Psychological Scaling Reveals a Single Parameter Framework For Visual Working Memory

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

Department of Psychology
University of California, San Diego
9500 Gilman Drive #0109
La Jolla, CA 92093

Please address correspondence to:

Mark Schurgin
Email: mschurgin@ucsd.edu

Abstract

Limits on the storage capacity of working memory have been investigated for decades, but the nature of those limits remains elusive. An important but largely overlooked consideration in this research concerns the relationship between the physical properties of stimuli used in visual working memory tasks and their psychological properties. Here, we show that the relationship between physical distance in stimulus space and the psychological confusability of items as measured in a perceptual task is non-linear. Taking into account this relationship leads to a parsimonious conceptualization of visual working memory, greatly simplifying the models needed to account for performance, allowing generalization to new stimulus spaces, and providing a mapping between tasks that have been thought to measure distinct qualities. In particular, performance across a variety of working memory tasks can be explained by a one-parameter model implemented within a signal detection framework. Moreover, despite the system-level distinctions between working and long-term memory, after taking into account psychological distance we find a strong affinity between the theoretical frameworks that guide both systems, as performance is accurately described using the same straightforward signal detection framework.

Keywords: visual working memory, recognition memory, psychological scaling, working memory capacity

Introduction

Working memory is typically conceptualized as a fixed capacity system, with a discrete number of items, each represented with a certain degree of precision (e.g., Cowan, 2001; Luck & Vogel, 2013). It is thought to be a core cognitive system (Baddeley, 2003; Ma, Husain & Bays 2014), with individual capacity differences strongly correlating with measures of broad cognitive function such as fluid intelligence and academic performance (Alloway & Alloway, 2010; Fukuda et al., 2010). As a result, many researchers are deeply interested in understanding and quantifying working memory capacity.

A task commonly used to probe the contents of visual working memory and to estimate its capacity is the continuous report procedure, where subjects are asked to report a remembered feature of a probed item (e.g., a color) by responding in a continuous space (e.g., a color wheel) (Wilken & Ma, 2004; Zhang & Luck, 2008). Responses are typically analyzed using a *mixture* model (Figure 1), the parameters of which represent the psychological status of the targets. In the simplest mixture model, the two parameters reflect distinct psychological states that correspond to (1) the proportion of trials in which subjects "guess" because the presented items are not available in working memory (represented by a uniform distribution), and (2) the precision of the representations associated with the items that are successfully represented in working memory (represented by a Gaussian-like distribution centered on the target color, with the standard deviation indicating precision). More elaborate mixture models allow for variability in the precision of target representations from item-to-item or trial-to-trial (Fougnie, Suchow & Alvarez, 2012; Ma, Husain & Bayes, 2014) and in some cases rather than assuming guessing, have some items represented with extremely low precision (van den Berg et al., 2012). Although the details of the models differ, they all assume that quantifying the relevant mixture distribution parameters provides an assessment of a fundamental limit on working memory capacity.

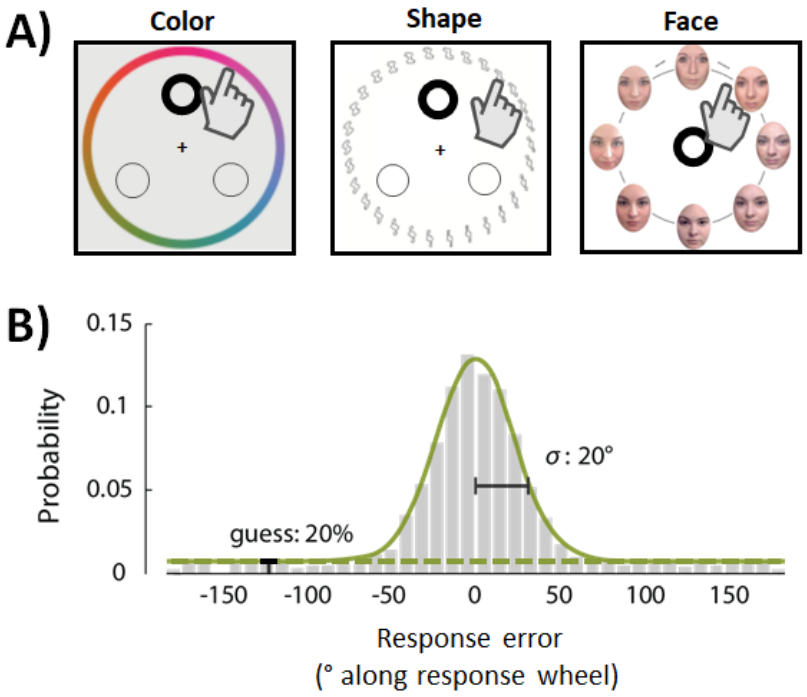

Figure 1. *(**A**). Examples of different continuous report tasks (portions adapted from Zhang & Luck, 2009; Zhou, Mondloch & Emrich, 2018). The participant sees a briefly presented stimulus display (e.g., 3 colors, 3 shapes, a single face, etc.) and after a delay is asked to use a wheel to report the exact color, shape or face identity of the single target item that occupied the cued location (indicated by the heavy black circle). (**B**) An example of a mixture model's fit to a histogram of continuous report data. Most responses fall within a relatively small range around the target item's true value, but some responses are far off. A standard model of performance in this task proposes two different psychological states underlie people's responses: target items that are encoded with a certain degree of precision (σ) and target items that were not encoded because they exceed the capacity of working memory (i.e. random guess responses distributed throughout the wheel). Some variations of this model assume a fixed capacity (i.e guess rate) but allow for variation in the precision of encoded targets, whereas others assume all targets are encoded, but guesses merely reflect items encoded with extremely low precision (i.e. an extremely large standard deviation).*

