Journal of Memory and Language 93 (2017) 67–81

Contents lists available at ScienceDirect

Journal of Memory and Language

journal homepage: www.elsevier.com/locate/jml

Does working memory have a single capacity limit?

Robert Taylor *, Hana Thomson, David Sutton, Chris Donkin

School of Psychology, University of New South Wales, Australia

ARTICLE INFO

Article history:
Received 11 April 2016
revision received 13 September 2016

Keywords:
Stimulus complexity
Visual working memory
Computational models

ABSTRACT

Debate continues over whether visual working memory has a single, fixed capacity. Empirically, performance in working memory tasks worsens as the complexity of stimuli increases. However, there exist two explanations for this result. One proposal is that visual working memory is capable of holding fewer complex stimuli. The alternative proposal is that visual working memory can store 3–4 items, irrespective of their complexity. According to this fixed-capacity explanation, performance is worse for complex items because discrimination between complex items is more difficult than discrimination between simple items. These so-called comparison errors are more likely with complex items, and when left unaccounted for, lead to an underestimate of the capacity of working memory. Previous attempts at resolving this debate have relied on clever empirical manipulations of the similarity between stimuli. However, such approaches change the task that is given to the participant, and so may also change the way that participants use their memory. Here, we use a standard change detection task, but use a measurement model to estimate both the capacity of memory, and the probability of comparison errors. We apply the model to two change detection experiments in which we varied the complexity of the stimuli that participants must remember. Critically, we found that capacity estimates, and not comparison error estimates, varied depending upon stimulus complexity. Our results suggest that the number of items that can be stored is dependent on the complexity of the stimuli.

Introduction

There is continuing debate within the field of visual working memory (VWM) regarding the processes through which items are stored and the capacity limits imposed by these processes. The discrete slots view suggests that only a limited number of items can be retained in memory (Luck & Vogel, 1997), whereas the continuous resource view suggests that a limited pool of resources may be flexibly allocated across all items, with no necessary constraint on the number of items that can be stored (Ma, Husain, & Bays, 2014; Wilken & Ma, 2004).

Early evidence suggested that VWM capacity is limited to around 3–4 visual items, and that this capacity was invariant to changes in the number of stimulus features (Fukuda, Awh, & Volgel, 2010; Luck & Vogel, 1997; Vogel, Woodman, & Luck, 2001). However, there have since been multiple demonstrations of a detriment to performance in visual working memory tasks due to additional stimulus complexity (Cowan, Blume, & Saults, 2013; Fougnie, Asplund, & Marois, 2010; Hardman & Cowan, 2015; Oberauer & Eichenberger, 2013; Olson & Jiang, 2002). Though it is now largely agreed upon that performance in visual working memory tasks is worse for more complex stimuli, the cause for this result remains an open issue.

Alvarez and Cavanagh (2004) were one of the first to demonstrate that performance worsened for more complex stimuli. They showed that there was an inverse linear

* Corresponding author at: School of Psychology, University of New South Wales, Kensington, NSW 2052, Australia.
E-mail address: taylor.rt17@gmail.com (R. Taylor).

http://dx.doi.org/10.1016/j.jml.2016.09.004

relationship between visual complexity and working memory capacity, as operationalized by visual search time and the conventional change detection task, respectively. They found that capacity estimates were smaller for items with slower search rates, and concluded that fewer visually complex items can be held in visual working memory. Alvarez and Cavanagh (2004) concluded that the limit on the number of items that can be held in visual working was not fixed, and is jointly affected by the number of items that must be remembered and the complexity of the stimuli.

Advocates of the slots model highlighted a critical flaw in Alvarez and Cavanagh's (2004) claim that people stored fewer complex items than simple items. Alvarez and Cavanagh (2004) used the standard method for calculating capacity, which attributes all incorrect responses to a failure to store an item. As such, increased errors in a change detection task are associated with smaller capacity estimates; however, if another type of error is contributing to incorrect responses in the change detection task, then capacity will be underestimated. Awh, Barton, and Vogel (2007) argued that errors in change detection tasks can also arise through perceptual confusion between the contents of memory and the test item, and that such confusion is increasingly likely for more visually complex items. Fundamentally, Awh et al.'s argument is that participants are able to store approximately 3–4 complex items in memory, but a reduction in the mnemonic resolution for complex items increases the probability of making a comparison error, which ultimately confounds estimates of item capacity.

To support this claim, Awh et al. (2007) conducted an experiment in which participants were asked to remember stimulus sets that included two types of complex items – shaded cubes and Chinese characters. At test two types of changes could occur. A within-category change involved a stimulus being swapped for another member of the same category; e.g., one Chinese character changes to another. Conversely, a cross-category change involved a stimulus being swapped for a member of the other category; e.g., a Chinese character being replaced by a shaded cube. The critical finding was that participants struggled to detect within-category changes. On the other hand, cross-category detection was excellent; in fact, participants were equally able to detect cross-category changes as they were able to identify changes between simple colors. Awh et al. (2007) concluded that within-category changes involving complex stimuli increased the rate of comparison errors.

Scolari, Vogel, and Awh (2008), Barton, Ester, and Awh (2009) and Umemoto, Scolari, Vogel, and Awh (2010) have also shown that cross-category change detection of complex items is superior to within-category discrimination, and have further argued for a fixed number of discrete slots. In Barton et al. (2009) and Umemoto et al. (2010), the authors also attempted to estimate the error rate associated with comparing complex within-category stimuli. They reasoned that a participant's ability to detect cross-category changes provided a pure estimate of item capacity. Thus, any additional errors committed in the within-category change condition could be attributed to the comparison of the test item with the contents of memory. To address this issue they proposed a model that could measure both the probability that an item was stored and the probability that a comparison error was made, given the item had been stored in memory. However, despite the excellent motivation behind the model, Morey, Morey, Brisson, and Tremblay (2012) showed that this particular implementation was incorrect.

Brady and Alvarez (2015) also raise concerns regarding the validity of capacity estimates derived from cross-category changes. They instead suggest that participants form global representations of the study array which facilitate detection of change. For example, on some arrays the Chinese characters (or cubes) may be clustered more closely together, whereas on others the items may be located more diffusely. They found that when items of one category were clustered together, cross-category changes could be made without having to attend to the item characteristics, because the change was quite obvious. Conversely, as the items became more dispersed, cross-category change detection grew poorer. To further support their claim, they demonstrated that when the study arrays were more heterogeneous (consisting of more complex items than just cubes and Chinese characters), and so limit the ability to form global representations, capacity estimates dropped to approximate 1–2 complex items for cross-category changes, consistent with the findings from Alvarez and Cavanagh (2004). Brady and Alvarez (2015) conclude that performance using large changes overestimates capacity, which challenges the assumption of a fixed working memory capacity.

Given the foregoing, it is unclear as to whether item capacity remains invariant to item complexity. However, the two positions are clear and distinct. One proposal is that the number of items stored in memory is invariant to stimulus complexity, but that complex items will lead to comparison errors. The alternative proposal is that working memory can hold fewer complex items than simple items, irrespective of comparison errors. We set out to discern between these two very specific hypotheses.

Our approach differs from previous attempts to answer the question of a fixed capacity of visual working memory. We develop and apply a computational model that simply infers both comparison errors and capacity from performance in a standard change detection task. As such, we do not require within- and cross-category changes, and thus leave the task for the participant as simple as possible, free from any potential issues with the encoding of items from different categories (Brady & Alvarez, 2015). The model we use is inspired by that of Barton et al. (2009), but corrects the issues raised by Morey et al. (2012).

We apply the model to two experiments in which we vary the particular stimuli that participants must remember. By estimating both capacity and comparison errors, we provide a direct test of the two claims over the influence of complexity on visual working memory. According to Alvarez and Cavanagh (2004), we should expect that the effect of complexity influences the number of items that can be held in memory (i.e., smaller capacity). According to Awh et al. (2007), we should expect that the same number of items will always enter into memory, but more complex items will yield a greater number of errors once the items are in memory (i.e., more comparison errors).

Experiment 1

In Experiment 1 we used the single-probe change detection paradigm (Cowan et al., 2005). Participants are presented with an array of items to study and must decide whether a single test item is the same as one of the study items, or has changed. Participants completed four sessions, and in each session they were presented with only one of four different types of stimuli: letters, words, colors, or complex shapes.

Participants

Twenty participants were recruited via an on-line recruitment system at the University of New South Wales. Participants completed four experimental sessions, in the lab, each of which lasted approximately one hour. Participants had the option to complete multiple sessions per day, though a minimum break of one hour was required between sessions. All participants were paid $15 per session.

