previously) elicit higher memory-match signals, modeled (in the simplest case) as a normal distribution with a mean of d' and standard deviation of 1. Performance is best in conditions where d' is high. 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 memory-match strength 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. We place the mean of each lure's memory-match distribution at its psychophysical distance to 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 memory-match signal than the green lure because it has a much shorter psychophysical distance 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. The TCC is simply an extension of the 10-AFC logic provided here, but adapted to every possible color probe (i.e. 360-AFC). If we assume that only a subset of colors are functionally probed at test (e.g., 60-AFC), the d' estimates would change, but the fit of the model would not be appreciably affected. (**B**) TCC fits to data at set size 1, 3, and 6. With only a single parameter (d'), the TCC accurately captures the response distribution typically attributed to multiple parameters / psychological states by existing frameworks and models of working memory. On the left, we plot the probability of a response relative to the absolute value of physical error. A more appropriate way to plot these data is relative not to the physical distance, where it appears to be a normal distribution on top of a long tail, but to plot them relative to psychophysical distance (as shown on the right; which depicts the same bars plot -- that is, how often each response option was chosen -- but now with equal steps in psychophysical distance along the x-axis). In this space, there is neither a discontinuity nor a long-tail, indicative of why the TCC is able to fit this data straightforwardly. Note that plotting the data this way results in a clustering of the response distribution bars in the 90° to 180° physical distance range because they all fall at almost the same psychophysical distance from the target.*

This straightforward model is broadly applicable across working memory paradigms. A distinct advantage of taking into account psychophysical distance is that it provides a unifying theory of working memory across many tasks previously thought to measure distinct psychological constructs. For example, tasks that measure large changes to a set of items across a delay (i.e., a change-detection task; see Figure 3A) are thought to be insufficient to capture working memory performance as they only measure 'capacity' (how many items are remembered) but not 'precision' (how precisely they are remembered)[17]. However, the TCC provides a natural generalization from change detection tasks to continuous report tasks and vice-versa, demonstrating that this seemingly fundamental distinction in the working memory literature between capacity and precision is illusory.

We had participants perform a standard change detection task at set size 2 and set size 5, with the probe colors being either an exact match to the item from that location or the maximal 180° away in physical color space (Figure 5A). We then obtained estimates of *d′* at both set sizes, and we used these *d′* estimates and the separately measured psychophysical scaling function to infer the distribution of responses from a continuous report task where all 360 colors were available as response options (Figure 3B). We found an excellent fit between this prediction and independently collected continuous report data from different participants at set size 2, R^2=0.991, p<0.00001, and at set size 5, R^2=.995, p<0.00001 (Figure 3C). Thus, only measuring how well participants can detect large (180-degree) changes was sufficient to completely explain the distribution from continuous report studies. This result demonstrates that change detection, even with extremely large and easily discriminable differences, is sufficient to fully characterize working memory. This discovery allows the integration of a huge literature on this task, with important theoretical[18] and clinical implications[19]. Critically, with an independent estimate of *d′* in hand from the change detection task, the TCC's fit to continuous color wheel performance is parameter free.

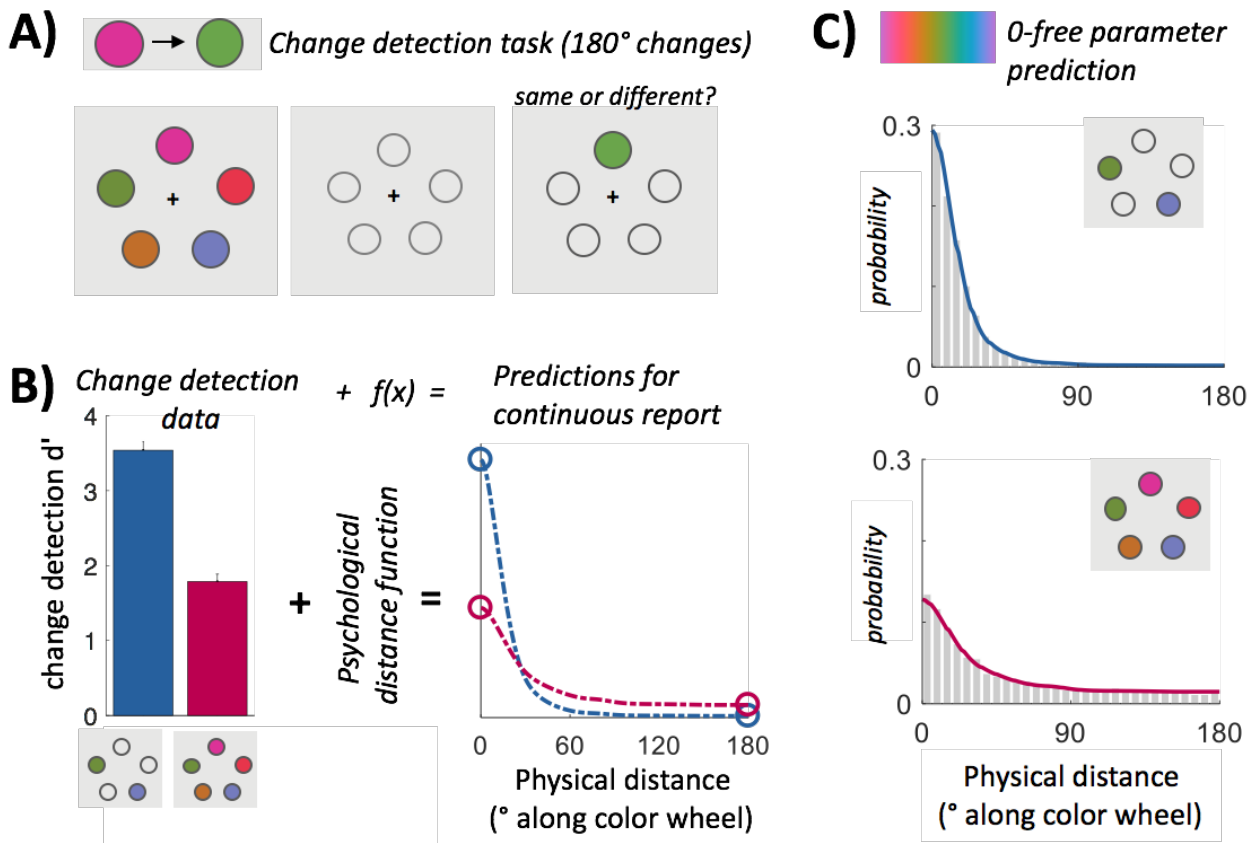

Figure 3. *(**A**) Illustration of change detection task. After a delay, participants are probed on a single item from the previous display, judging whether it was the same or different. If it was different, the foil color was always 180° away in color space, and thus according to existing models should only measure capacity, not the precision of working memory. Participants performed this task at set size 2 and set size 5. (**B**) Using the change detection data (error bars represent within-subject SEM) and psychophysical scaling data, we generated a predicted continuous report distribution for both set size 2 (blue) and set size 5 (pink). (**C**) Remarkably, the predicted distributions accurately fit separately collected continuous report data from an independent set of individuals. Thus, taking into account psychophysical distance allows the TCC to accurately predict continuous report data, even though such data are typically assumed to reflect multiple parameters that cannot be measured in change detection with large-scale color changes (i.e. capacity, precision, etc.).*

Generalization across different stimulus spaces. We next applied the TCC to stimulus spaces outside of color. We consider first the case of orientation, another space widely used to study working memory performance. We measured the psychophysical distance function for orientation using the quad MLDS task[13], which is designed to minimize the possibility of relational encoding in the psychophysical scaling task (Figure 4A). Again, we found the TCC fit

previously observed memory data from different individuals remarkably well across set sizes 3 and 6 (set size 3: R^2=0.995, set size 6: R^2=0.987, all p<0.001; see Figure 4B; data courtesy of Rademaker et al.[20]).