Variants of the mixture model have been extremely influential in shaping our current understanding of the nature of working memory. For example, experimental data quantified in terms of the parameters of the mixture model have been interpreted to mean that as the retention interval increases, representations in working memory do not decay gradually but instead "die" in an all-or-none fashion (Zhang & Luck, 2009), as 'guess rate' changes much more than 'precision' with delay. Other findings have been interpreted as showing that when working memory capacity is pressured, the precision of these representations plateaus near estimates of long-term memory precision, suggesting a shared constraint on performance (Brady et al., 2013). Mixture models have also been broadly applied to interpret the effect of many other variables on memory capacity, including aging (Peich, Husain & Bays, 2013), Parkinson's disease (Zokaei et al., 2014), sleep deprivation (Wee, Asplund & Chee, 2012), sleep disorders (Rolinski et al., 2015), video game training (Blacker, Curby, Klobusicky & Chein,

2014), transcranial magnetic stimulation (Rademaker, van de Ven, Tong & Sack, 2017), mental imagery (Keogh & Pearson, 2011), language and perception (Souza & Skora, 2017), auditory stimuli, (Kumar et al., 2013), speech (Joseph et al., 2015), visual motion (Zokaei et al., 2011), audio-visual events (Olivers, Awh & Van der Burg, 2016), medial temporal lobe damage (Pertzov et al., 2013), and the race of a face (Zhou, Mondloch & Emrich, 2018).

Here we show that the mathematical framework of fitting mixture models to characterize representations in working memory is based on a foundational assumption that is incorrect. Specifically, the models assume that degrees of error along the response wheel is a linear measure (the measure plotted on the *x*-axis of Figure 1B) that can be directly modeled to understand memory. However, we demonstrate that, consistent with nearly all cases of discriminability and generalization (Fechner, 1860; Shepard, 1987), the '*psychological distance*' that is relevant for understanding which items will be confused in memory is not linear with respect to physical distance. As a result, the assumptions guiding these models are untenable (i.e. the x-axis of Figure 1B is not directly interpretable). Taking into account the relationship between physical distance and psychological distance leads to a parsimonious single-parameter conceptualization of working memory, greatly simplifying the models needed to account for performance, allowing generalization to new stimulus spaces, and providing a mapping between tasks that have been thought to measure distinct qualities.

Results

We take memory for visual features, in particular color, as our primary case study because it has become a dominant tool for formalizing working memory capacity. When using color memory to investigate working memory capacity, researchers focus on the distribution of errors people make measured in degrees along the response wheel, *x*, where *x* ranges from 0° (for the color that matches the target) to 180° (for the most distant color from the target on the response wheel).

Existing models of working memory implicitly assume that because the perceptual space is locally perceptually uniform (i.e., any two nearby points on the color wheel are approximately equally discriminable; Brainard, 2003; Allred & Flombaum, 2014; Bae et al., 2015), the internal psychological distance between items relevant for memory performance is also linear as a function of (non-local) distance in this physical space. In reality, as with nearly all cases of discriminability (Fechner, 1860), perceptual differences (Maloney & Yang, 2003) and generalization (Shepard, 1987), we show that there is an approximately exponential fall off in psychological distance with physical distance in color space. Critically, the long-tails that are typically observed in working memory tasks when performance is plotted as a function of physical distance (Figure 1B) -- and which have been interpreted to reflect guessing associated with non-represented items -- do not exist when performance is more appropriately plotted as a function of psychological distance, *f(x)* (i.e., after correcting for the exponential fall-off with distance).

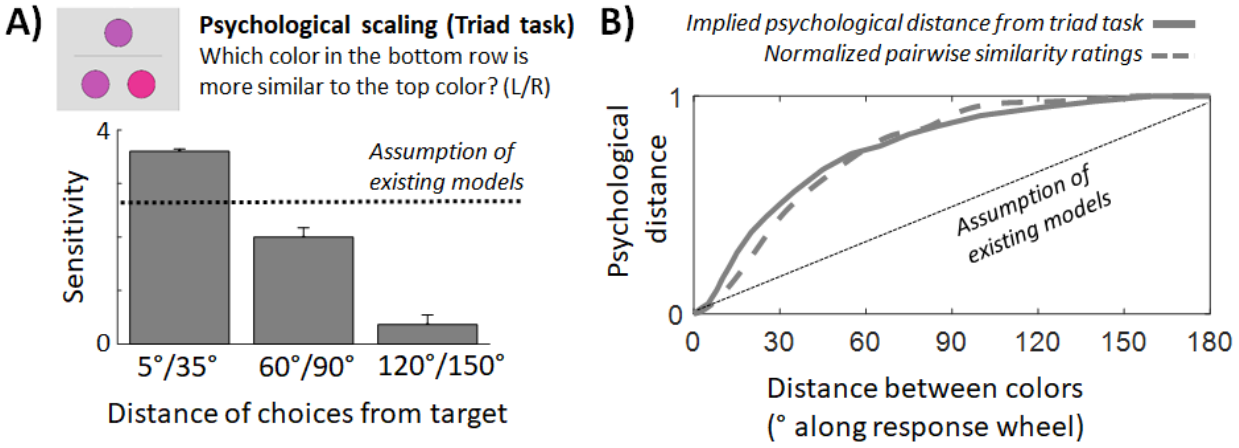

Figure 2. *(**A**). In our psychological scaling task, participants had to say which of two colors in the bottom row was more similar to the top (target) color. Despite the difference between the two choice colors always being 30° on the color wheel, sensitivity (d') dramatically decreased as the choices became more distant from the target.* ***(B)*** *We can use the data from a larger-scale triad task to infer the psychological distance of colors at different physical distances along the color wheel (solid curve). This function is highly non-linear, and thus these data contradict the linear scaling assumption of existing working memory models. Unlike two colors that are both close to a target color (e.g., 5° and 35°), colors that are both far from a target color (e.g., 120° and 150°) are approximately equal in psychological distance to the target, causing participants to pick both equally often in the triad task and when they serve as foils in a memory task. Rather than using the triad task and performing psychological scaling, we can also ask participants to rate the similarity of two colors on a Likert scale (1-7) and normalize their ratings. This similarity rating task provides another measure of psychological distance that is in close agreement with the results of the triad task (dashed line).*