Stimuli

The experimental stimuli consisted of four sets of stimuli, as shown in Fig. 1. There was a list of 10 letters (B, D, G, H, K, L, M, P, R, and S), and a list of 10 words (BRIM, DUSK, GLUE, HERB, KNOT, LUMP, MAST, PORK, RAIL and SEAM). The words were approximately matched for frequency according to the Kucera-Francis written frequency norms (Kucera & Francis, 1967). The visual stimuli were composed of either 10 highly-discriminable colors (dark-blue-green, purple, cyan, orange, yellow, green, blue, red, white and black), or 10 novel complex shapes. Stimuli were presented using a 24 inch LCD monitor located approximately 60 cm from the participant. All stimuli were made to be of a similar size, having an approximate visual angle of .75 × .75 degrees. On a given trial, the set of study items were randomly positioned within a 8.8 × 7.3 degree array, subject to the constraint that there be at least a 2 degree separation between all items, and from any item to the center of the array.

Design

All participants completed a single session with each stimulus type. For each session participants completed 540 trials, which were broken into 9 blocks of 60 trials. Within each session both set size, $N = [2, 5, 8]$, and the proportion of trials on which the test item changed from the study item, $\pi = [0.3, 0.5, 0.7]$, were manipulated. Set size was manipulated across trials and probed equally often for same and change trials. Change proportion was manipulated across blocks, subject to the constraint that all change proportion conditions were completed before they were repeated. Within any given session the stimulus type remained unchanged, though the order participants received each stimulus type was randomized.

Procedure

Participants were instructed that they would undergo a simple memory task. At the start of each session, participants were introduced to the stimulus set to be shown during that hour. Prior to each block the participant was informed of the proportion of change trials in the current block (with a verbal description, and a pie graph). Trials began with a fixation cross presented in the center of the screen for 500 ms. Next, a study array consisting of N items was presented on the screen for 1000 ms. Following the study array a black screen was displayed for 500 ms, which was then followed by a 500 ms screen containing multicolored mask stimuli at the location of each of the study items. A single test item was then presented in one of the locations in which a study item was presented. On change trials, the test item was a stimulus that had not been presented during the study phase. A circular cue surrounded the test item. The participant was asked to respond as to whether the test item matched the study item presented in that location by pressing the "J" key, or whether the test item had changed by pressing the "F" key. Responses were self-paced. Once a response had been made a 1000 ms feedback screen indicated whether the response was correct or not, after which a 1000 ms blank screen was presented before the next trial began.

A measurement model for capacity and comparison errors

Model overview

The model that we apply was designed to reflect the two assumptions that are fundamental to the Awh et al. (2007) theory. First, we assume that the observer is capable of storing a subset, k, of the N items they are asked to remember. Second, if an item has been stored in memory, then there is some probability that the observer will make a comparison error, and thus give an incorrect

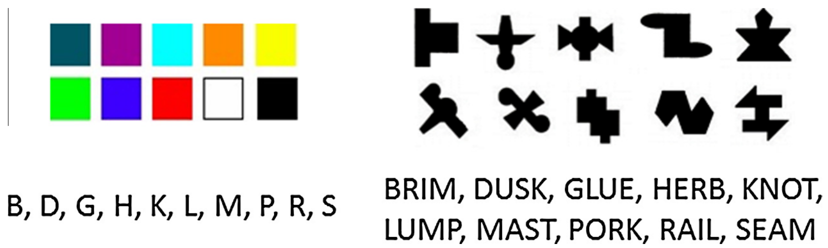

Fig. 1. The four sets of stimuli used in each experiment. The top left are the color stimuli, the top right are the complex shape stimuli, while the bottom left and right are letters and words, respectively. Note that the white color stimulus did not have a black border. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

response. In other words, the model does not assume that items stored in memory can be recalled with perfect precision.

In certain instantiations of the slots model there is an assumption that responses based upon memory will be perfect. This assumption has been referred to as the 'certainty assumption' (Rouder, Province, Swagman, & Thiele, submitted for publication). However, the model detailed here is based upon that outlined in Rouder et al. (submitted for publication) and relaxes this assumption. Specifically, we assume that responses based on memory may be incorrect due to a poor mapping between the test item and the contents of memory, which results in a comparison error. Like Barton et al. (2009), our model proposes two independent processes that influence an observers ability to detect change: the probability that an item has been stored in memory, and the probability of making a comparison error.

Model specification

We assume that study items are allocated to a limited number of slots. The average number of available slots, denoted by the capacity parameter, k, is assumed to remain fixed across changes in both set size, N, and change proportion, π. Note that the slots model assumes that capacity is a discrete quantity. However, conventionally capacity is estimated as if it were a continuous quantity, which approximates variability in capacity across trials. While our model does not explicitly assume that the number of slots varies across trials, we adopt the convention of estimating an average capacity value.

The probability that the probed item is successfully stored in memory, given the ith set size, is given by $d_i = \min\left(\frac{k}{N_i}, 1\right)$. The 'min' function simply says that if the number of study items is smaller than the available capacity, then all items are stored with probability 1. If an item is stored in memory, then there is some probability of a comparison error at test, though this probability depends on whether the trial is a change or same trial. If the test item changed, then a correct *change* response is made with probability a_j. If the test item is the same as the study item, then an incorrect *change* response is given with probability b_j. If the test item at the probed location was not stored in memory, the participant will be forced to guess. We estimate the probability that the participant guesses *change* in the jth change proportion condition to be g_j. That is, we assume that participants will adjust their guessing behavior in response to the base-rate of change trials (see Fig. 2).

The model outlined in Fig. 2 says that for the ith set size and the jth change proportion, the probability of making a correct *change* response on a change trial, θ^c_{ij}, or an incorrect *change* response on a same trial, θ^s_{ij}, is given by

$$\begin{aligned} \theta^c_{ij} &= d_i a_j + (1 - d_i) g_j \\ \theta^s_{ij} &= d_i b_j + (1 - d_i) g_j \end{aligned} \tag{1}$$

Implementation of comparison errors

Note that both a_j and b_j are conditional upon the item having successfully made it into memory. They do not affect the probability that an item *is* stored. In essence, they are conditional hit and false alarms rates, respectively, and determine the probability that a correct response is made given that an item is in memory. Note also that the probability of a comparison error, given that an item is in

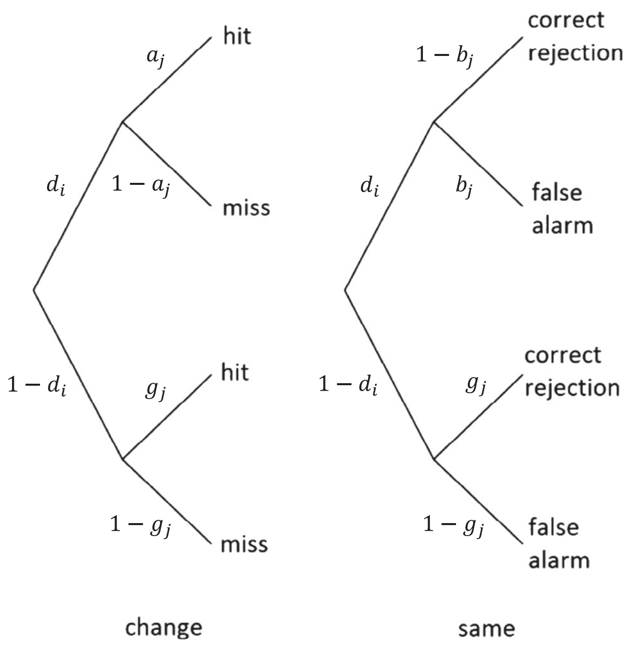

Fig. 2. Description of the comparison error slots model for the change detection task.

memory, depends on the change proportion condition. The probability of a comparison error depends on change proportion because a_j and b_j represent errors in change and same trials, respectively, and are simply a proportion of change responses. So, if the proportion of change trials is large, then both a_j and b_j will be larger than if the proportion of change trials is smaller. This assumption allows participants to require different levels of evidence of a match between test items and the contents of memory depending on the probability of a change trial. Note also that because both a_j and b_j can vary across change proportion, it is possible to produce curvilinear ROC functions. This behavior is typically not possible in the discrete-slots framework.

We also assume that the probability of a comparison error is constant across all set size conditions. In other words, we assume that an item in memory is stored with the same resolution, regardless of how many items were presented to the participant. Such an assumptions follows directly from the simplest versions of slot-based models of visual working memory. However, the slots plus averaging model in Zhang and Luck (2008) proposes that items can be stored with higher resolution when the capacity of memory is larger than N. The idea is that items may be stored in multiple slots, and these multiple representations can be aggregated to yield better memory. As such, we also fitted a version of our model that estimated one set of a and b parameters for the $N = 2$ conditions, and another for the $N = 5$ and $N = 8$ conditions. The results for this alternative model were extremely similar to those presented here, and certainly yield the same conclusions. We have placed the details and results for this model in the Supplementary Materials.