Next we examined the case of face identity. Using a previously created face identity continuous report procedure[21], we collected memory data for set size 1 and 3. We also measured the psychophysical distance function using the quad task and MLDS (Figure 4C). Again, we found the TCC fit observed memory data remarkably well across both set sizes 1 and 3 (set size 1: R^2=0.99, set size 3: R^2=0.99, all p<0.001; see Figure 4D).

Thus, the TCC accounts for data across a variety of stimulus spaces: as long as the perceptual space of the stimuli is measured using psychophysical scaling, the TCC's straightforward signal detection account with only a singular d' parameter accurately captures the data.

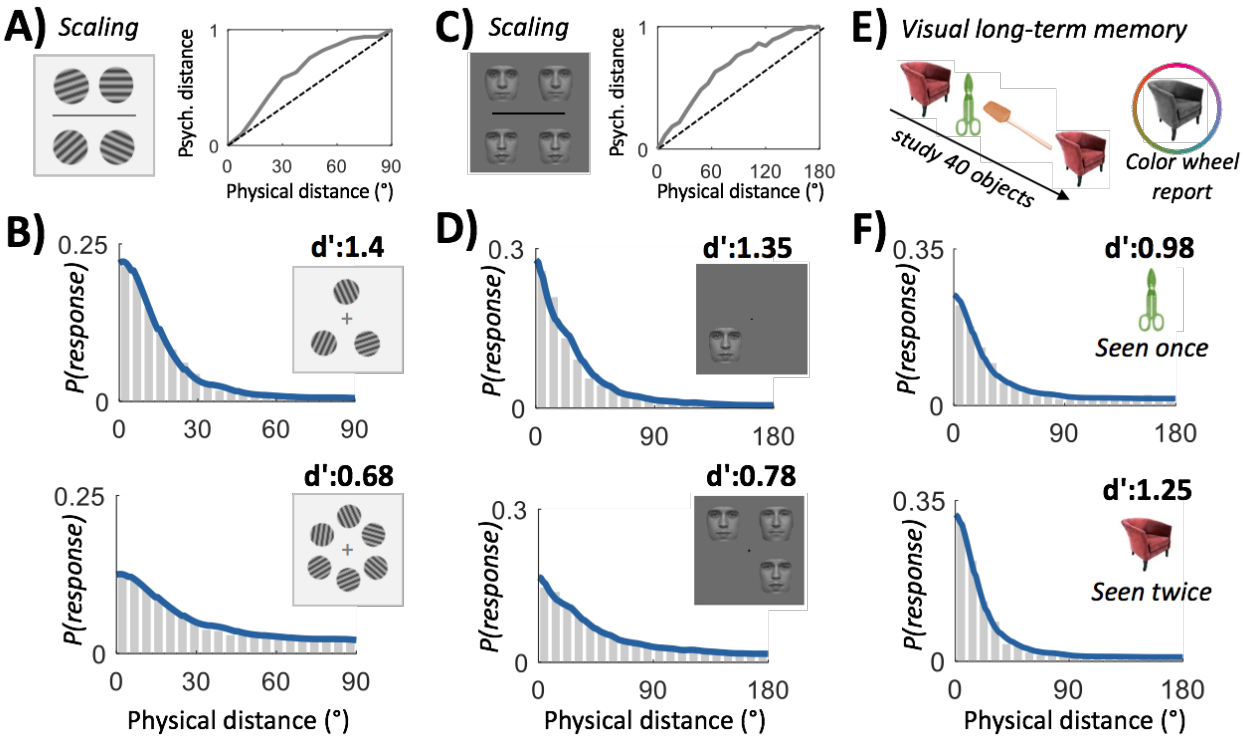

Figure 4. *(**A**) Using the same MLDS method as we did for color, we collected a psychophysical distance function for orientation. (**B**) TCC fits to working memory data using orientation at set size 3 and 6. (**C**) Using the same MLDS method, we collected a psychophysical distance function for face identity. (**D**) TCC fits to working memory data using face identity at set size 1 and 3. The TCC accurately captures face identity data, demonstrating its generalizability across diverse stimulus spaces. **(E)** To show generalization to other memory systems, we collected data on a visual long-term memory continuous report task with colors. 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. **(F)** TCC fits to visual long-term memory data for items seen only once and for items repeated twice. The TCC accurately captures visual long-term memory data.*

Generalization across different memory systems. To demonstrate the TCC's applicability to multiple memory systems, not just visual working memory, we collected data on a visual long-term memory continuous report task with colors (similar to that previously used[22]). Participants performed blocks where they sequentially saw 40 real-world objects' that were randomly colored (a number of colors far greater than working memory capacity), 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. Some items were seen only once, and

some repeated twice in the same color within a block (Figure 4E). Again, we found the TCC fit the observed memory data remarkably well across both the unrepeated and repeated items (seen once: R^2=0.978, seen twice: R^2=0.985, all p<0.001; see Figure 4F). Thus, unlike working memory modeling frameworks which propose system-specific mechanisms (e.g., population coding and divisive normalization[23]), the TCC naturally accounts for data across both working memory and visual long-term memory.

Discussion

Previous theories and models of visual working memory have not considered the non-linear relationship between physical stimuli and the psychological representation of those stimuli. This oversight has led to models based on "mixtures" of remembered and not-remembered items that quantify 'precision' and 'capacity' separately. Mixture models have been highly influential, leading to applications across a variety of domains beyond working memory, including attention[24], iconic memory[25] and long-term memory[26]. In fact, strong claims about deficits in clinical populations[27-29], and even the underlying nature of consciousness[30] have been made based on the supposedly discrete failures of memory proposed by these models. The TCC shows that the apparent dissociation between capacity and precision is an illusion of modeling the data in physical space, as the same aspects of the data arise naturally from a single parameter model with no distinction between 'capacity' and 'precision' once psychophysical scaling is taken into consideration. Existing data that have been interpreted in terms of these models -- and closely related models that propose some items are represented with extremely low precision rather than supposing those items are not remembered[8] -- must be reconsidered in light of this new framework (Supplemental Material, section 1). Thus, in addition to providing a unifying theory of visual memory capable of bridging distinct tasks and stimulus conditions that would not be possible using these previous models, the TCC suggests a new psychological interpretation for many previous results, providing evidence that distinctions between 'capacity'

and ‘precision’ arise from changes in overall memory strength combined with non-linear psychophysical scaling.

