Cognitive Psychology 121 (2020) 101305

Contents lists available at ScienceDirect

Cognitive Psychology

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

Is there a K in capacity? Assessing the structure of visual short-term memory

Maria M. Robinson[a,*], Aaron S. Benjamin[b], David E. Irwin[b]

[a] Department of Psychology, University of California, San Diego, United States
[b] Department of Psychology, University of Illinois at Urbana-Champaign, United States

ARTICLE INFO

Keywords:
Visual short-term memory
Quantitative models of memory
Capacity and resource limits
Signal detection theory

ABSTRACT

Visual short-term memory (VSTM) is a cognitive structure that temporarily maintains a limited amount of visual information in the service of current cognitive goals. There is active theoretical debate regarding how limits in VSTM should be construed. According to *discrete-slot* models of capacity, these limits are set in terms of a discrete number of slots that store individual objects in an all-or-none fashion. According to alternative *continuous resource* models, the limits of VSTM are set in terms of a resource that can be distributed to bolster some representations over others in a graded fashion. Hybrid models have also been proposed. We tackled the classic question of how to construe VSTM structure in a novel way, by examining how contending models explain data *within* traditional VSTM tasks and also how they generalize *across* different VSTM tasks. Specifically, we fit theoretical ROCs derived from a suite of models to two popular VSTM tasks: a change detection task in which participants had to remember simple features and a rapid serial visual presentation task in which participants had to remember real-world objects. In 3 experiments we assessed the fit and predictive ability of each model and found consistent support for pure resource models of VSTM. To gain a fuller understanding of the nature of limits in VSTM, we also evaluated the ability of these models to jointly model the two tasks. These joint modeling analyses revealed additional support for pure continuous-resource models, but also evidence that performance across the two tasks cannot be captured by a common set of parameters. We provide an interpretation of these signal detection models that align with the idea that differences among memoranda and across encoding conditions alter the memory signal of representations in VSTM.

1. Introduction

Visual short-term memory (VSTM) is a fundamental cognitive structure that temporarily maintains visual information in the service of active goals. Although there is consensus that VSTM is capacity limited, the mental currency of these capacity limits is a topic of active theoretical debate. According to the *discrete-slot* view, VSTM capacity is limited by a discrete number of slots that can store individual items in an all-or-none fashion (e.g., Cowan, 2001). According to an alternative *continuous resource* view, VSTM is limited by a continuous resource that can be distributed across memory representations to prioritize some input over other (e.g., Bays & Husain, 2008). These models differ fundamentally in how they represent information in VSTM. One key prediction of standard discrete-slot models is that capacity is set in terms of the number of objects a person can store in memory. Strong versions of these models predict that remembered items are stored with complete fidelity or not at all, and that there is no tradeoff between the quality

* Corresponding author at: Department of Psychology, University of California, San Diego, 9500 Gilman Dr., La Jolla, CA 92093, United States.
E-mail address: mrobinson@ucsd.edu (M.M. Robinson).

https://doi.org/10.1016/j.cogpsych.2020.101305
Received 20 September 2018; Received in revised form 8 April 2020; Accepted 13 April 2020

of item representations and the number of items that can be maintained. In contrast, continuous resource models predict that features and items can be stored with variable resolution and resources can be distributed in such a way so as to bolster some representations over others. This capacity can lead to a trade-off between the number of stored items and the fidelity with which they are coded in memory. Hybrid models that combine assumptions from the discrete-slot and resource models have also been proposed (e.g., Zhang & Luck, 2008).

Despite ongoing disagreement regarding which model of VSTM provides the best description of VSTM architecture, the discrete-slot model remains the dominant model in cognitive psychology (e.g., Cowan, 2001; Luck & Vogel, 2013; Luria, Balaban, Awh, & Vogel, 2016). For example, it is very common for researchers to report VSTM capacity in terms of a measure known as Cowan's K, despite the fact that this measure is only meaningful if the assumptions of the discrete-slot model are correct. How we characterize performance from data in VSTM tasks has significant implications both for understanding the structure of VSTM and also for the choice of proper measurements of VSTM capacity in experimental research. In the current manuscript we review and discuss some limitations of previous studies that have formally tested models of VSTM representation and capacity. The first aim of our research is to replicate these studies while redressing some of these methodological and analytic limitations.