Measuring psychological distance. To determine the relationship between physical distance along the color wheel, x, and psychological distance, $f(x)$, we tested how accurately participants could determine which of two colors was closest to a target color using a triad task (Torgerson, 1958; Maloney & Yang, 2003). This triad task is a perceptual task, but 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). In the triad task, even with a fixed 30° distance between the two colors that had to be compared to the target, participants were much more accurate on this task the closer the two colors were to the target (Figure 2A shows a subset of data; ANOVA $F(12,384) = 71.8$, $p<0.00001$). In other words, participants largely cannot tell, even in a perceptual task, 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. Using additional triad task data, we determined the implied psychological distance of colors at different physical distances along the color wheel using the psychological scaling technique of Maloney and Yang (2003). Psychological distance falls off in a nonlinear, exponential-like function (Figure 2B). The nonlinear relationship between physical distance, x, and psychological distance, $f(x)$, is consistent with classic psychophysical studies of discrimination and generalization in a wide variety of domains (Fechner, 1860; Shepard, 1987). A second, non-psychological factor is also relevant to this nonlinearity. In particular, physical distance in color space should be a function of the linear distance between two colors in 3D

color space (e.g., in CIELa*b*), not distance along the circumference of the response wheel (Figures S1-3 [Supplemental Material]).

A key implication of these scaling considerations is that the "long tail" of errors from continuous report data – often assumed to reflect distinct no- or low-information guessing – is a measurement artifact. What appears to be a long tail in physical space (e.g., Figure 1B) is a very short distance in psychological space. In particular, since participants are incapable, even 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. As an analogy, consider a person standing on a road aligned in both directions by telephone poles, with adjacent poles equally spaced in physical distance. As the person looks one way or the other, each telephone pole in the receding distance becomes less and less distinctive from its neighbors. Even though the local separation between the poles remains constant, the *perceived* (psychological) separation between them falls off with distance (Crowder, 1976). 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 2). Thus, colors far from the target on the color wheel are by necessity equally confusable with the target. The fact that participants are approximately equally likely to confuse the target with all of these far away colors, resulting in the "long tail" of errors, need not reflect a distinct guessing or low-information state, but instead is a natural consequence of psychological scaling.

Incorporating psychological distance into an equal-variance signal detection model. The results of our psychological scaling experiment provide the opportunity to investigate whether there is a simple relationship between psychological distance and memory that has been obscured by relying on physical distance (error in degrees) to model memory performance. Indeed, we find that an extremely straightforward signal detection model that treats the memory-match signal of each item as arising from its psychological distance to the target (Figure 3) accurately fits the key data from visual working memory studies.