The TCC provides a compelling connection between working memory and long-term recognition memory, which is often conceptualized in a signal detection framework. In this framework, memory strength necessarily affects both how much evidence we have for our responses (e.g., how confident we will be in them, how likely we are to believe we are guessing, etc.) and how ‘precise’ they are (e.g., how likely we are to confuse nearby items with the target item) in a connected manner. In addition, the TCC (like all signal detection models) is based fundamentally on the idea of sampling (and potentially re-sampling[31]) 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. This signal detection framework 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[20,32] 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’[31]. Thus, despite research on working and long-term memory operating largely independent of one another, the TCC provides a unified framework for investigating the distinctions and similarities in memory across both domains.

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.

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. Mechanical Turk users form a representative subset of adults in the United States[33,34], and data from Turk are known to closely match data from the lab on visual cognition tasks[35,36], including providing extremely reliable and high-agreement on color report data[31]. 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. This circle was centered in the color space at (L = 54, a = 18, b = -8) with a radius of 59. 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.

2. Psychophysical scaling quad experiment. N=100 participants on Mechanical Turk judged which of two pairs of colors presented were more distinct (e.g., which pair had constituent items that were more different from each other). On each trial, we chose 4 colors without replacement from the set of 360 colors along the response wheel at steps of 5 degrees, sorted them and presented these in pairs, according to the logic of Maloney and Yang's[13] Maximum Likelihood Difference Scaling method.

Participants were asked to make their judgments solely based on intuitive visual similarity, rather than the use of knowledge of color categories and the color wheel. To encourage an intuitive strategy, we asked participants to repeat the word 'the' for the duration of the trial to minimize the use of verbal strategies, and set a response deadline of 2 seconds for each trial. If participants took greater than 2 seconds to respond, the trial was terminated and a new trial was presented. Each participant completed 210 trials, the first 10 of which were practice trials. Using our a priori exclusion rule, we excluded any participant whose overall accuracy was 2 standard deviations below the mean (M=49%) leading to the exclusion of 9 of the 100 participants.

To compute psychophysical distance from these data, we used the model proposed by Maloney and Yang[13], the Maximum Likelihood Difference Scaling method. This method can be used to measure the appropriate psychophysical scaling of distances between colors as a function of the physical placement of colors along the wheel per se. In particular, any given trial has two pairs of colors, where their physical color 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 distance to which this physical 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 $\psi_{ij} + \varepsilon > \psi_{kl}$. We set the standard deviation of the Gaussian ε noise to 1, so that the model has 36 free parameters, corresponding to the psychophysical scaling values for each possible physical interval length (e.g., how long a physical interval 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 distance ranges from 0 to 1.

3. 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 0, 3, 5, 8, 10, 13, 15, 20, 25, 30, 35, 45, 55, 65, 75, 85, 100, 120, 140, 160, or 180 away from the target color. 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.

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 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 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 distance from this data, we used the model proposed by Maloney and Yang[13], the Maximum Likelihood Difference Scaling method as explained in the quad experiment above. The only difference is that in the triad experiment, S_i was always the same as S_k (e.g., both pairs included the target color).

4. 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 a commonly used response circle in the CIE *L*a*b** color space. This circle identical to that used in the fixed distance triad experiment. 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.

5. Description of Target Confusability Competition 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. 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, $m_x \sim N(d_x, 1)$ (Figure 3A), (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 distance to the target according to the empirical psychophysical distance function ($f(x)$), such that $d_x = d'(1-f(x))$ (Figure 2).

As $f(x)$, the psychophysical distance function, we use the smooth function estimated from the similarity experiment rather than the model fits from the Maloney and Yang[13] model for the triad or quad task. In the similarity experiment, 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 $f(x)$, the psychophysical distance function, we simply normalize this data to range from 0 to 1, giving a psychophysical similarity metric; and then subtract it from 1 to turn it into psychophysical distance, such that $f(x) = 1 - ((S_x - S_{min}) / (S_{max} - S_{min}))$.

Thus, according to the model, the mean memory-match signal for a given color x on the working memory task is given by $d_x = d'(1-f(x))$, where d' is the model's only free parameter. When x =

0, $f(x) = 0$, so $d_0 = d'$. By contrast, when $x = 180$, $f(x) = 1$, 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, $m_x \sim N(d_x, 1)$. The response on a given trial is made to the color on the wheel that generates the maximum memory-match signal, $r \sim argmax(m)$.

Thus, once the psychophysical distance function has been collected (Figure 2A) and interpolated to have a value between 0 and 1 for every 1 degree in distance from the target, the full code for sampling an absolute value of error from the model is only two lines of MATLAB:

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

Although the model assumes that every possible color probe elicits a memory-match signal depending on how well it matches the color of the encoded item from the probed location (e.g., the task is a 360-AFC), the predictions of the model are not significantly affected if we instead assume that subjects randomly sample a subset of the colors on the color wheel or if we scale the number of options to the limit (e.g., $n = 36$ or n= 36,000) when trying to decide which color appeared in the cued location (only the inferred d' changes as a result, not the shape of the predicted distributions).

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.[17] 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 included, e.g., the response on a given trial, rather than being *argmax(m)*, can be assumed to include motor noise of some small amount, for example, 2°:

$$r \sim N(argmax(m), 2°)$$

In MATLAB notation:

```
reportedError = randn*motorNoise + memoryError;
```

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.

Fits of the free parameter of the model to datasets were performed using the MemToolbox[38] making use of maximum likelihood.

6. Continuous color report data (set size 1, 3 and 6). The continuous color report data used for fitting the model at set size 1, 3 and 6 was previously reported in Brady and Alvarez[37]. In that study, N=300 participants each performed 48 trials of continuous color report at set size 1, a different group of N=300 participants did so at set size 3 and yet another group of N=300 at set size 6. At both set size 3 and 6, participants were probed on 3 of the colors on each trial sequentially. Thus, the effects of delay and response interference can be probed using these data as well (see Supplemental Material, section 11). This study used the same color wheel used in the current studies and was also conducted on Mechanical Turk to allow for direct comparison to the current similarity data.

7. Change detection experiment and associated set size 2 and 5 continuous report data. N=60 participants on Mechanical Turk performed a change detection task. These data were used to estimate *d′* at set size 2 and 5 in this particular set of task conditions. Then a distinct set of N=50 participants performed a continuous response task, using similar task conditions and also at set size 2 and 5. The *d′* from the change detection task was then used to a priori predict the response distribution of the participants in the continuous report task. Both tasks used the standard fixed luminance CIELa*b* color wheel.

Change detection task. Participants performed 200 trials of a change detection task, 100 at set size 2 and 100 at set size 5. The display consisted of 5 placeholder circles. Colors were then presented for 1000ms, followed by an 800ms ISI. At set size 2, the colors appeared at 2 random locations and placeholders remained 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. We used these data to calculate a *d′* (i.e., d_{180}) separately at set size 2 and set size 5.

Continuous report task. The task was identical to the change detection task except that rather than a probe color reappearing at test, one of the placeholder circles 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.

Fitting data from continuous report with change detection data. If its assumptions are correct, the TCC can use the measured change detection *d′* for each set size (involving test trials with discrete colors either 180° or 0° from the target) to predict the corresponding continuous report data (involving a color wheel with colors ranging from +/-180° to 0° from the target). The change detection *d′* is directly related to, but is not identical to, the *d′* value used by TCC. The reason is that the actual TCC *d′* value for the continuous report task is determined by the unknown number of colors sampled at test (which is not necessarily equal to 360). To estimate the TCC *d′* from the change detection *d′*, we followed these steps:

(1) Evaluate the probability density function of TCC at -180° and 0° (the two options in the change detection task).