Model inference

We estimate the parameters of the model separately for each of the four different stimulus types. So, for the mth participant, completing the task with the sth stimulus type, the relevant parameters are: k_{sm}, a_{jsm}, b_{jsm}, and g_{jsm}. Each individual produces a number of correct *change* responses, or hits h_{ijsm}, and a number of incorrect *change* responses, or false alarms f_{ijsm}. We assume that a Binomial process generates these observed hits and false alarms, such that

$$\begin{aligned} h_{ijsm} &\sim Bin(\theta^c_{ijsm}, n^c_{ijsm}) \\ f_{ijsm} &\sim Bin(\theta^s_{ijsm}, n^s_{ijsm}) \end{aligned} \tag{2}$$

where n^c_{ijsm} and n^s_{ijsm} are the number of change and same trials completed by an individual in each condition.

We use hierarchical Bayesian estimation to fit the model to data. The hierarchical model assumes that the parameter values of individual participants are drawn from a population-level Normal distribution. For example, we assume that the mth individual's capacity for the sth stimulus type comes from a Normal distribution with mean K_s and standard deviation σ_K, such that $k_{sm} \sim N(K_s, \sigma_K)$. For the remaining parameters, we chose to set the standard deviation of the population-level parameters to be σ, in order to constrain the model. For example, we assumed that the mth individual's guessing rate for the sth stimulus type in the jth change proportion condition was $g_{jsm} \sim N(G_{js}, \sigma)$.[1] Since our focus was on parameter estimation we used vague priors for our model parameters. We placed a Beta distribution $Be(1, 1)$ prior on the population-level means for rates (A_{js}, B_{js}, and G_{js}) and a Uniform $U(0, 8)$ prior on the population-level mean capacity (K_s). Finally, we placed a Uniform $U(.01, 10)$ on both σ_K and σ. Finally, all posterior distributions were obtained using JAGS by running 6 chains of 5000 samples (after 500 burn-in samples).

Model validation

Because our conclusions depend entirely upon the inferences drawn from our model, we ran a number of validation analyses. The details of these are reported in full in the Supplementary Materials section. Briefly, we first report a simulation study that demonstrates that the design of our experiment allows us to estimate the parameters of our model with sufficient precision to discriminate between the two focal hypotheses. That is, our design allows us to determine whether changes in performance for more complicated stimuli are due to increased comparison errors or reduced capacity.

Second, we show that the comparison error parameter in the model captures what should be an increased rate of comparison errors in actual data. We report the results of an experiment in which the change between study and test stimuli, when it occurred, was either small or large. Critically, since all of the stimuli were colored squares, there should be no difference in capacity estimates for small- and large-change trials. Rather, we should expect the number of comparison errors to increase when the change between study and test items is smaller. We fit our model to the small- and large-change trials separately, and found large changes in the comparison error parameter, while the capacity parameter (and all other parameters) remain relatively constant between the two conditions. That is, we find that our model attributes the difference between small- and large-change conditions to the comparison error parameter, and not to the capacity parameter.

Model predictions

We focus our inference on the means of the population-level distributions, K_s, A_{js}, B_{js}, and G_{js}. The values of K_s will tell us how many items of each stimulus type are stored in memory. A_{js} and B_{js} will tell us about the rate of comparison errors that occur for each of the different types of stimuli. The way that the K, A and B parameters change as a function of stimulus complexity will allow us to distinguish between our two hypotheses. According to Alvarez and Cavanagh (2004), we expect that K will be smaller for more complex items, while Awh et al. (2007) would

[1] For rate parameters we truncated the normal distributions between 0 and 1. Another approach is to model individual differences as normally distributed latent variables that are transformed into probabilities through a logistic function. We considered this latter approach also. The fits to the data were virtually identical. We subsequently report the fits based upon the original truncated normals. Plots and code for both model implementations are available on the OSF.

predict that K will be constant for all stimuli, but that A and $1 - B$ will be smaller for more complex stimuli.

Our primary analysis will be to estimate parameters from the data. However, to facilitate a more qualitative interpretation of our data, we plot the predicted hit and false alarm rates for a range of parameter settings of our model in Fig. 3. In the top panels of the figure, we hold the capacity parameter constant and change the probability of a comparison errors. If we imagine that moving from the top left to the top right panel is equivalent to increasing the complexity of a stimulus, then we see the pattern of data we expect based on Awh et al. (2007). The bottom panels of the figure demonstrates the effect of reducing the capacity of working memory. The difference between the top left and the bottom left panels shows the effect of increasing stimulus complexity according to Alvarez and Cavanagh (2004). The key difference between the two accounts is that when the capacity for complex stimuli is reduced, we expect to see a greater difference between the small set size and the larger set sizes. That is, we expect performance to remain good when there are only 2 stimuli, but participants should get much worse for the $N = 5$ and $N = 8$ conditions.

Results

Turning first to the model parameter estimates, Fig. 4 plots the posterior distributions for the mean of the population-level parameter distributions in our model. The posteriors are plotted as violin plots, which are composed of a central box-plot that is surrounded by a smoothed density. These plots represent the relative likelihood of the values that each parameter could take to account for the observed data. There are two critical results to take away from these posterior distributions. First, we find that the capacity estimates, K, for letters, words, and colors are consistent with the standard upper limit for short term memory of 3–4 items. However, the capacity for complex shapes is markedly lower, at around 2 items. This finding is consistent with previous arguments that there is a smaller capacity for more complex items (Alvarez & Cavanagh, 2004; Brady & Alvarez, 2015).

Second, we observe negligible changes in the comparison error parameters, A and B, across the different stimulus types. This result implies that items stored in memory, regardless of their complexity, yield the same probability of a correct or incorrect change response. More specifically,

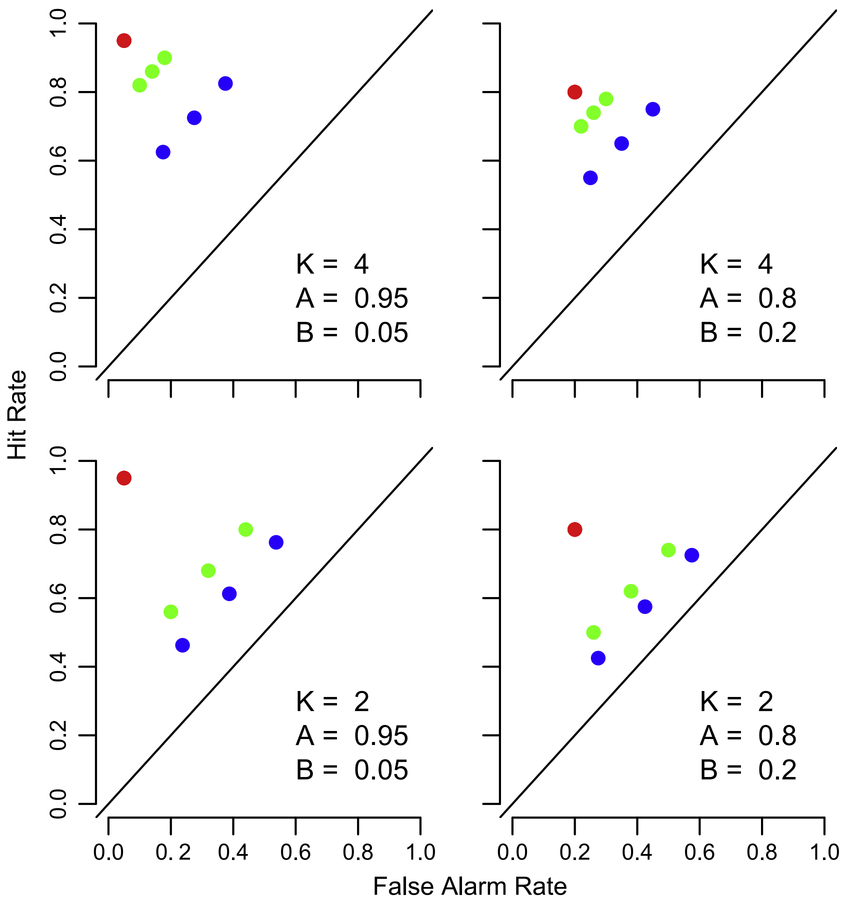

Fig. 3. Predicted hit and false alarm rates for four different parameter settings of our model. The parameters in the top left panel reflect those that one would expect for simple color stimuli. The top right panel shows the predictions for more complex stimuli according to Awh et al. (2007). The bottom figures reflect the predictions for complex stimuli that are consistent with Alvarez and Cavanagh's (2004) account. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

we see that changes to items that make it into memory are identified with near perfect accuracy.

Finally, though not of particular theoretical relevance, we see a typical pattern of guessing behavior. Participants appear to do something akin to probability matching, where they increase the rate at which they guess change as the proportion of change trials increases (as opposed to the more rational approach of maximizing, such as always guessing change when the probability of a change trial is greater than 50%). This pattern is observed for all stimulus types, though there seems to an inflated rate of responding *change* to color stimuli.

To assess the fit of the model to the data we generated posterior predictive distributions. We do this by simulating hit and false alarm rates, defined by Eq. (1), using the values contained within the population-level posterior distributions. So, we generate distributions of hits and false alarms based upon our model's most likely parameter values. We then compare the posterior model predictions to the observed data. We estimated the observed hit and false alarm rates by fitting a hierarchical data model that simply estimates a binomial rate across all conditions. This ensures that the observed rates have also undergone the shrinkage imposed by hierarchical models, as did our computational model. As before, the model assumes that individual-participant hit and false alarm rates are drawn from population-level normal distributions.