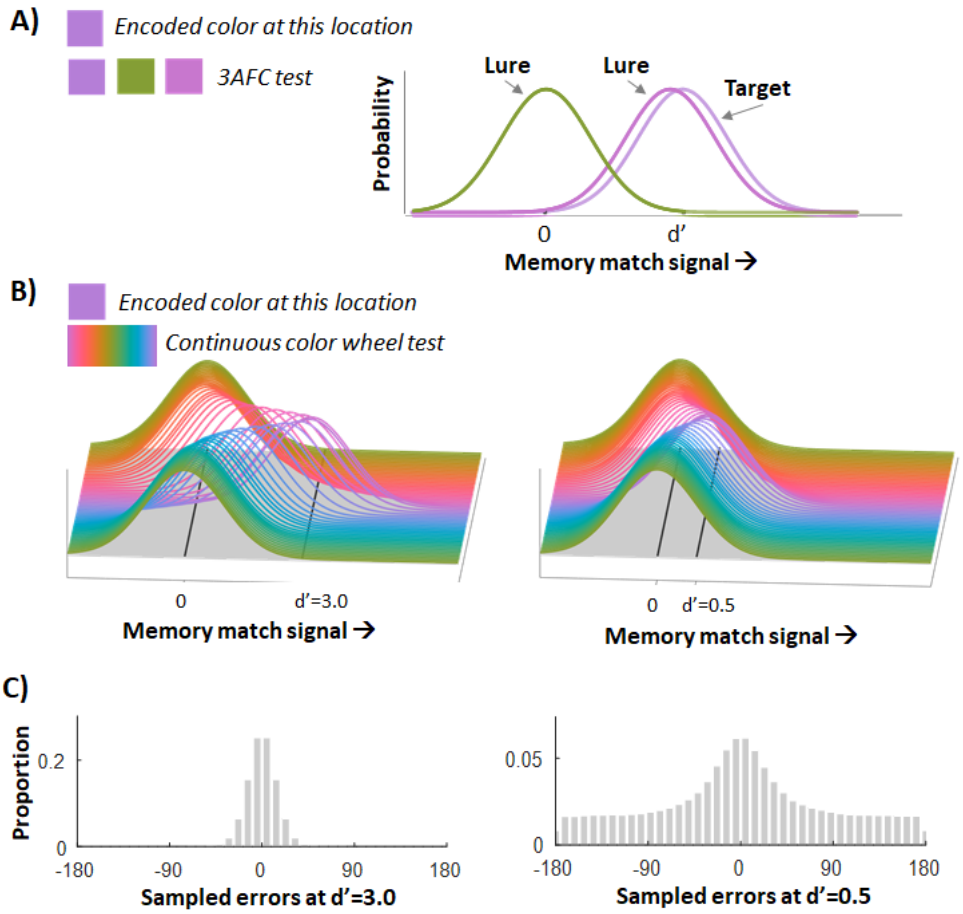

Figure 3. *(**A**) Signal detection model applied to memory in a 3-alternative forced-choice test. In standard long-term memory applications of signal detection, a "new" item sometimes feels familiar or less familiar, and thus elicits a memory match signal centered on 0 but with SD=1. By contrast, an "old" item (encoded previously) elicits a higher memory match signal, modeled as a normal distribution with SD=1 centered on d', and thus performance is best when d' is high. If asked to choose which is old between these choices, the simplest strategy is to sample from each distribution and choose whichever generates a higher memory signal sample. If we add a "similar" lure, here the second purple item, we expect people to choose it much more often than the unrelated lure (the green), and thus we conceive of it as having a memory match signal that is centered higher than 0 -- in other words, the lure gets some increased memory match signal from its shorter psychological distance to the target. (**B**) To extend this 3AFC case to the full continuous color wheel case, we think of every possible color probe (360 colors) eliciting a memory match signal, with the strength of each signal scaling based on its psychological distance to the target color (as measured in Figure 2B).On each trial, when faced with the continuous report display, participants draw samples from each of the 360 memory match signal distributions and report the color with the maximum memory match value. At high d' (left), this is very likely to lead to a color near the target. At low d' (right), it may lead to a far away color. However, because of the non-linearity in psychological distance, all items far from the target are roughly equally likely to be sampled, leading to the long-tails typically interpreted as evidence of low precision or non-represented items in working memory. (**C**). Example of generated data for d'=3 (left) and d'=0.5 (right), which closely resembles the data usually found in continuous report working memory tasks. Note that d' is the only parameter in this model; the means of the lure distribution are specified by the empirically-determined psychophysical function, f(x).*

This model is the same model used to understand how people perform 2-AFC in long-term memory, with two additional considerations: (1) other colors on the color wheel serve as additional lures (e.g., continuous report is like a 360-AFC task), and (2) the mean of the memory strength distributions for these lures corresponds to their measured psychological distance to the target, f(x). In particular, according to this model, explained in Figure 3, when participants are probed on the color of an item and shown a color wheel to choose their response, each color x on the wheel (0 degrees <= x <= 180 degrees) generates a memory signal, m_x, conceptualized as a random draw from that color's memory-match distribution, which is centered on d'_x (i.e. mean memory match signal strength for color x). That is, $m_x \sim N(d'_x, 1)$. Participants then respond by choosing the color corresponding to the maximum m_x. The mean memory-match signal for a given color x on the color wheel is given by that color's psychological distance to the target, i.e., $d'_x = d'(1\text{-}f(x))$, where d' is the model's only free parameter and $f(x)$ is the empirically determined psychological scaling function. When x = 0° (minimum distance), $f(x) = 0$ so $d'_0 = d'$. By contrast, when x = 180° (maximum distance), $f(x) = 1$, so $d'_{180} = 0$. Because of the nonlinear scaling, colors in the ~90° to 180° physical distance range (i.e., the long-tail range) do not cover a great expanse but instead all cluster near $f(x) \approx f(x)_{max}$ such that $d' \approx 0$. In other words, all of these colors are essentially equally distant from the target color (Figure 2). Importantly, this model makes a strong novel prediction: rather than the continuous report distributions reflecting 2 or 3 distinct parameters (guessing, precision and, in many models, variation in precision), they reflect a single d' parameter -- exactly as memory strength is conceived of in long-term memory models -- which is then combined with the *a priori* measured psychological distance function, $f(x)$.

Remarkably, this straightforward Target Confusability Model (TCM) 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 4), including both the relative height of the "long tails" of the distribution compared to the width of the central distribution, and aspects of the "peakiness" of the central distribution (thought to be the hallmark of variability in precision of working memory; see van den Berg et al., 2012). This is demonstrated by the 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).

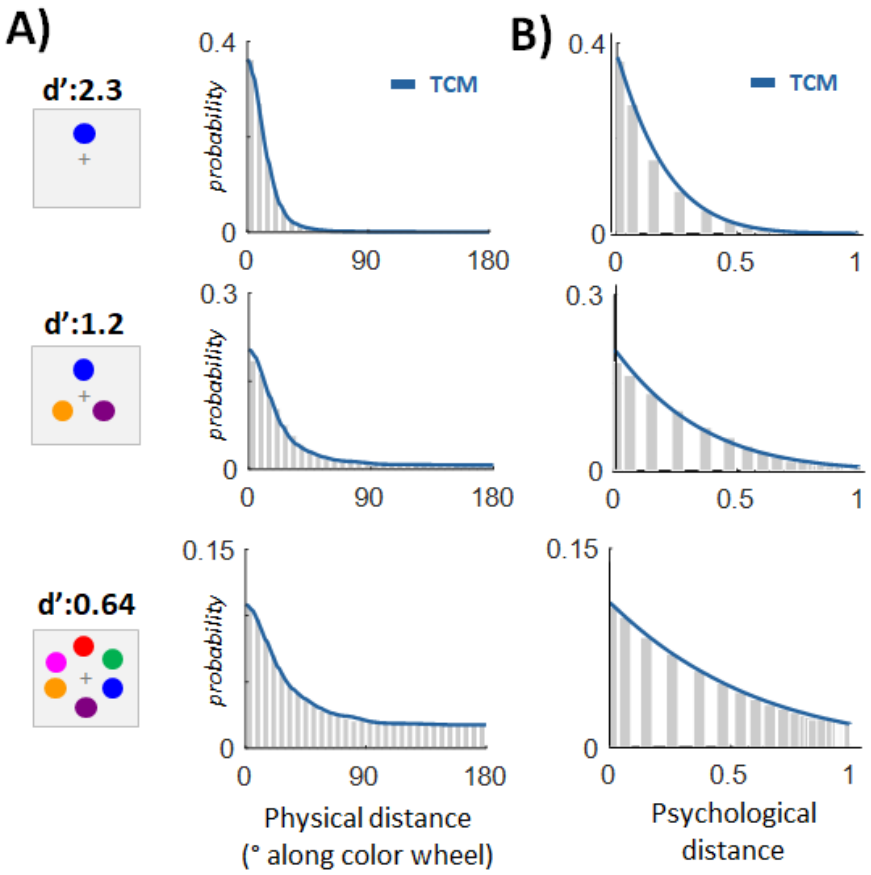

Figure 4. *TCM fits to data at set size 1, 3, and 6.* ***(A)*** *With only a single parameter, the Target Confusability Model accurately captures the response distribution typically attributed to multiple parameters / psychological states by existing models. Here we plot the absolute value of physical error.* ***(B)*** *A more accurate way to plot this data is relative not to the physical distance, where it appears to be a normal distribution on top of a long tail, but in psychological distance (e.g., with equal steps in psychological distance along the x-axis). In this space there is no discontinuity nor a long-tail, indicative of why the TCM is able to fit this data straightforwardly (note that plotting this way results in a clustering of the response distribution bars as psychological distance increases, since many of the options in the response wheel are at far psychological distances).*