(2) Assume that in a forced-choice task where participants were given those two options, they would select them proportional to their probability density function values. Thus, we use the ratio of the probability density values at 0° and -180° to calculate a *d′* value for a 2-AFC task that we would expect for this TCC *d′* value.

(3) To adjust for the fact that our participants performed an old/new (change detection) task, not a 2-AFC task, divide the corresponding *d′* by √2 to arrive at the corresponding change detection *d′* for this TCC *d′* value[39].

Thus, to map from a *d′* value in the change detection task (d'_{cd}) to a *d′* in the TCC model, we make use of the following relationship:

$$p = \frac{g_{d'}(0°)}{g_d(0°) + g_{d'}(-180°)}$$

$$d'_{cd} = \frac{\Phi^{-1}(p) - \Phi^{-1}(1-p)}{\sqrt{2}}$$

Where $g_{d'}(x)$ is the probability density function of TCC for a given TCC d' value, and Φ^{-1} is the inverse cumulative normal distribution.

8. Orientation continuous report data (set size 3 and 6). The continuous color report data used for fitting the model at set size 3 and 6 was previously reported in Rademaker et al.[20] and made available by van den Berg et al.[40].

9. 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[21]. 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 considered practice and not included in the analysis.

10. Visual long-term memory color report task. 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. Participants were highly accurate at old/new (d': 2.92, SEM: 0.20), and our main

analysis consists of analyzing the pattern of errors at color report for old items that were reported old (e.g., the only items they saw and made a color report for).

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. Brainard, D.H. (2003) Color appearance and color differences specification. In *The Science of Color* (2nd edn) (Shevell, S.K., ed.), pp. 191–216, Optical Society of America.

10. Fechner, G. (1860). *Elements of psychophysics*.

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

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.

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

15. 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.

16. Crowder, R. G. (1976). *Principles of learning and memory*. Hillsdale, NJ: Erlbaum

17. 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.

18. Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. Nature, 390(6657), 279.

19. 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.

20. 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.

21. 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.

22. 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.

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

24. 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.

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

26. 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*.

27. 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.

28. 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.

29. 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

30. 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.

31. 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.

32. 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), 10-65.

33. Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). Evaluating online labor markets for experimental research: Amazon. com's Mechanical Turk. *Political Analysis*, *20*(3), 351-368.

34. Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data?. *Perspectives on psychological science*, *6*(1), 3-5.

35. 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.

36. 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.

37. 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.

38. 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.

39. Macmillan, N. A., & Creelman, C. D. (2004). *Detection theory: A user's guide*. Psychology press.

40. van den Berg, R., Awh, E., & Ma, W. J. (2014). Factorial comparison of working memory models. Psychological review, 121(1), 124.

41. 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.

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

43. 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.

44. Zhang, W., & Luck, S. J. (2009). Sudden death and gradual decay in visual working memory. *Psychological science*, *20*(4), 423-428.

45. 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.

46. Fougnie,D., Cormiea, S. M., Kanabar, A., & Alvarez, G. A. (2016). Strategic trade-offs between quality and quantity in working memory. *Journal of Experimental Psychology: Human Perception & Performance, 42*(8), 1231.

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

48. Zawadzka, K., Higham, P. A., & Hanczakowski, M. (2017). Confidence in forced-choice recognition: What underlies the ratings?. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(4), 552.

49. Fougnie, D., Brady, T. F., & Alvarez, G. A. (2014). If at first you don't retrieve, try, try again: The role of retrieval failures in visual working memory. *Journal of Vision*, *14*(10), 851-851.

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

Acknowledgements

We thank Viola Störmer, Rosanne Rademaker, Jeremy Wolfe, Daryl Fougnie and Talia Konkle for comments on these ideas and on the manuscript. 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)

Supplemental Material

(1) The TCC captures existing working memory data.

Contrary to previous frameworks guiding working memory research, the TCC predicts there is only one underlying parameter (*d′*) underlying working memory performance. However, a huge variety of existing literature has measured multiple parameters of the error distributions and claimed that not only can multiple parameters be measured, but these parameters can be dissociated. Thus, we sought to further support the TCC's claim that there is in fact only a single parameter responsible for working memory response distributions by predicting the observed relationship between *d′* in TCC and the parameters of the most widely-adopted model currently used to analyze working memory performance across a variety of previously published papers. This model has two parameters, precision (in terms of SD) and guessing rate, used to quantify capacity[7]. The TCC framework claims that the parameters of previous models are not psychologically interpretable because they are based on an uninterpretable x-axis; nevertheless, they do provide a way to summarize response distributions from previous working memory experiments.

We searched for papers that used these mixture models techniques by finding papers that cited the most prominent mixture modeling toolboxes, Suchow, Brady, Fougnie & Alvarez[38] and Bays et al.[41]. 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 in CIE $L^*a^*b^*$ color space (as different color spaces may have different scaling). 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[42,43].

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 15 papers and 56 data points were included.

After simulating data obtained from the TCC over a range of *d′* values, and fitting the simulated data from the TCC with Zhang & Luck's[7] model, we obtained estimates of this model's two parameters ("precision" and "guess rate") for each level of *d′*. We then compared this to the reported values from studies utilizing this mixture model, and found that manipulating TCC's one parameter (*d′*) accurately predicted the previously observed data. That is, 'guessing' and 'precision' seem to be highly correlated with each other in exactly the manner predicted by the TCC (Figure S1). This provides additional evidence that once psychophysical distance is taken into account, one parameter is sufficient to capture much of the data observed in working memory tasks.

So how has previous literature found dissociations between these parameters? The majority of these dissociations are based on findings that the precision (SD) does not change when the 'guess rate' (or capacity) does change[22,44]. This is naturally explained by the TCC, as at low *d′* values, 'guess rate' can change by a huge amount with SD changing only a few degrees -- and a difference between SD=21 and SD=24, for example, is visually indistinguishable and would require high power to detect. 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 changes in 'guess rate'[45,46]. 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 comparing across different sets of stimuli with different psychophysical scaling functions (e.g., the Zhang and Luck manipulation adds color noise to the items, making them less perceptually distinct), which is perfectly consistent with the TCC, as different psychophysical scaling functions result in a different relationship between the central distribution and the tail of the distribution. Thus, in addition to providing a way to generalize across stimuli and tasks, the TCC also naturally explains existing working memory data and does so more parsimoniously than previous models.

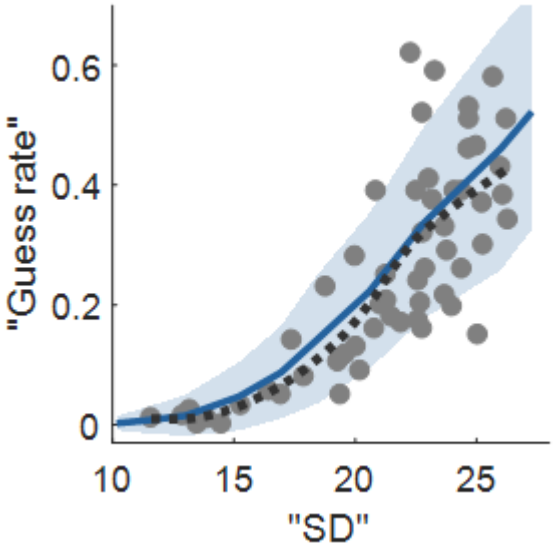

Figure S1. Analysis of previous literature measuring the most widely used model parameters currently used to analyze working memory performance (precision [SD] and guessing). Gray dots are values reported in papers found in the literature; the dashed black line is a LOESS (local regression) smoothed version of these points. The solid blue line reflects the average "guess" and "SD" parameters when fitting data sampled from TCC with a mixture model, 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.