Fig. 5 displays the empirical data (in white) and the model fits (in color) for each stimulus condition. Looking first at the data, we see that the shape stimuli show a dif-

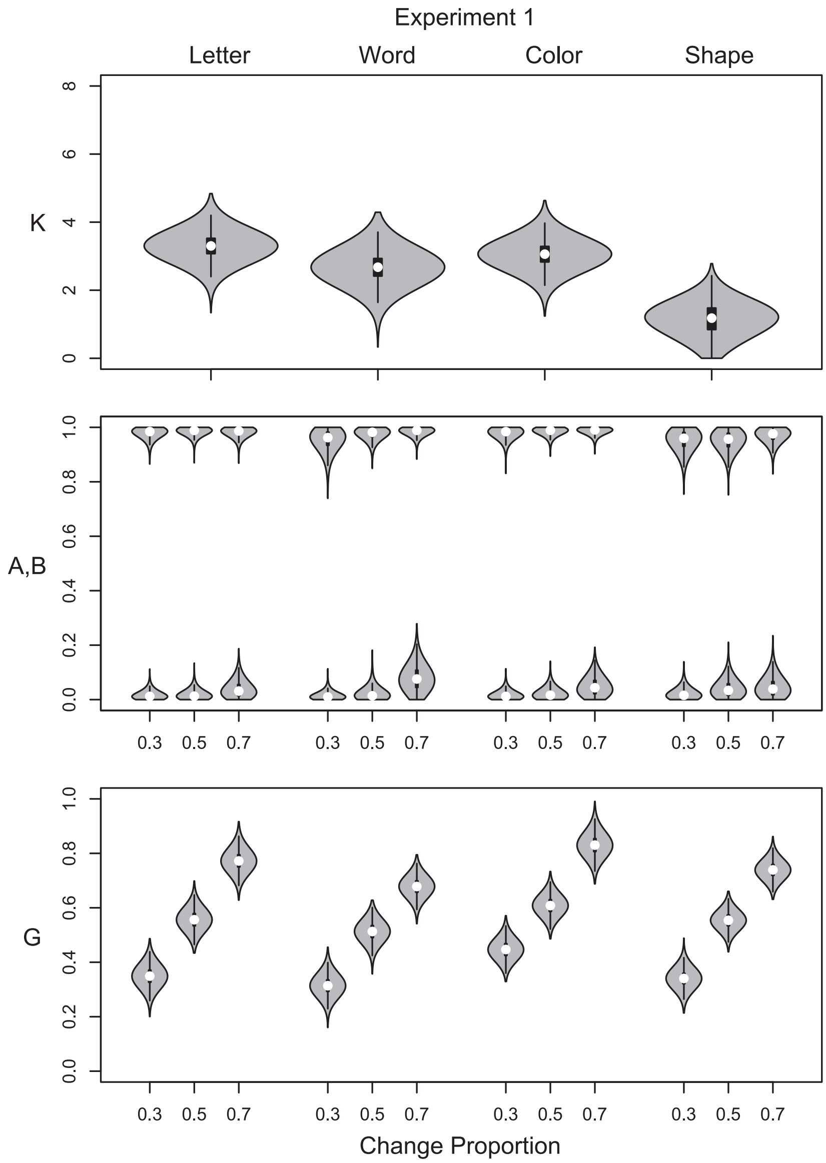

Fig. 4. Posterior distributions for the population-level parameter means. The top panel displays the capacity means across each stimulus type; the middle panel displays the mean comparison error probability across both stimulus type and change proportion for change, *A*, and same, *B*, trials; the bottom panel displays the guessing rates across each level of change proportion for each stimulus type.

ferent qualitative pattern to the other stimuli. Compared with the letter, word, and color stimuli, the shapes show a smaller difference between the $N = 5$ and $N = 8$ conditions than between the $N = 2$ and $N = 5$ conditions. Recall that this is the hallmark pattern of a reduction in capacity, based on Fig. 3.

The posterior predictive distributions in Fig. 5 suggest that the model provides a reasonable fit to the data. The model is able to capture the general pattern in the observed data, in terms of the effect of set size, change proportion, and stimulus type. Though the model certainly does not capture all of the patterns in the data, recall that we aim to use this model as a measurement tool. As such, we are more tolerant of some of the misfits, as we do not think that these speak directly to the measurement of comparison errors and capacity. For example, the model fails to capture the false alarm rates when $\pi = 0.3$ for color and shape stimuli. However, to foreshadow, we do not see such patterns in Experiment 2, suggesting that such misfits may not be a large problem.

Discussion

Our results indicate that participants stored fewer of the complex shape items than they did letters, words, or colors. However, interestingly, we find that participants made almost no comparison errors with any of the stimuli that were contained in memory. These results differ from previous suggestions that comparison errors provide the key limit upon capacity estimates for complex stimuli (Awh et al., 2007; Barton et al., 2009). That is, our results are more in line with the conclusions of Alvarez and Cavanagh (2004). What's more, we are able to draw these conclusions while having allowed for the possibility of comparison errors.

We now demonstrate the same empirical effect in a second change detection experiment. In the previous experiment we had participants complete a change detection task with stimuli that varied in complexity. In the change detection paradigm, participants are presented with all stimuli simultaneously, and so are allowed to choose how to allocate items into memory. However, participants only have a limited amount of time during which item encoding can be preformed. For example, one could reason that there was not enough time to encode the complex stimuli, and so the number of items that could be stored was limited by encoding limitations.

In Experiment 2 we sought to replicate our previous findings, though allow for more time to view the study array when there are more items in the array. Specifically, we allowed 250 ms encoding time per item. Accordingly, when participants were shown an array of five items they had a total of 1250 ms (5×250 ms) encoding time. Additionally, we also incorporate the knowledge we have gained from Experiment 1 to update our computational model. When fitting the model to the data from Experiment 1 we had very little substantive knowledge about what values the model parameters might take. As a result we sought only to estimate the model parameters and so placed uninformative priors on the population-level parameter means.

The resulting posterior distribution reflects the reallocation of credibility to the most likely parameter values. That is, it encapsulates what has been learned from the observed data and how we should revise our beliefs about the value of a specific model parameter. Moreover, this new information can be used to update the model specification. Specifically, we can use the posterior distributions estimated in Experiment 1 as priors on the population-level model parameters in Experiment 2 (Kary, Taylor, & Donkin, 2016). Using informed priors also allows us to perform Bayesian inferential tests on differences in the model parameters across conditions.

Experiment 2

Method

Twenty new participants were recruited into Experiment 2. All methodological details regarding participants, design, stimuli, and modeling were the same as in Experiment 1. The only procedural difference was the length of time the study items were presented to the participant. We allowed 250 ms encoding time per item within the study array. The total presentation time was thus equal to the number of study items in the array multiplied by 250 ms. Accordingly, for the $N = [2, 5, 8]$ set size conditions, total presentation time was equal to 500 ms, 1250 ms, and 2000 ms, respectively.

Model updating

Our model is hierarchical, and so participant-level parameters are draws from population-level normal distributions. Each population-level distribution has a mean and a standard deviation parameter; for example, the population-level capacity parameter for a given stimulus type has a mean of K_s and a standard deviation of σ_K. Recall that in Experiment 1, we placed uninformative priors on the mean and standard deviations of these normal distributions. For example, we placed a $Uniform(0, 8)$ prior on the mean of the population-level capacity parameter, K_s. In Experiment 2, we will instead use informative priors, based upon the posterior distributions from Experiment 1.[2]

We need an informative prior for the mean and standard deviation of each population-level normal distribution. First, for the means, we defined our informative priors by fitting Normal and Beta distributions to the population-level mean posterior distributions shown in Fig. 4. We assumed that the population-level mean capacity parameter, K_s would be normally distributed. So, for each stimulus type, we fitted normal distributions to the posterior distributions shown in the top row of Fig. 4, and used the maximum-likelihood estimates as our prior. For example, for Letter stimuli, the best-fitting normal distribution to the posterior distribution plotted in the far left

[2] It is worth noting that the pattern of parameter estimates, and the conclusions from this experiment, are practically identical when vague priors are used to analyze the data from Experiment 2. All of the materials necessary to carry out this analysis are available on the OSF.