Thus, taking into account psychological distance metrics permits a greatly simplified model of working memory performance - simply the standard signal detection model with added lures whose mean memory strengths are fixed based on psychological distance. This model also allows us to address other important theoretical questions about working memory. For example, an important debate in the working memory literature is centered on the question of whether all items are represented or whether participants have a fixed capacity limit of ~3-4 items after which they remember nothing about other items (e.g., Luck & Vogel, 2013). Our model holds that all items are approximately equally represented, even at set size 6, as we only need a single *d'* parameter for all items at this set size to fit the data. A hybrid model based on the TCM but mixed with 'guessing' (e.g., that assumes only a subset of items are represented and the remainder have *d'*=0) is both more complex and provides a worse fit when adjusted for

complexity (of 296 participants at set size 6, 93% have a lower AIC for the standard TCM than the one with the guess parameter). Across participants, we find a reliably better fit for the no-guess model ($t(295)=20.6$, $p<0.00001$). Thus, with only a single parameter, no concept of capacity, and no variability in precision, our model accurately and parsimoniously characterizes working memory response distributions across a range of set sizes and other factors (see Supplemental Material).

Model broadly applicable across working memory research. An advantage of taking into account psychological distance is that it provides a unifying theory of working memory across many tasks previously thought to measure distinct psychological constructs. To demonstrate the strength of this model and its claim, we next illustrate several ways in which, despite having only one parameter, the TCM generalizes to new tasks in ways that have not been previously considered possible.

One extremely important task for the literature on working memory is detecting large changes to a set of items across a delay (i.e., a change-detection task). For example, participants see 5 colors in a study display, and after a brief delay see a single probe color that may have changed or not changed from the study display (Figure 5A). This task has been widely used in the literature (e.g., Luck & Vogel, 1997), including in important work on clinical populations (Gold et al., 2003; Pisella, Berberovic & Mattingley, 2004; Olson, Moore, Stark & Chatterjee, 2006; Lee et al., 2010; Parra et al., 2010; Moriya & Sugiura, 2012). However, this task has recently been rejected as insufficient because, when working in physical space rather than psychological space, change detection appears not to fully characterize working memory performance (e.g., it does not measure 'precision', only 'guess rate'; Fougnie et al. 2010). Our psychological scaling findings, and the need for only a single d' parameter in the TCM, raise an alternative possibility that the change detection task yields the same d' parameter that characterizes continuous report data (once psychological distance is taken into account). If so, the TCM should naturally generalize from change detection tasks to continuous report tasks and vice-versa.

To test this prediction, we had participants perform a change detection task at both set size 2 and set size 5, with the probe colors being either an exact match to the item from that location or 180° away in physical color space (Figure 5A). We then obtained estimates of d' at set size 2 and 5 from memory performance data. Importantly, this simple measurement of d' in a change detection task allowed us to generalize and infer the distribution of responses from a continuous report task with the same stimuli. In particular, by scaling up from a standard 2 distribution signal detection model to a 360 distribution signal detection model (e.g., TCM) with the additional 358 foil color memory match signals centered at their psychological distance (a known and previously measured quantity, $f(x)$), we could predict what the entire histogram of data from a continuous report task using the same stimuli should look like (Figure 5B).

We found an excellent fit between this prediction and collected continuous report data at set size 2, $R^2=0.991$, $p<0.00001$, and at set size 5, $R^2=.995$, $p<0.00001$ (Figure 5C). The TCM provides an accurate fit to the data without any information other than how well participants can detect large (180-degree) changes and the measured psychological distance space $f(x)$. This

demonstrates that (1) the single parameter, *d'*, is sufficient to completely explain the distribution from continuous report studies, and (2) importantly for the field at large, shows that change detection, even with extremely large and easily discriminable changes, is sufficient to fully characterize working memory, thereby allowing the integration of a huge literature on this task that has been generally discarded as insufficient to assess working memory (e.g., Fougnie et al. 2010). Note also that, with an independent estimate of *d'* in hand, the TCM is parameter free.