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 15 papers.

(2) Natural extensions of the Target Confusability Competition Model

In this manuscript we present a straightforward version of the TCC, a signal detection model where each lure's memory-match strength is simply its psychophysical distance from the target. This model captures a wide variety of data without any additional factors included. However, some of the psychological concepts identified as parameters in previous models of working memory (for example, variability in encoding across trials) are likely to be true and potentially important aspects of working memory, even if the TCC can accurately account for the pattern of data without considering them.

The role of "swap" errors. A common component typically included in more recent adaptations of working memory models is "swapping," or errors related to misreporting an incorrect item in the remembered display. Adding a swap component to models of working memory generally improves the fit of these models, although it is not critical to be able to tell different models apart from one another, as such a parameter can be built into all models. The continuous report data we use here, because they provide long encoding times and reliable locations from trial-to-trial with placeholders for where the items will appear, do not contain a significant amount of "swaps"[37]. However, in other contexts this component could become more relevant. In the context of the TCC, as in all other models, a swap component could be added as an additional parameter. In particular, rather than the memory-match signal that is entered into the signal detection model being based purely on psychophysical distance to the target, this signal could be a weighted average of psychophysical distance to the target as well as to the distractor items on the trial. This would naturally allow the TCC to account for swaps.

Variability in encoding across trials. 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 remembered with extremely low precision), this does not

mean that some variability in memory strength is not present in working memory tasks. In particular, there is direct evidence that there is at least some variability in memory performance across items and trials[20,47]. Certain working memory tasks may also contain manipulations that would make variability in d′ across items more likely, such as pre-cueing a particular item.

Importantly, however, the TCC can naturally account for the 'long tails' of the error distribution without reference to extremely low precision or un-represented items. Thus, rather than assuming that item-by-item variability explains all aspects of the "long-tails" in response distributions and is a necessary component to fit even aggregate data[8], this variability should instead be considered in light of the TCC framework. The TCC is easily amenable to allow variability in d′ between items and/or trials, although we do not expand it in this direction in the current work. Variability across items in previous models was primarily added to previous models to account for the relative "peakiness" of response distributions[8], which for the current data is naturally captured under the single parameter of the TCC because it is captured by the shape of the psychophysical distance function. However, the data we use here may have little variability between items and trials, since we provided long encoding times and reliable locations from trial-to-trial with placeholders between trials. Variability in *d′* would be more likely to occur under certain circumstances / task designs. Expanding the TCC to allow variability in *d′* between items could potentially better elucidate the dynamics of how variability in encoding affects working memory performance, while remaining in the same underlying signal detection framework.

(3) Implications of the signal detection framework for working memory

The TCC provides strong evidence in favor of a signal detection framework for working memory. Such a signal detection framework has profound implications for our theoretical conception of working memory. A key component of previous frameworks guiding working memory research

has been that observers are either in a “memory state,” with relatively precise information about a previously presented item available, or they are in a “guess state” with no- or (extremely) low-information available. Under signal detection theory, there is no distinction between “memory” and “guess” states as previously theorized, but rather both high and low confidence and high and low error responses arise naturally through the same underlying process.

A signal detection model does not deny the obvious fact that “guessing” occurs, but it does offer a different interpretation of what it means to guess. In a threshold model, guessing is not based on a psychological evidence variable and is akin to a coin flip. In a signal detection model, guessing refers to a decision based on a psychological variable that happens to be largely non-diagnostic. The distinction is subtle but critical. Given a signal detection framework with $d' > 0$, “guessing” will occur when the sampling process at decision-making yields a maximum memory-match signal that falls below a low criterion (e.g., all the colors return weak memory-match signals[48]). This will result in low-confidence responses associated with near-chance accuracy. Subjectively, these decisions ‘feel like’ (and essentially are) guessing because the memory-strength evidence variable does not strongly support any item being old. Other trials may well yield higher memory-match signals that fall above a high criterion, strongly supporting the decision that an item is old and thereby resulting in high-confidence responses. On these trials, observers will usually make the correct decision, and the strong memory signal will ‘feel like’ an above-threshold memory state. They key point is that all decisions are based on the strength of a memory signal, whether that signal is largely non-diagnostic on a given trial because it is very weak or is extremely diagnostic because it is very strong.

Guessing could also naturally arise if d' is 0 if, for example, a “target” item was not even perceived at encoding (e.g., because the participant did not attend to one of the colors, or if one color was not even shown but was then probed). Because at low d', and especially at $d' = 0$, all

the memory-match signal distributions are likely to lead to very low samples and trials will subjectively feel like guessing (and, objectively, overall performance will be no better than chance). But it is still not guessing in the sense assumed by previous models of working memory postulating a “guess state”. In other words, the decision is based on a psychological random variable (i.e., it is based on a memory-match signal) drawn from a distribution, not on a coin flip.

An advantage of conceptualizing working memory using signal detection theory is that it explains phenomena that cannot be naturally explained by existing frameworks of working memory. For example, imagine a working memory experiment with eight color positions, but where only 4 colors are presented during each trial. Thus, at response observers should have 4 positions with $d' > 0$ and 4 positions with $d' = 0$ (no information). If a no-information position is cued, the subject will likely feel like they are guessing most of the time, and performance will be at chance (with responses distributed across the color wheel). But sometimes on no-information trials, the memory signal will be such that subjects sense a non-trivial (albeit false) match for one of the colors on the wheel, and they’ll have a corresponding degree of confidence that they are right. On rare occasions, the spurious match signal will be strong and they will have high confidence (because according to signal detection theory the brain sometimes yields a strong memory-match signal due to “noise,” even when there is no memory trace). As an analogy from perception, you may sometimes feel your cell phone vibrate (announcing the arrival of a text message) even when it did not vibrate. When you are on high alert for a very weak signal, you sometimes “detect” that signal even when it is not there. Previous models of working memory assuming a no- or low- information “guess” state assume this never happens. As a result, previous models of working memory have no principled way to account for high-confidence decisions that sometimes occur in the absence of a true signal, whereas a signal detection framework naturally explains such phenomena.

Thus, the TCC's signal detection framework allows us to interpret data showing that even when people respond very far from the true color (thought of as a 'guess' or 'extremely low precision item' in previous models), they often still have information if they are asked to respond again. In a signal detection framework, it is natural to assume that on trials with large errors that are typically thought of as "guesses" or extremely low precision items, asking participants to try to respond again should result in a new set of samples from the signal detection model and thus should frequently result in a more accurate response. This prediction has been borne out by previous work[31,49], and is strongly supportive of a signal detection framework, giving further credence to the TCC.

Broadly, the signal detection framework provides a tool for understanding how responses are generated without requiring participants to have distinct psychological states like "guessing".