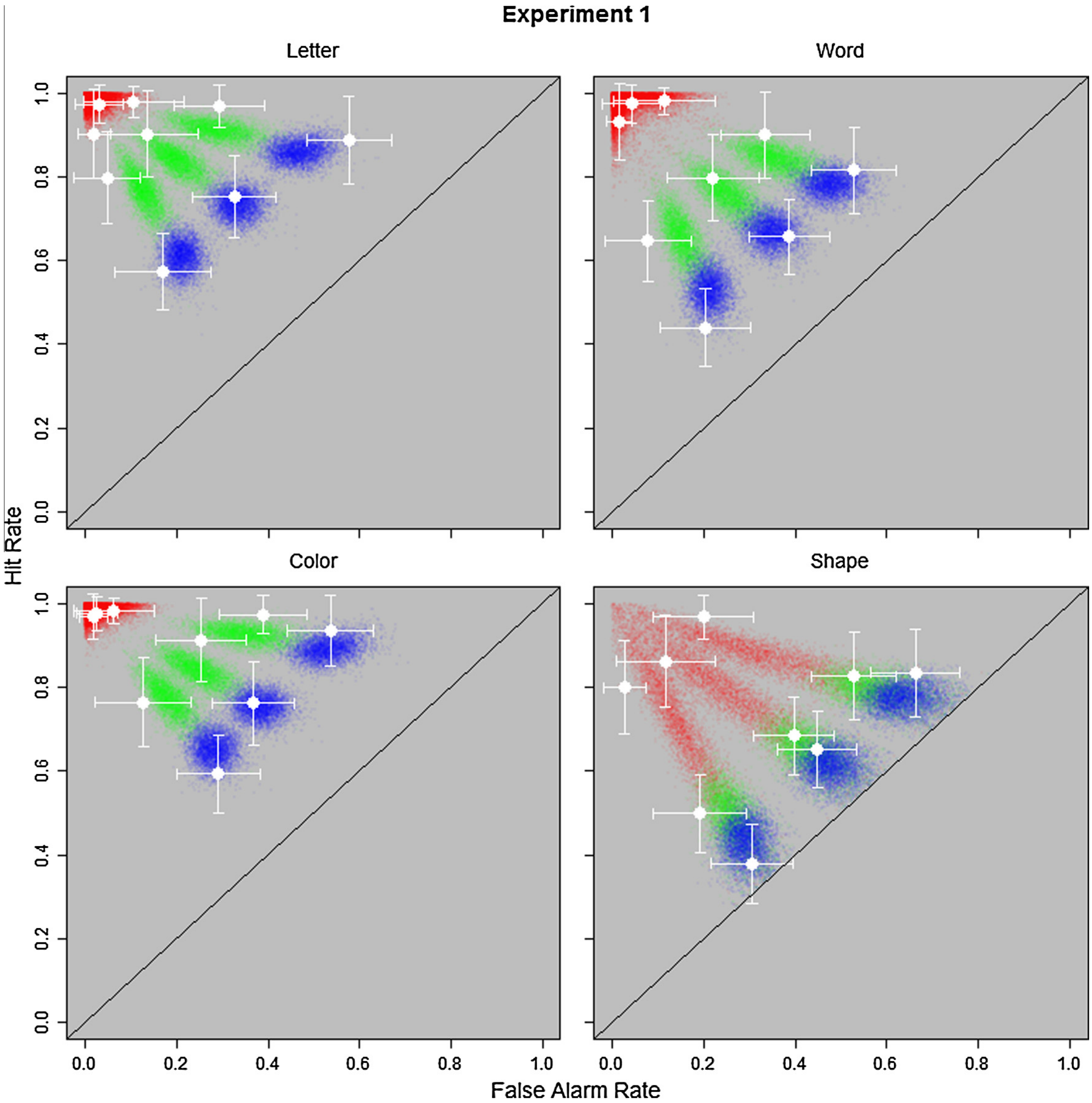

Fig. 5. Posterior predictive distribution for the comparison error model across each stimulus condition in Experiment 1. Each color band corresponds to a different set size condition: red = set size 2; green = set size 5; and blue = set size 8. Within each band there are three distinct clusters, each of which correspond to the different change proportion conditions. The white data points are the mean hit and false alarms rates for the observed data, with 2 standard deviations indicated. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

of the top row of Fig. 4 had a mean of 3.3 and a standard deviation of 0.35, and so our prior on K_L was $N(3.3, 0.35)$. The prior distributions for all population-level mean capacity parameters are given in the top row of Table 1.

The same basic approach applied when estimating the distributions placed on the population-level rate parameters (i.e., A_j, B_j, and G_j), only here the parameters must lie in the interval $[0, 1]$. We thus used a Beta distribution for each of these parameters. So, for the jth change proportion condition we assumed that A_j, B_j, and G_j were distributed as a $Be(\alpha, \beta)$, where α and β were estimated by fitting the posterior distributions from Experiment 1. The full set of distributions used to inform the population-level mean parameters are contained in Table 1.

We also used informative priors for the population-level standard deviation parameters. We used normal distributions to characterize the posterior distributions for the variance parameters. We required only two such priors: one for the standard deviation on the rate parameters, σ, and one for the standard deviation of the capacity parameter σ_K. For the variance of the rate parameters, the updated prior was $\sigma \sim N(.11, .01)$, whereas the variance for the capacity parameter was $\sigma_K \sim N(1.43, .18)$.

Bayesian hypothesis tests

We applied the Savage-Dickey density ratio test to assess the differences between the capacity posterior distribution, denoted as δ, for each stimulus type. The Savage-Dickey test

Table 1
Prior distributions placed on K_s for Experiment 2.

Parameter	Stimulus type			
	Letter	Word	Colour	Shape
K	$N(3.30, .35)$	$N(2.67, .40)$	$N(3.05, .35)$	$N(1.15, .46)$
$A_{.3}$	$Be(51.55, 1.08)$	$Be(25.41, 1.23)$	$Be(51.55, 1.10)$	$Be(26.17, 1.30)$
$A_{.5}$	$Be(69.64, 1.08)$	$Be(46.53, 1.13)$	$Be(71.16, 1.06)$	$Be(28.09, 1.45)$
$A_{.7}$	$Be(59.33, 1.08)$	$Be(67.30, 1.05)$	$Be(82.11, 1.06)$	$Be(38.00, 1.16)$
$B_{.3}$	$Be(1.10, 67.59)$	$Be(1.05, 77.23)$	$Be(1.07, 66.48)$	$Be(1.11, 52.03)$
$B_{.5}$	$Be(1.08, 60.85)$	$Be(0.99, 48.10)$	$Be(1.12, 50.84)$	$Be(1.29, 31.13)$
$B_{.7}$	$Be(1.25, 31.83)$	$Be(2.13, 24.77)$	$Be(1.47, 28.99)$	$Be(1.29, 27.17)$
$G_{.3}$	$Be(71.05, 132.52)$	$Be(66.63, 145.42)$	$Be(107.34, 132.89)$	$Be(92.76, 179.66)$
$G_{.5}$	$Be(120.01, 95.74)$	$Be(117.79, 112.19)$	$Be(140.81, 90.64)$	$Be(160.33, 129.14)$
$G_{.7}$	$Be(112.97, 33.19)$	$Be(147.54, 70.13)$	$Be(82.74, 16.82)$	$Be(161.35, 56.93)$

Note: $N(\mu, \sigma)$ denotes Normal distribution and $Be(\alpha, \beta)$ denotes Beta distribution.

uses the prior and posterior distribution on δ to test whether the difference is reliably different from zero. Accordingly, the test corresponds to a conventional two-tailed hypothesis test insofar as the null hypothesis assumes that there is no difference between the capacity posterior distributions, $H_0; \delta = 0$, and the alternative hypothesis assumes that there is a difference, $H_1; \delta \neq 0$. However, unlike conventional hypothesis tests we can obtain evidence for both the null and alternative hypothesis. This permits an evaluation of the relative odds that the data were generated by the null hypothesis relative to the alternative.

The prior placed on the population-level difference between stimulus capacity, $p(\delta)$, was simply the difference between the priors placed on the population-level mean capacity for each stimulus type (i.e., K_s). So, as an example, for the capacity differences between letters and colors, denoted as $p(\delta_{LC}) = p(K_L) - p(K_C)$, where $p(K_L)$ and $p(K_C)$ are the informative priors as described in the previous section (also see Table 1), samples were first drawn from $p(K_L)$ and $p(K_C)$ and then the difference between each pair of samples was calculated. This procedure yields the prior for the difference in capacity between color and letter stimuli. Similarly, the posterior distribution on the capacity differences is the difference between the obtained posterior distributions for K_L and K_C. Finally, we used a logspline density estimator to obtain the prior and posterior density estimates of the capacity differences (Wagenmakers, Lodewyckx, Kuriyal, & Grasman, 2010).

The Savage-Dickey density ratio is the ratio of the prior and posterior densities at the given point of interest. In the present case, the point of interest is $\delta = 0$. This ratio is called the Bayes factor and it quantifies the degree of support for the null hypothesis relative to the alternative, denoted here as BF_{01}. A Bayes factor greater than 1 provides support for the null hypothesis, whereas a Bayes factor less than 1 indicates support for the alternative. Conveniently, we may also express the degree of support for the alternative hypothesis relative to the null hypothesis in the following way: $BF_{10} = 1/BF_{01}$.