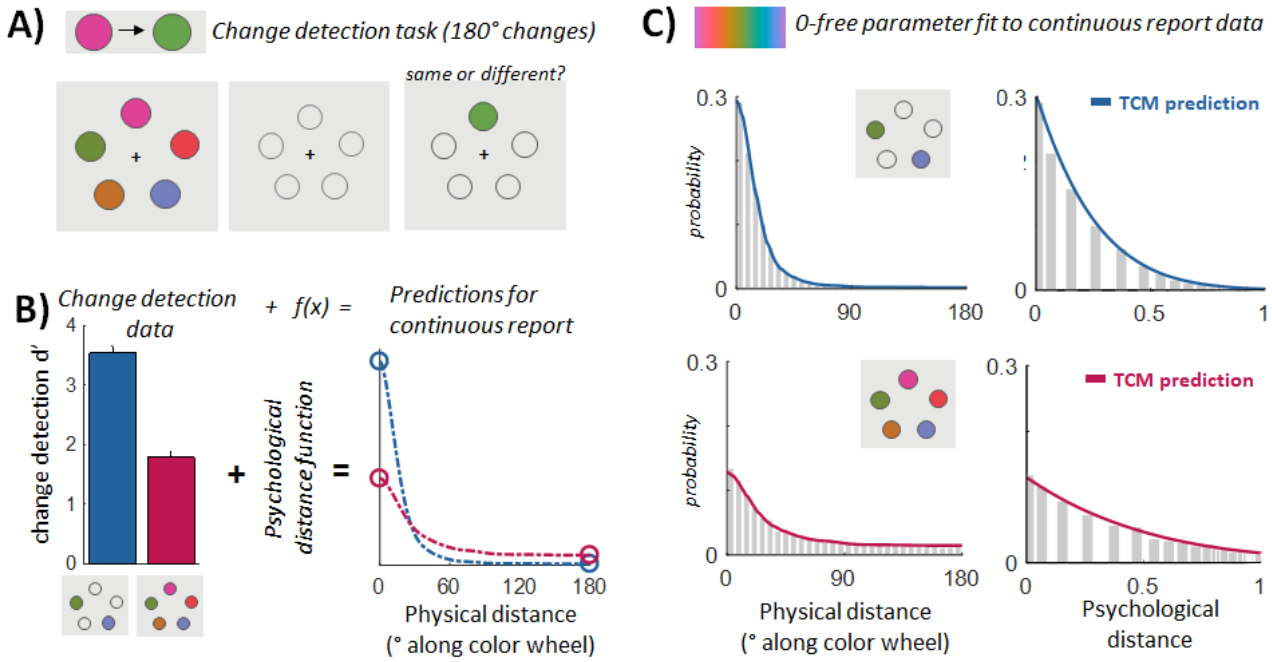

Figure 5. (***A***). *Illustration of change detection task. After a delay, participants were probed on a single item in the previous display, judging whether it was the same or different. If it was different, the foil color was always 180° away in color space, and thus according to existing models should only measure guess rate, not any of the precision parameters. Participants performed this task at set size 2 and set size 5.* (***B***). *Using the change detection data and similarity data, we generated a predicted continuous report distribution for both set size 2 (blue) and set size 5 (pink).* (***C***). *Fit of change detection predicted distribution to separately collected continuous report data from an independent set of individuals. Despite the assumption that change detection only reflects one parameter (capacity, i.e. guessing), taking into account psychological distance allows the TCM to accurately predict continuous report data, even though this data is typically assumed to reflect multiple parameters (i.e. guessing, precision, variability in precision).*

Generalization across stimulus variation. The TCM predicts that we should not only be able to generalize across different tasks previously thought to measure distinct constructs, but also predicts that we should be able to generalize to new stimulus spaces as well, as long as we take into account the psychological distance of the relevant stimulus space.

To assess this prediction, 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 participants to

report their memory on this smaller color wheel (see Figure S8 [Supplemental Material]). According to the TCM, the same model and *d′* parameters should apply without modification to this distinct task once we take into account the different psychological 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 TCM (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; see Figure S10 [Supplemental Material]). This generalization across stimulus space is not captured or predicted by previous theoretical frameworks and models guiding working memory.

Thus, in a working memory task using a distinct color space, after taking into account the new psychological distance structure of this color space, the TCM generalizes with no free parameters -- effectively showing that the same memory strength (*d′*) is evident in both tasks even though people make quite different distributions of errors in physical space (which would otherwise result in very different parameter values, and different psychological interpretations, under previous models).

The TCM harmoniously captures existing working memory data. The TCM makes a strong prediction that previously observed trade-offs between the parameters of traditional mixture modeling analyses actually reflect variation in only a single underlying parameter (*d′*). To evaluate this prediction, and to ask whether the TCM accurately characterizes the parameters that researchers have obtained by fitting the mixture model in other studies, we conducted an analysis of studies that reported precision and guessing parameters estimated using the dominant two-parameter mixture model developed by Zhang and Luck (2008). These studies vary considerably in set size, encoding time and other manipulations, and generally treat the two parameters obtained as estimates of distinct psychological concepts. We compared the trade-off between parameters values obtained by fitting the standard mixture model to empirical data to the trade-off in those same parameter values, but this time obtained by fitting the mixture model to simulated data obtained from the TCM over a range of *d′* values. As in our main analysis, we focused on studies that tested memory for color (15 papers, 56 data points; Table S1 [Supplemental Material]). We find that across a wide swath of papers, the two mixture model parameters indeed trade-off in the manner predicted by *d′* changes in simulated TCM data (Figure 6). This is also true when fitting data from individual papers that claim to find manipulations of a single parameter (e.g., Zhang & Luck, 2009; Brady et al. 2013), which, when mapped onto this plot, clearly show a decrease in *d′* is sufficient to account for their findings. This provides additional evidence that once psychological distance is taken into account, one parameter is sufficient to capture much of the data observed in continuous report tasks.

These results also serve to test the TCM's assumption that no items are poorly represented (i.e. low precision) or unrepresented (i.e. guesses). In particular, a capacity-limited model in which only a subset of items is represented predicts that the TCM should require multiple *d′* parameters to adequately characterize the data, especially at set sizes that are thought to exceed a fixed capacity of ~3-4 items. Using a set size of 6 as an example, imagine that a fit of the standard mixture model indicated that 4 items are represented and 2 are unrepresented

(guess rate = 2 / 6 = .333). To accurately fit data from a condition in which 2 items are not represented, the TCM should require a mixture of at least two *d′* values (with *d′* for some items set to 0 and *d′* for other items set to a value greater than zero) rather than a single *d′* value for all items of a given set size. When fit with just a single *d'* value, if two items are in fact unrepresented, the TCM should underestimate the height of the tails (and thus the 'guess' parameter). We found no evidence of this in our analysis of the literature (mean "guess rate" in literature reviewed with set size greater than four: 0.393, SEM: 0.037; mean predicted by the single-*d′* TCM from the corresponding SD parameters: 0.385). These results support the idea that all items are equally represented regardless of set size.