(4) How to measure psychophysical scaling of a given stimulus space

Our psychophysical distance function naturally captures two key aspects of how stimuli are perceived: The relationship between the physical stimulus and the psychological representation of that stimulus is rarely linear[10], and the similarity between stimuli as a function of distance is additionally non-linear[11]. It is necessary for the psychophysical distance function to capture both these points - a function that only measures one of these factors will fail to accurately capture perception of the stimulus space, possibly resulting in the artifact of "long-tails" erroneously being preserved.

In the current manuscript, we present several examples of tasks that naturally captures both these insights and can be translated to a psychophysical distance function, including the triad task, the quad task and a subjective similarity judgment (see Methods). It is important note that

depending on the number of trials, a large number a data points (i.e. subjects) may be necessary in order to obtain reliable estimates of a given stimulus space in the triad and quad tasks. A similarity task may be sufficient to capture this function under some circumstances, like for color in the current study. Participants can be asked to rate the similarity of two items (varying in distance from one another) on a Likert scale from 1 to 7, and these ratings can then be normalized. We observed this similarity rating task in color provided another measure of psychophysical distance that is in close agreement with the results of the quad and triad tasks and requires considerably less data to estimate. However, it is important to note that depending on the stimulus space, observers may utilize different strategies in such subjective similarity tasks, and that ultimately an objective task like the quad task may be best to understand the psychophysical distance function. In particular, 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[50] and Torgerson[12].

(5) Capturing non-uniformities in feature spaces.

Many stimulus spaces (whether low- or high-level) contain non-uniformities, which may affect subsequent working memory performance. Indeed, Bae et al.[15] discovered non-uniformities in working memory for color, where responses for targets tend to 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[14,15].

A key advantage of the TCC is that by taking into account the psychophysical distance function, non-uniformities within whatever feature space being probed can be automatically captured if psychophysical distance data is measured separately from each relevant starting point in the feature space. In the current work, we use only a single psychophysical distance estimate averaged across possible starting points. However, this is not necessary to the TCC framework, and the framework naturally accommodates non-uniformities in feature spaces. For example, rather than using a psychophysical distance function that averages over all targets, one could collect a psychophysical distance matrix that provides the specific function for every possible target relative to every possible lure of that target. This would naturally capture any potential non-uniformities or biases within a feature space, which would be particularly relevant for high dimensional features spaces, such as faces. Then any potential biases arising from non-uniformities in color space would be corrected for in the resulting d' estimate.

(6) 3D color space partially explains the non-linearity in confusability

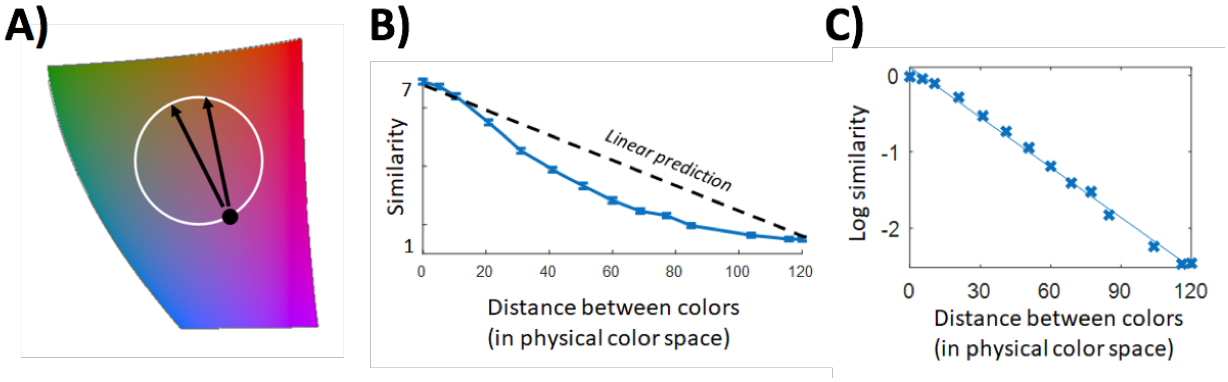

*Figure S2. (**A**) Despite being conceived of as a color wheel, in reality participants internal representation of color -- and thus the confusability between colors -- is a function of their distance in an approximately 3D color space, not their distance along the circumference of an artificially imposed 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 partially attributable to psychophysical distance being a non-linear function even of physical distance across 2D color space. (**B**) The similarity function, as a proxy for psychophysical distance, remains non-linear even in 2D physical 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. (**C**) 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 conceivably reflects perceptual noise/lack of perceptual discriminability between nearby colors.*

(7) Fixed distance triad experiment details

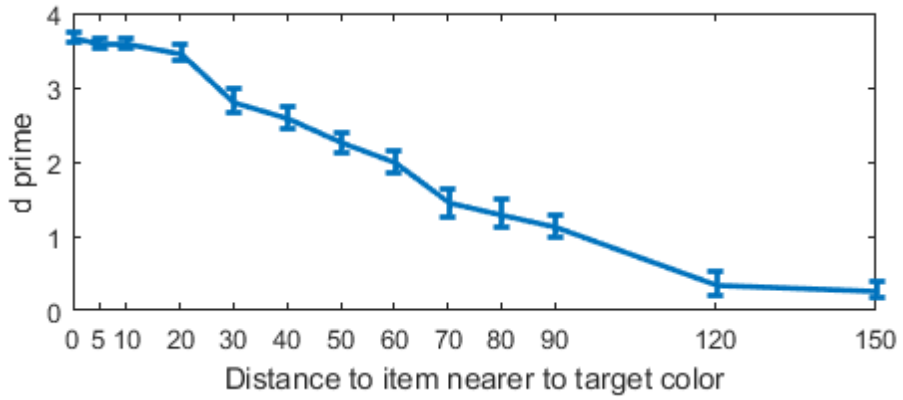

Figure S3. Data from all distances in the fixed distance triad task (Figure 1C). On each trial, there was a target color, here always at 0°, and participants task was to choose which of two other colors was closer to the target color in color space. The two choice colors always differed by 30°. The x-axis 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 this physical space as a psychophysical space.