Results

Fig. 6 displays the posterior distributions of the model parameters for each stimulus type. The immediate impression from the figure is that the population-level means exhibited a very similar pattern to the estimates from Experiment 1. There was, however, a slight increase in the capacity estimates across stimulus type. However, we again see that the capacity for complex shapes was smaller, at around 2 items. Also, the comparison error parameters again remain fairly constant across all stimuli types. Our results also suggest that, while ostensibly participants may have been able to use the additional presentation time to encode a little more information into memory, they do not appear to use that extra time to remember a large number of shapes, at a cost of more comparison errors. That is, even with more time to encode items, the number of complex items that can be stored is smaller than that for simpler items, and this limit is below the assumed fixed capacity of 3–4 items. Crucially, the results from Experiment 2 replicate the results reported for Experiment 1.

Turning next to the Bayesian hypothesis tests, our immediate inferential focus was on the differences between the capacity estimates for complex shapes relative to all other stimulus types. The Bayes factors for all capacity differences are displayed in Table 2, though we present only the prior and posterior distributions for the shape stimuli comparisons in Fig. 7. It is worth noting that because we used informative priors, we expect that capacity is different for shapes and other stimuli – the prior distributions fall to the right of the dotted horizontal line in Fig. 7. The posterior distributions from Experiment 2 indicate a solidification of such beliefs, such that the posterior distributions shift slightly further away from 0. The posterior distributions also have smaller variance than the priors, indicating increased certainty about the difference between the capacity for shape and other stimuli. When the difference between prior and posterior at 0 was quantified via a Bayes factor, we found that the alternative hypothesis was 608 times, 62 times, and 16 times more likely than the null model, which expects no difference in capacity, respectively (cf: Table 2). These results provide decisive evidence for capacity differences between the shape and other stimuli.

Finally, we performed a posterior predictive check to assess the fit of the model predictions to the empirical data. The empirical data (white) and model fits (color) for Experiment 2 are displayed in Fig. 8. The data exhibited a

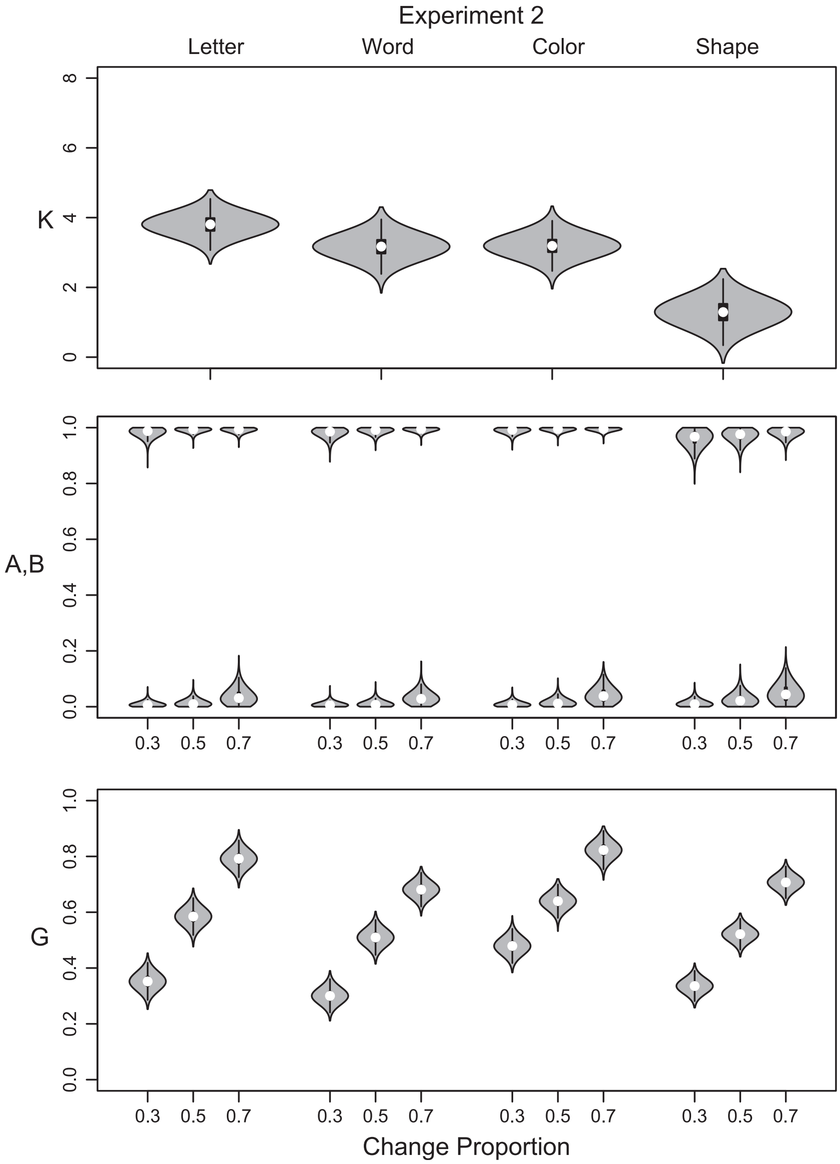

Fig. 6. Posterior distributions for the population-level parameter means. The top panel displays the capacity means across each stimulus type; the middle panel displays the mean comparison error probability across both stimulus type and change proportion for change, *A*, and same, *B*, trials; the bottom panel displays the guessing rates across each level of change proportion for each stimulus type.

Table 2
BF_{10} for the posterior difference between capacity estimates across stimulus type.

	Letter	Word	Colour	Shape
Letter	0	1.37	2.68	608.84
Word	–	0	0.57	62.48
Colour	–	–	0	15.97
Shape	–	–	–	0

very similar pattern to Experiment 1. The model again appeared to capture the general pattern of results across both set size and stimulus type. Most critically, the model describes the attenuation in the hit and false alarms rates for set sizes 5 and 8 for shape stimuli. Analogous to Experiment 1, the model did not perfectly capture all the empirical data patterns. For example, we see that the model predicts worse performance than was observed for letter and word stimuli. This mis-prediction is likely because performance for letters and words was better in Experiment 2 than Experiment 1, and our model used informative priors that were based on Experiment 1. That said, we do not think that these mis-predictions necessarily threaten the validity of our conclusions about capacity and comparison errors. For example, with the exception of the set size 2 condition for shapes, none of the misfits are systematic across the two experiments. That is, while the model may under predict performance in a particular

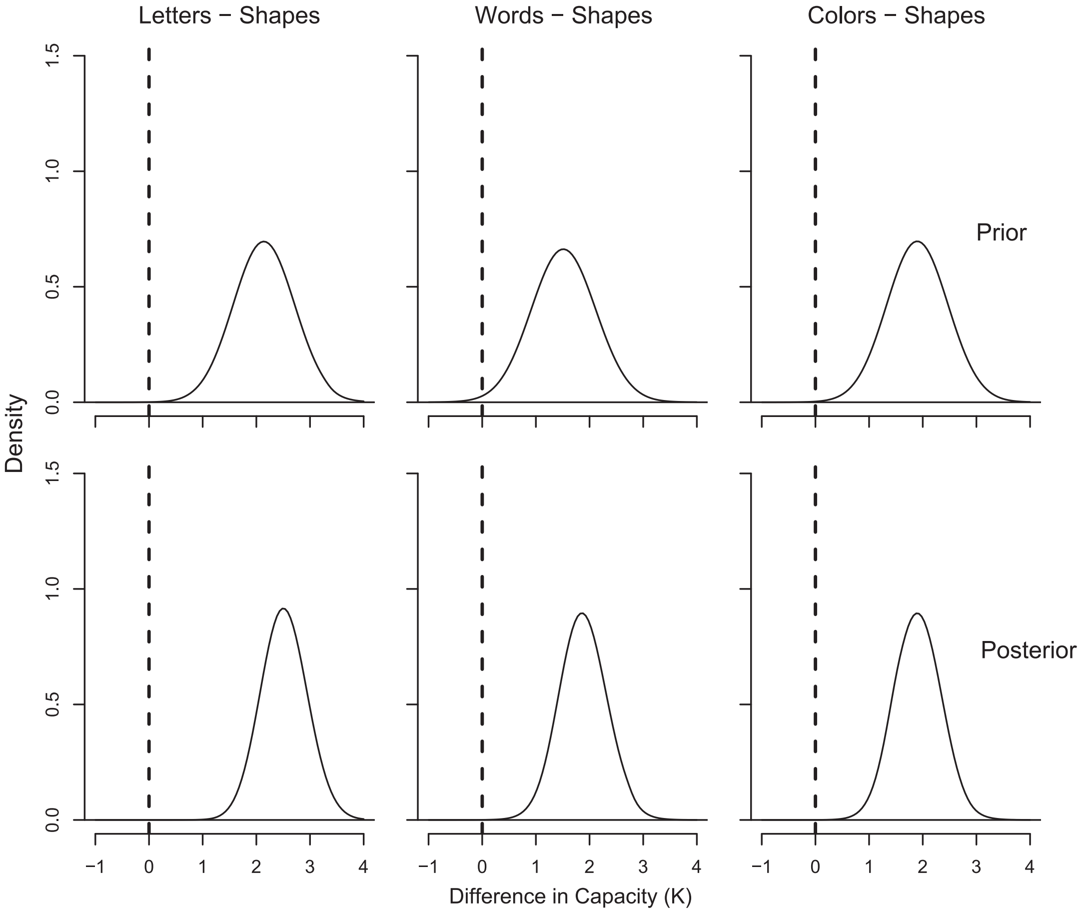

Fig. 7. Top Panel: Prior on the difference between the population-level mean capacity estimates across stimulus type in Experiment 2. These prior distributions are derived by taking the difference between the informative prior distributions placed upon the population-level means. **Bottom Panel**: Posterior distribution of the difference between the population-level mean capacity estimates across stimulus type. These distributions are derived by taking the difference between the posterior population-level mean estimates across stimulus type having conditioned on the data from Experiment 2.

condition in Experiment 1, no such misfit appears in Experiment 2. Furthermore, the misfits relate less to the pattern of attenuation (or squashing) in the rates for the set size 5 and 8 conditions for the shape stimuli, which was the signature pattern of smaller capacity we highlighted in Fig. 3.