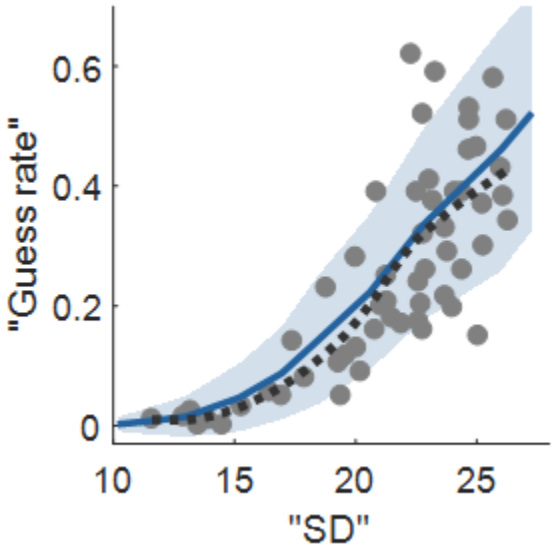

Figure 6*. Analysis of previous literature measuring the most commonly reported parameters of mixture models (precision [SD] and guessing). Gray dots are values reported in papers found in the literature; the dashed black line is a LOESS (local regression) smoothed version of these points. The solid blue line reflects the average "guess" and "SD" parameters when fitting data sampled from TCM with a mixture model, as a function of the d′ of TCM. 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 TCM with the mixture model.*

Discussion

For more than half a century, and especially over the last decade, efforts have been made to quantify the capacity of working memory. The major debate in the field is whether or not there is a finite limit on the number of items that working memory can hold (Miller, 1956; Luck & Vogel, 1997; Luck & Vogel, 2013; van den Berg, Awh & Ma, 2014), and in the last decade, this debate has been addressed largely by fitting mixture models to continuous report data. Although their details differ, all of the models agree that the items within a large working memory set are represented heterogeneously in terms of precision, including some items that are totally or almost totally unrepresented. However, our findings suggest these models are fundamentally incorrect. Once the nonlinearity of psychological distance with respect to physical distance is taken into account, neither guessing nor extremely low precision items are needed to account

for working memory performance. Instead, the Target Confusability Model (TCM) parsimoniously explains working memory data under the assumption that all items are represented with a single memory strength parameter (*d'*) that varies across different set-size conditions. Furthermore, taking into account the non-linear relationship between physical distance and psychological distance leads to a new conception of working memory performance, one that greatly simplifies the models needed to account for performance and that also generalizes between stimulus spaces and different tasks that were previously thought to measure distinct concepts.

At a broader level, the TCM provides a compelling connection to the literature on long-term recognition memory, which is often conceptualized within a signal detection framework. In this framework, there is no inherent capacity limitation, item memory strengths are assumed to vary continuously over a significant range, and mean memory strength is influenced by a variety of different variables (including similarity of the lures to the targets, or other factors known to affect working memory performance, such as encoding time or interference). This signal detection framework can also be naturally adapted to explain a number of findings that are in common between the working memory and long-term memory literature but have been difficult to explain with previous working memory models, like the relationship between confidence and accuracy (Rademaker et al. 2012; see also Wixted & Wells, 2017) and the ability of participants to respond correctly when given a second chance even if their first response was a 'guess' (Fougnie, Brady & Alvarez, in revision).

The fact that the TCM assumes no inherent upper limit on how many items are represented -- although *d'* does decrease with set size -- places it in agreement with some variable-precision models of working memory. However, in contrast to those models, TCM assumes that the items in a working memory set – like the items on a list in a study of long-term memory – can be represented as being drawn from single-precision distributions (that is, no items are assumed to be unrepresented or very weakly represented), which calls into question the basis of most inferences about a fundamental working memory capacity limit. Rather than a single specific constraint on the system generally attributed to set size, performance in working memory, like long-term memory, is likely limited by a variety of factors (such as lure similarity, encoding time, delay). It is important to emphasize that our claim is not that working memory and long-term memory are indistinguishable; instead, our claim is that the principles that govern working memory and long-term memory are much more similar than they are generally considered to be. Despite research on working and long-term memory operating largely independent of one another, we have provided a potential unified framework for investigating the distinctions and similarities in memory across both domains.

Methods

Fixed distance triad experiment. N=40 participants on Amazon Mechnical Turk judged which of two colors presented were more similar to a target color. Mechnical Turk users from a representative subset of adults in the United States (Berinsky, Huber, & Lenz, 2012; Buhrmester, Kwang, & Gosling, 2011), and data from Turk are known to closely match data from the lab on visual cognition tasks (Brady & Alvarez, 2011; Brady & Tenenbaum, 2013), including providing extremely reliable and high-agreement on color report data (Brady & Alvarez, 2015). 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 prevent the usage of words or other verbal information. Each participant did 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. 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.

Variable distance triad experiment. N=100 participants on MTurk 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 psychological 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. 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 psychological distance from this data, we used the model proposed by Maloney and Yang (2003), the Maximum Likelihood Difference Scaling method. This method assigns scaled values for each stimulus (in this case, each offset) designed to accurately predict observers' judgments, such that equally different stimuli in the scaled space are discriminated with equal performance.

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 commonly used response circle in the CIE *L*a*b** color space. This circle identical to that used in the 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.

Description of Target Confusability 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 psychological distance to the target according to the empirical psychophysical psychological distance function ($f(x)$), such that $d'_x = d'(1-f(x))$ (Figure 2).

Rather than using the outputs of the Maloney and Yang (2003) model for $f(x)$, the psychological distance function, we instead use the directly measured data from the similarity experiment. In that experiment, similarity between two colors separated by $x°$ was measured using a 7-point Likert scale, where $S_{min} = 1$ and $S_{max} = 7$ (Figure 2B). To generate $f(x)$, the psychological distance function, we simply normalize this data to range from 0 to 1, giving a psychological

similarity metric; and then subtract it from 1 to turn it into psychological 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 psychological 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. 2010 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.

Set size 1, 3 and 6 continuous report data. The continuous color report data used for fitting the model at set size 1, 3 and 6 was previously reported in Brady and Alvarez (2015). 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 (see Figure S11 [Supplemental Material]). The data used the same color wheel used in the current studies and was also conducted on MTurk to allow for direct comparison to the current similarity data.