(8) The space of TCC predictions

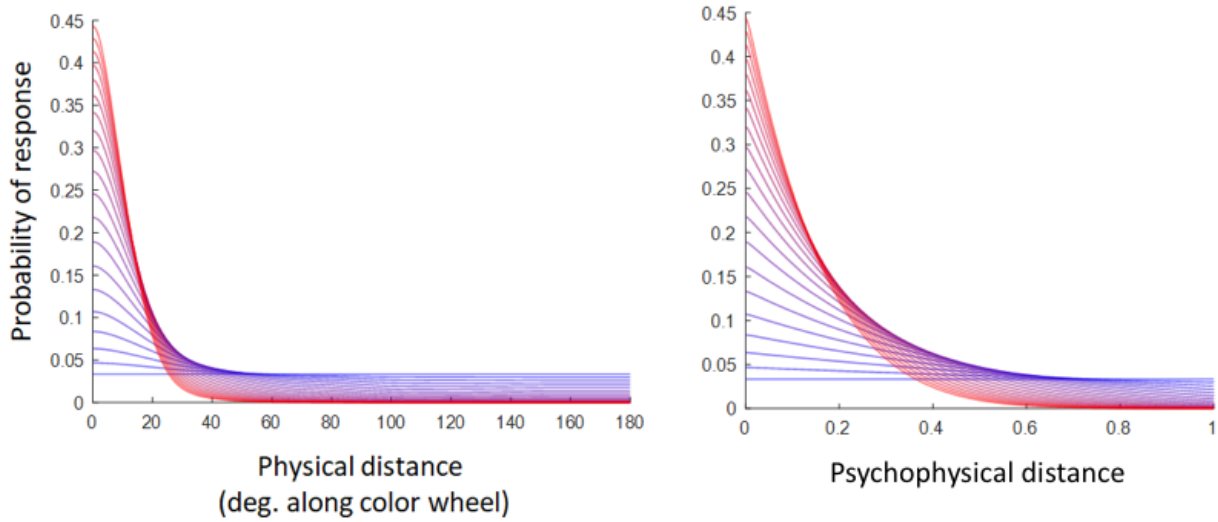

Figure S4. Model predictions from low (blue) to high (red) d', across both physical distance (degrees along the color wheel) and psychophysical distance. Note how the entire function is fully constrained by knowing the height of the 0 point -- this is why a simple change detection task with maximum (180 deg) changes can accurately predict the entire shape of the distribution. In addition, note that while previous models like mixture models can potentially predict a huge variety of functions (as they allow independent modulation of the width of the central distribution, the peakiness of this distribution, and the height of the tails), the TCC can predict only a single relationship between those parts of the function.

(9) Peakiness of the predicted distribution without any variability in precision

A significant literature has used 'mixture models' and in particular a two component mixture model with a uniform distribution and a von Mises distribution to fit the data from working memory tasks and measure 'capacity' and 'precision'[7]. Additional evidence has shown that the Zhang & Luck[7] two component mixture model has significant residuals when fit to the data[8,40,47]. In particular, the distribution of real data is "peakier" in the central part of the distribution than the von Mises distribution predicts.

For example, the main continuous report data reported here, from Brady & Alvarez[37], has tell-tale residuals when compared to this two-component mixture model, with too much data right at the peak of the distribution and too little data just off the peak (see Figure S5A).

To account for this, models now assume that there is variability in the precision of items, e.g., that different items are represented with different width von Mises distributions. Mixing together distributions of different widths can account for this shape, although it adds another free parameter to the models (e.g., variability in precision[47]). Models that incorporate variable precision thus have a peakier distribution, and when samples from these models are fit with the typical two component mixture model, they have similar residuals to the data, with the telltale shape of a peakier-than-normal distribution (see van den Berg et al.[8] for an example of this).

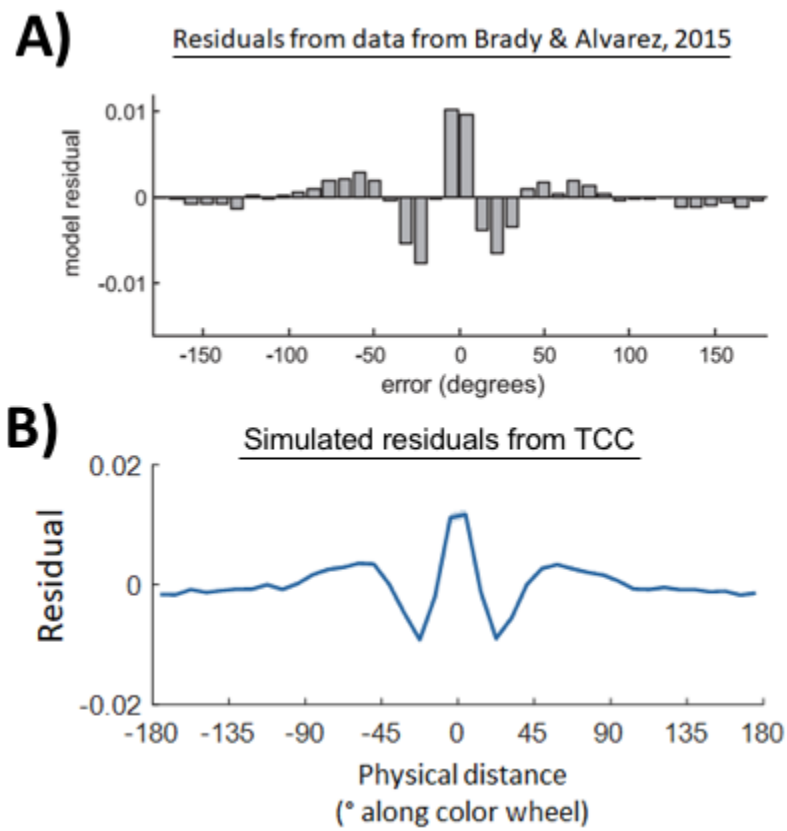

*Figure S5. (**A**) Residuals of a typical visual working memory experiment relative to the standard two component mixture model of Zhang & Luck[7]. (**B**) Residuals of simulated data sampled from the TCC (at d'=1) when fit with the standard two component mixture model. Notice that the residuals look the same as (1) real data residuals when fit with that model, and (2) the residuals used to argue for variability in precision, despite the fact that the TCC has no variability in precision and was used to generate this data.*

One interesting aspect of the TCC, which incorporates measured psychophysical distance data, is that it largely predicts this peakiness without any additional free parameters. That is, the functional form of participants' psychophysical distance judgments naturally results in a peakier-than-normal distribution, suggesting that this peakiness is not direct evidence for variability in precision.

(10) Further generalization across stimulus space (different color)

We collected new data from a continuous color report experiment where rather than using the standard color wheel that spans a large range of color space, we instead showed participants

only colors in the “green/yellow” family and asked them to report their memory on this smaller color wheel.

Small color wheel continuous report data. This experiment was identical to the previous, standard color wheel experiment. The only change was the color space for the items presented, and the color response wheel colors, which were all drawn from the new, smaller color space (slice through L=54, centered at a*=-20 and b*=30, with a radius of 25, focused largely on greens and yellows).

Small color wheel similarity experiment. N=50 participants on MTurk 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 a smaller circle 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 minimize verbal strategies. Each participant completed 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.

According to the TCC, the same model and *d′* parameters should apply without modification to this distinct task once we take into account the different psychophysical distance structure of the green/yellow foils offered to participants. Indeed, we find that the best fitting *d′* parameters from the standard color wheel at set sizes 1, 3 and 6 generalize extremely well to this new color wheel using the TCC (with no free parameters; set size 1: R^2=0.986, set size 3: R^2=0.970, set size 6: R^2=0.900, all p<0.001).