General discussion

Our results suggest that changes in stimulus complexity produce negligible changes in the rate of comparison errors across stimulus types. Instead, the present findings indicate that item capacity is sensitive to changes in the complexity of the stimuli, and suggest that the original interpretation of the result in Alvarez and Cavanagh (2004) was correct – observers store fewer complex stimuli. Specifically, we find that when people undergo change detection tasks using perceptually simple stimuli (i.e, letters, words, and colors), capacity estimates tend to hover about the accepted item limit of approximately 3–4 items. However, when participants are subjected to stimuli that are more complex (i.e., abstract polygons) we find capacity estimates drop to between 1 and 2 items. Our results are thus more consistent with the conclusions made by Alvarez and Cavanagh (2004) and Brady and Alvarez (2015) regarding capacity limits for complex stimuli.

Our conclusions differ from those of Awh et al. (2007), who showed that cross-category change detection for complex stimuli was as good as change detection for simple stimuli. However, our experiment differs from that of Awh et al. (2007), in that participants only ever needed to remember items from just one stimulus category. In the Awh et al. (2007) experiments, perhaps when participants were asked to remember items from multiple categories, they encoded only enough information to be able to identify the particular type of stimulus. In that way, participants would have been able to make accurate cross-category decisions, but would have made more errors on within-category decisions (Brady & Alvarez, 2015). The participants in our experiments, however, appear to have encoded more information about a given stimulus, allowing for more accurate within-category decisions, but were thus only able to store a smaller number of those items. Such an explanation for our results suggests that participants are capable of allocating their mnemonic resource in a flexible manner. That is, they can store fewer items

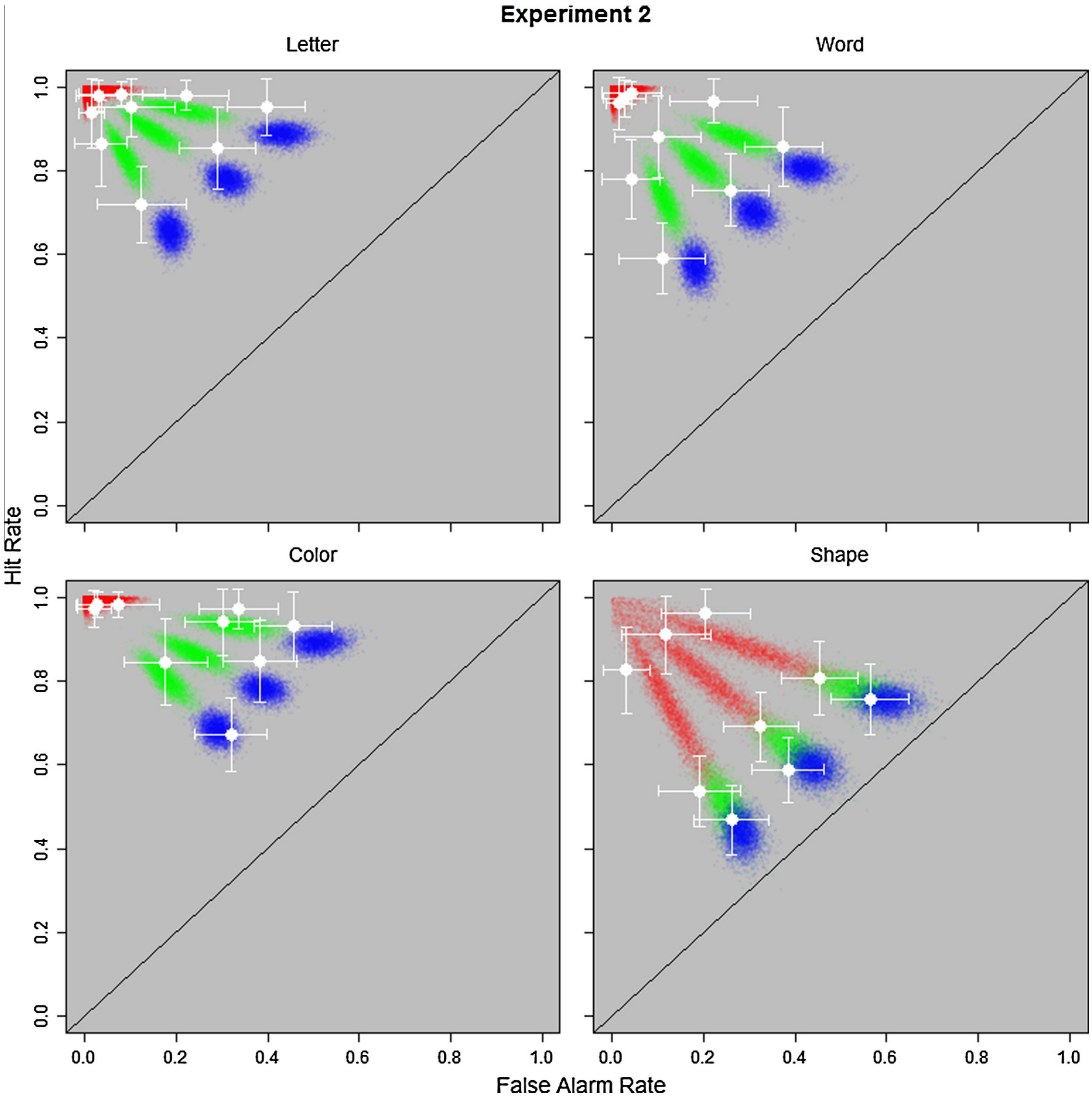

Fig. 8. Posterior predictive distribution for the comparison error model across each stimulus condition in Experiment 2. Each color band corresponds to a different set size condition: red = set size 2; green = set size 5; and blue = set size 8. Within each band there are three distinct clusters, each of which correspond to the different change proportion conditions. The white data points are the mean hit and false alarms rates for the observed data, with 2 standard deviations indicated. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

but with higher resolution, or more items with poorer precision.

Our results, on the surface, appear inconsistent with recent results reported in Jackson, Linden, Roberts, Kriegeskorte, and Haenschel (2015). Jackson et al. (2015) independently manipulated the similarity and complexity of visual stimuli, and found that it was only the similarity between stimuli, but not complexity, that increases the rate at which errors occur in a change detection task. However, we note that the capacity estimates for the stimuli used in this study hovered about 1.5–2 items, lower than the usual capacity of 4 items. As such, despite manipulating the relative complexity of their stimuli, we expect that Jackson et al. used a fairly complex set of visual stimuli. In other words, we expect that their simplest stimuli were as difficult to remember as the stimuli that are generally considered to be visually complex stimuli.

Our model makes no theoretical distinction between whether comparison errors occur due to an increase in similarity between items, or because the items become more 'complex'. However, this point is irrelevant with respect to our fundamental conclusion. That is, if the complex stimuli lead our participants to commit comparison errors they might have occurred because the shapes were more similar to one another, or because they were more visually complex. Our model, however, is agnostic to such descriptions; moreover, we found no evidence for these types of errors in our data. If comparison errors did exist in the data, our model has a demonstrated capacity to measure them. Simply, regardless of how one might describe the process for comparison error, our model classified the poorer performance with complex stimuli as being due to fewer items being stored, and not because of an increase in comparison errors.

It is worth being explicitly clear that we are not proposing the model we used here as a theoretical account of visual working memory capacity. Rather, this model was developed and used to distinguish between two very clear and distinct explanations for an empirical effect. The model was developed as a direct instantiation of Awh et al.'s (2007) alternative explanation of the empirical result reported in Alvarez and Cavanagh (2004). We then used this model to learn that the original interpretation of the result in Alvarez and Cavanagh (2004) was correct – observers store fewer complex than simple stimuli. There remains, however, a question of what particular process gives rise to such differences in capacity estimates.