Change detection experiment and associated set size 2 and 5 continuous report data. N=60 participants on MTurk performed a change detection task. This data was 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 this data to calculate a d′ 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 response 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, TCM 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 TCM. The reason is that the actual TCM d′ value is determined by the (unknown) number of colors sampled at test. To estimate the TCM d′ from the change detection d′, we followed these steps:

(1) Evaluate the probability density function of TCM 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, use the ratio of the probability density values for 0° and -180° to calculate a d′ value for a 2-AFC task that we would expect for this TCM 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 TCM d′ value (Macmillan & Creelman, 2004).

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

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

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

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

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 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 techniques by finding papers that cited the most prominent mixture modeling toolboxes, Suchow, Brady, Fougnie & Alvarez (2013) and Bays et al. (2009). We used a liberal inclusion criteria in order to obtain as many data points as possible. Our inclusion criteria was: 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 non-uniformities). For papers that did not report SD/guess values, these values were obtained by digitizing Figures with clear axis labels (Rohatgi, 2011; also see Mukherjee, Seok, Vieland & Das, 2013).

This 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 (56 data points) were included.

Acknowledgements

We thank Viola Stormer, Rosanne Rademaker and Daryl Fougnie for comments on these ideas and on the manuscript. For funding, we would like to acknowledge NSF CAREER (BCS-1653457) to TFB.

References

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.

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

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

Bae, G. Y., Olkkonen, M., Allred, S. R., & Flombaum, J. I. (2015). Why some colors appear more memorable than others: A model combining categories and particulars in color working memory. *Journal of Experimental Psychology: General*, *144*(4), 744.

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.

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.

Blacker, K. J., Curby, K. M., Klobusicky, E., & Chein, J. M. (2014). Effects of action video game training on visual working memory. *Journal of Experimental Psychology: Human Perception and Performance*, *40*(5), 1992.

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.

Brady, T. F., & Alvarez, G. A. (2015). No evidence for a fixed object limit in working memory: Spatial ensemble representations inflate estimates of working memory capacity for complex objects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(3), 921.

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.

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.

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.

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.

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

Crowder, M. J. (1976). Maximum likelihood estimation for dependent observations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 45-53.

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.

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

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.

Fechner, G. (1860). 1966. *Elements of psychophysics*.

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.

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.

Joseph, S., Iverson, P., Manohar, S., Fox, Z., Scott, S. K., & Husain, M. (2015). Precision of working memory for speech sounds. *The Quarterly Journal of Experimental Psychology*, *68*(10), 2022-2040.

Keogh, R., & Pearson, J. (2011). Mental imagery and visual working memory. *PloS One, 6*(12), e29221.

Kumar, S., Joseph, S., Pearson, B., Teki, S., Fox, Z. V., Griffiths, T. D., & Husain, M. (2013). Resource allocation and prioritization in auditory working memory. *Cognitive Neuroscience*, *4*(1), 12-20.

Lee, E. Y., Cowan, N., Vogel, E. K., Rolan, T., Valle-Inclan, F., & Hackley, S. A. (2010). Visual working memory deficits in patients with Parkinson's disease are due to both reduced storage capacity and impaired ability to filter out irrelevant information. *Brain, 133*(9), 2677-2689.

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

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.

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

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

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

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*(2), 81.

Moriya, J., & Sugiura, Y. (2012). High visual working memory capacity in trait social anxiety. *PloS One*, *7*(4), e34244.

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.

Olivers, C. N., Awh, E., & Van der Burg, E. (2016). The capacity to detect synchronous audiovisual events is severely limited: Evidence from mixture modeling. *Journal of Experimental Psychology: Human Perception and Performance*, *42*(12), 2115.

Olson, I. R., Moore, K. S., Stark, M., & Chatterjee, A. (2006). Visual working memory is impaired when the medial temporal lobe is damaged. *Journal of Cognitive Neuroscience*, *18*(7), 1087-1097.

Parra, M. A., Abrahams, S., Logie, R. H., Méndez, L. G., Lopera, F., & Della Sala, S. (2010). Visual short-term memory binding deficits in familial Alzheimer's disease. *Brain*, *133*(9), 2702-2713.

Peich, M. C., Husain, M., & Bays, P. M. (2013). Age-related decline of precision and binding in visual working memory. *Psychology and Aging*, *28*(3), 729.

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.

Pisella, L., Berberovic, N., & Mattingley, J. B. (2004). Impaired working memory for location but not for colour or shape in visual neglect: a comparison of parietal and non-parietal lesions. *Cortex*, *40*(2), 379-390.

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.

Rademaker, R. L., van de Ven, V. G., Tong, F., & Sack, A. T. (2017). The impact of early visual cortex transcranial magnetic stimulation on visual working memory precision and guess rate. *PloS one*, *12*(4), e0175230.

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

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.

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

Souza, A. S., & Skóra, Z. (2017). The interplay of language and visual perception in working memory. *Cognition*, *166*, 277-297.

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.

Torgerson, W. S. (1958). Theory and methods of scaling.

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

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.

Wee, N., Asplund, C. L., & Chee, M. W. (2013). Sleep deprivation accelerates delay-related loss of visual short-term memories without affecting precision. *Sleep*, *36*(6), 849-856.

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

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.

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

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

Zhou, X., Mondloch, C. J., & Emrich, S. M. (2018). Encoding differences affect the number and precision of own-race versus other-race faces stored in visual working memory. *Attention, Perception, & Psychophysics*, 1-11.

Zokaei, N., Gorgoraptis, N., Bahrami, B., Bays, P. M., & Husain, M. (2011). Precision of working memory for visual motion sequences and transparent motion surfaces. *Journal of Vision*, *11*(14), 2-2.

Zokaei, N., McNeill, A., Proukakis, C., Beavan, M., Jarman, P., Korlipara, P., ... & Husain, M. (2014). Visual short-term memory deficits associated with GBA mutation and Parkinson's disease. *Brain*, *137*(8), 2303-2311.