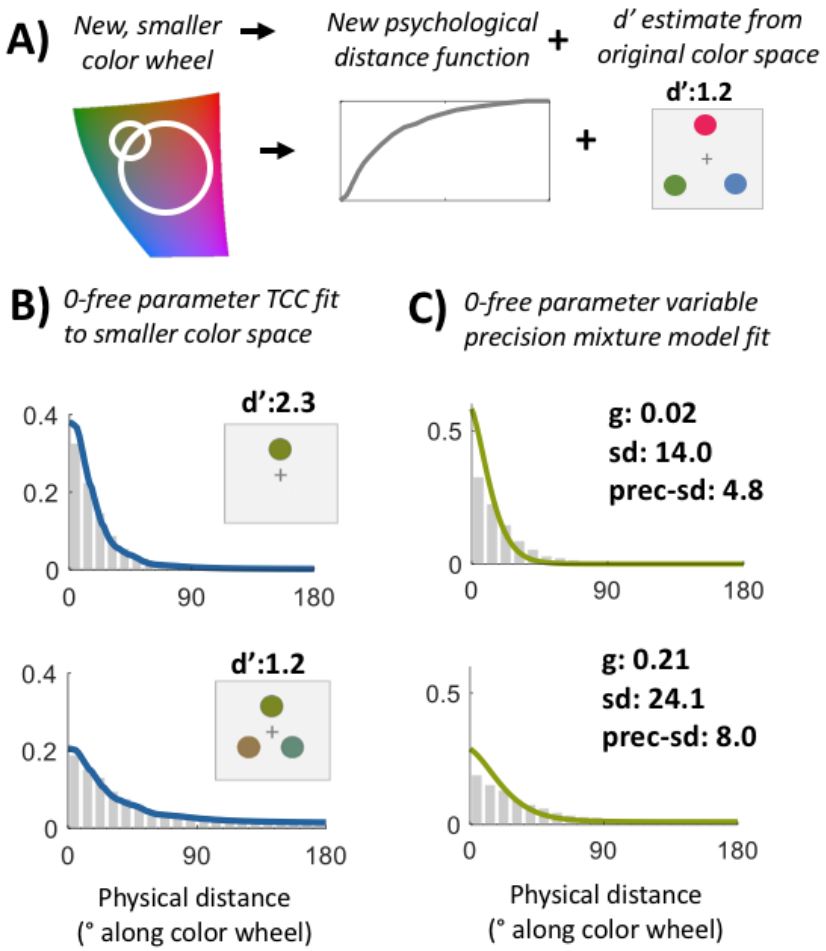

*Figure S6. (**A**) We created a new, smaller color space (consisting of green/yellow colors) and obtained a new psychophysical distance function for this different color space. We then combined this new psychophysical distance function with d' estimates obtained from the original color space. (**B**) 0-free parameter TCC fits to the smaller color space. Across set sizes 1 and 3 the TCC accurately predicts separately collected data from a new, distinct color space. (**C**) 0-free parameter mixture model fits[47] to the smaller color space. A mixture model fit (including capacity, precision, and variability in precision parameters) from the original color space fails to generalize to the smaller color space, demonstrating the limitations of previous working memory models.*

(11) Continuous report delay & response interference

An analysis of continuous report data with different response orders found the TCC's *d'* decreases significantly with increased response delay and interference; Consistent with the long-term memory literature, this result indicates that much more than set size affects *d'* in the

TCC. For example, the *d′* difference between the 1st and 3rd item probed at set size 6 is greater than the difference between set size 3 and 6 overall.

In particular, our continuous report data comes from Brady & Alvarez[37]. In that data, N=300 participants performed 48 trials each of continuous color report at set size 1, a distinct set of N=300 participants at set size 3 and a distinct set of N=300 at set size 6. At both set size 3 and 6, participants were probed on 3 of the colors on each trial. Thus, the effects of delay and response interference can be probed in this data as well.

What we observe at both set sizes is that the first, second and third items probed are all well captured by our model, with *d′* decreasing with increasing delay. Thus, in addition to *d′* decreasing with set size, it is also affected in a smooth manner by other factors, like delay and response interference. This is also evident in the fit of the model to data from a variety of other experiments (Figure S1) which vary along many of these dimensions yet all seem consistent with the TCC.

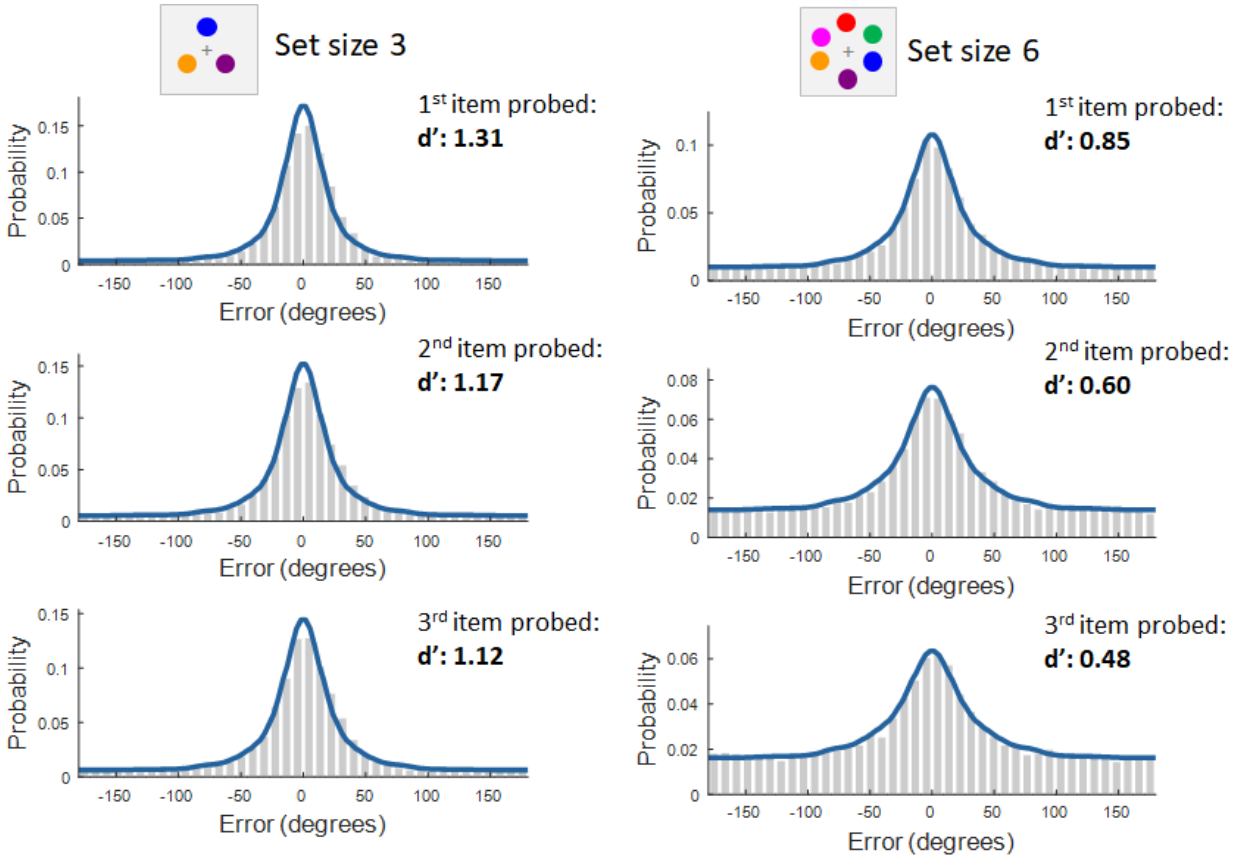

Figure S7. The TCC Model fits for data from Brady & Alvarez[37] with different response orders. Three items were always probed in both set size 3 and set size 6 (top: first item probed; middle: second item probed; bottom: third item probed). TCC d' decreases significantly with increased response delay and interference; This data clearly indicates that much more than set size affects d' in the model. For example, the difference between the 1st and 3rd item probed at set size 6 is greater than the difference between set size 3 and 6 overall.