One potential explanation may be the degree with which each stimulus type can be verbally encoded. While it is surely the case that the abstract shapes used here are not amenable to verbal encoding, it might be the case that categorical labels supplement encoding and improve performance with the color stimuli. For example, capacity for color stimuli was virtually identical to the capacity for letters and words. Accordingly, the capacity observed for complex shapes suggests that perhaps the *true* capacity of visual working memory is closer to 1–2 items and that simpler stimuli readily exceed this limit via the recruitment of verbal encoding mechanisms.

With respect to this line of reasoning, a recent paper by Sense, Morey, Prince, Heathcote, and Morey (2016) demonstrated that verbal recoding of visual stimuli is unlikely to occur during change detection tasks. Sense and colleagues found that whether articulatory suppression was included or not during the study phase had no differential effect upon change detection performance with simple colors. Additionally, the null effect persisted regardless of whether study items were presented in an array or sequentially. A state-trace analysis further suggested that there was no interaction between articulatory suppression and performance. That is, performance could be adequately summarized by invoking a single latent dimension relating to visual memory. Sense et al. (2016) conclude that if individuals had employed any verbal strategies then these strategies had no demonstrable effect upon performance. In light of these findings it is unlikely that the differences between capacity estimates reported here are merely an artifact of comparing stimuli that differ in how well they can be verbalized. Instead, the complexity of the stimulus imposes a limit that is fundamentally different to the limit imposed by simpler items.

Our belief is that visual working memory is a continuous resource, but that not all items may be encoded into memory on a given trial (Donkin, Kary, Tahir, & Taylor, 2016; Donkin, Nosofsky, Gold, & Shiffrin, 2013; Donkin, Tran, & Nosofsky, 2014; Nosofsky & Donkin, 2016). That is, though we have presented a model that is firmly rooted within the "slots" framework, our position is that these changes in capacity estimates point to the allocation of a flexible resource. Simple stimuli require less of that resource, and so more of those stimuli can be encoded. Complex stimuli, however, require a larger amount of mnemonic resource to be encoded with the same precision, and so fewer items can be remembered. It is interesting to note that such an account does allow for the possibility that participants would allocate less resource to more complex items, and so be able to remember more complex items. Had participants used such an encoding strategy, then our results may have turned out to have been in agreement with Awh et al. (2007). Thankfully, it appears that participants prefer to have high-precision representations for fewer complex items, thus allowing us to discriminate between our two competing hypotheses.

An alternative explanation for our results may be provided by a "slots + averaging" account. According to the slots + averaging theory, it is possible to encode an item into multiple slots. An item that has been stored in multiple slots will yield a more precise representation. One could argue that participants choose to allocate multiple slots to more complex items, in order to gain a precise representation of those items. Though possible, we think that the distinction between slots and continuous resource models becomes largely semantic once such possibilities are introduced. If one allows resource to be allocated within a slots framework then the resolution, or granularity, of that resource is simply a matter of degree. Further, slots + averaging models have not fared well when their quantitative predictions have been contrasted with continuous resource alternatives (e.g., van den Berg, Awh, & Ma, 2014; van den Berg, Shin, Chou, George, & Ma, 2012). Moreover, our Supplementary Materials contains an instantiation of this type of model. We found that the model provided virtually identical results and loaded the poorer performance for the complex stimuli onto the capacity parameter. The model also yielded noisier posterior predictions and did not provide an improved description of the empirical data.

In summary, we presented a model that allowed for the simultaneous estimation of capacity and comparison errors in a simple working memory task. Across two such change detection experiments we found that participants were able to store fewer complex items than simpler objects. Our results are inconsistent with the notion of a single, fixed capacity of visual working memory.

Acknowledgements

Thank you to Robert Nosofsky, with whom the data from our pilot experiment was collected. All data and code is available on the OSF: https://osf.io/rhmu2/. Chris Donkin's contribution to this research was supported by the Australian Research Council (DP130100124; DE130100129).

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jml.2016.09.004.

References

Alvarez, G. A., & Cavanagh, P. (2004). The capacity of visual short-term memory is set both by visual information load and by number of objects. *Psychological Science, 15*(2), 106–111.

Awh, E., Barton, B., & Vogel, E. K. (2007). Visual working memory represents a fixed number of items regardless of complexity. *Psychological Science, 18*, 622–628.

Barton, B., Ester, E. F., & Awh, E. (2009). Discrete resource allocation in visual working memory. *Journal of Experimental Psychology: Human Perception and Performance, 35*, 1359–1367.

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*, 921.

Cowan, N., Blume, C., & Saults, J. S. (2013). Attention to attributes and objects in working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*, 731–747.

Cowan, N., Elliott, E. M., Saults, J. S., Morey, C. C., Mattox, S., Hismjatullina, A., et al. (2005). On the capacity of attention: Its estimation and its role in working memory and cognitive aptitudes. *Cognitive Psychology, 51*, 42–100.

Donkin, C., Kary, A., Tahir, F., & Taylor, R. (2016). Resources masquerading as slots: Flexible allocation of visual working memory. *Cognitive Psychology, 85*, 30–42.

Donkin, C., Nosofsky, R. M., Gold, J., & Shiffrin, R. M. (2013). Fixed slots models of visual working memory response times. *Psychological Review, 120*, 873–902.

Donkin, C., Tran, S., & Nosofsky, R. M. (2014). Landscaping analyses of the ROC predictions of discrete-slots and signal-detection models of visual working memory. *Attention, Perception & Psychophysics, 76*, 2103–2116.

Fougnie, D., Asplund, C. L., & Marois, R. (2010). What are the units of storage in visual working memory? *Journal of Vision, 10*, 1–11.

Fukuda, K., Awh, E., & Volgel, E. K. (2010). Discrete capacity limits in visual working memory. *Current Opinion in Neurobiology, 20*, 177–182.

Hardman, K. O., & Cowan, N. (2015). Remembering complex objects in visual working memory: Do capacity limits restrict objects or features. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41*, 235–247.

Jackson, M. C., Linden, D., Roberts, M. V., Kriegeskorte, N., & Haenschel, C. (2015). Similarity, not complexity, determines visual working memory performance. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41*, 1882–1892.

Kary, A., Taylor, R., & Donkin, C. (2016). Using Bayes factors to test the predictions of models: A case study in visual working memory. *Journal of Mathematical Psychology, 72*, 210–219.

Kucera, H., & Francis, N. (1967). *Computational analysis of present-day american english*. Brown University Press.

Luck, S. K., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature, 390*, 279–281.

Ma, W. J., Husain, M., & Bays, P. M. (2014). Changing concepts of working memory. *Nature Neuroscience, 17*, 347–356.

Morey, R. D., Morey, C. C., Brisson, B., & Tremblay, S. (2012). A critical evaluation of c as a measure of mnemonic resolution. *Journal of Experimental Psychology: Human Perception and Performance, 38*(4), 1069.

Nosofsky, R. M., & Donkin, C. (2016). Response-time evidence for mixed memory states in a sequential-presentation change-detection task. *Cognitive Psychology, 84*, 31–62.

Oberauer, K., & Eichenberger, S. (2013). Visual working memory declines when more features must be remembered for each object. *Memory & Cognition, 41*, 1212–1227.

Olson, I. R., & Jiang, Y. (2002). Is visual short-term memory object based? rejection of the "strong-object" hypothesis. *Perception & Psychophysics, 64*, 1055–1067.

Rouder, J. N., Province, J. M., Swagman, A. R., & Thiele, J. (submitted for publication). From ROC curves to psychological theory. http://pcl.missouri.edu/sites/default/files/p_6.pdf, http://pcl.missouri.edu/sites/default/files/sup.pdf.

Scolari, M., Vogel, E. K., & Awh, E. (2008). Perceptual expertise enhances the resolution but not the number of representations in working memory. *Psychonomic Bulletin & Review, 15*(1), 215–222.

Sense, F., Morey, C. C., Prince, M., Heathcote, A., & Morey, R. D. (2016). Opportunity for verbalization does not improve visual change detection performance: A state-trace analysis. *Behavior Research Methods*, 1–10.

Umemoto, A., Scolari, M., Vogel, E. K., & Awh, E. (2010). Statistical learning induces discrete shifts in the allocation of working memory resources. *Journal of Experimental Psychology: Human Perception and Performance, 36*(6), 1419.

van den Berg, R., Awh, E., & Ma, W. J. (2014). Factorial comparison of working memory models. *Psychological Review, 121*, 129–149.

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*, 8780–8785.

Vogel, E. K., Woodman, G. F., & Luck, S. J. (2001). Storage of features, conjunctions, and objects in visual working memory. *Journal of Experimental Psychology: Human Perception and Performance, 27*, 92–114.

Wagenmakers, E. J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the savage–dickey method. *Cognitive Psychology, 60*, 158–189.

Wilken, P., & Ma, W. J. (2004). A detection theory account of change detection. *Journal of Vision, 4*, 1120–1135.

Zhang, W., & Luck, S. J. (2008). Discrete fixed-resolution representations in visual working memory. *Nature, 453*(7192), 233–235.