The second aim of our research is to address a fundamental and, surprisingly, unexplored question in the VSTM modeling literature, namely: how well do the contending models generalize? The major way in which we address this question is by evaluating models of VSTM capacity across different experimental paradigms. The first way we do this is by using multiple and materially different VSTM tasks. These tasks differ in both the type of stimuli and the presentation format, and we fit models separately to data from each type of task. This analysis elucidates whether discrete-slot or resource models are good descriptive models of VSTM processing in both types of tasks. The second way in which we assessed performance across these markedly different tasks is by jointly modeling performance across them. Such models allow us to directly test the prediction that a single parameter of capacity or resource is shared across these tasks. A similar type of joint modeling analysis was recently applied in the long-term memory literature to examine the degree to which common cognitive processes can be captured with specific model parameters across different long-term memory tasks (Cox, Hemmer, Aue, & Criss, 2018). Joint modeling of this kind goes beyond prior work that has examined whether there is overlap between different working memory and attention tasks with covariance-based techniques (Unsworth, Fukuda, Awh, & Vogel, 2014), because it directly assesses the stronger prediction of how the common latent construct should be construed. The computational modelling approach employed here addresses a novel and fundamental question that, to our knowledge, has not yet been addressed with formal quantitative methods in the VSTM modeling literature: Is there evidence that different VSTM tasks can be understood with a single model of capacity or resource?

We also assess how well the models generalize by guiding model selection with cross-validation as well as measures of model fit. Cross-validation analysis is routinely applied in machine learning and statistics (e.g., Murphy, 2012), though it is rarely used to evaluate models in the social sciences. This approach to model selection is desirable because models may provide a superior fit to a sample of data by erroneously fitting noise rather than variance generated by a psychological process of interest (e.g., Pitt & Myung, 2002). A consequence of overfitting is that such models do a poor job of generalizing to new samples of data. As such, assessing models based on their predictive rather than explanatory capability may help produce theories that provide an accurate understanding of and prediction of behavior (Yarkoni & Westfall, 2017).

1.1. Review

1.1.1. Genesis of the discrete-slot model of VSTM

The discrete-slot model of VSTM gained popularity following a study by Luck and Vogel (1997; see also Vogel, Woodman, & Luck, 2001), who used the change detection task to measure people's memory for a range of stimuli that varied in complexity. In those experiments, participants were briefly shown a visual array that contained stimuli, such as shapes (squares) of different colors in different locations. A brief period after the first array offset, participants were shown a second test array in which one of the squares may have changed color and were asked to indicate whether a change did or did not occur. The authors found that, on average, performance declined sharply as the number of to-be-remembered items surpassed three or four. They also suggested that the capacity of VSTM was not limited by the number of features, but rather by the number of objects (i.e., bound features) that could be maintained. That is, participants could maintain information about three to four objects defined by a single feature (stimuli that differed only in color or only in orientation), or three to four objects that were defined by a conjunction of features (stimuli that could differ in either color or orientation, or both). This finding has been used to motivate discrete-slot models of VSTM because it suggests that the unit of storage in VSTM is set in terms of slots, each of which can be filled with objects of bound features and thus arbitrary complexity. The results of follow-up studies examining the effects of stimulus complexity on memory capacity have been mixed, however, sometimes showing diminished capacity for more complex items (contrary to the discrete slot model prediction) and sometimes not (Alvarez & Cavanagh, 2004; Awh, Barton, & Vogel, 2007; Fougnie, Cormiea, Kanabar, & Alvarez, 2016; Fukuda, Vogel, Mayr, & Awh, 2010; Wheeler & Treisman, 2002).

1.1.2. Model-based measures of capacity

One limitation of studies that have failed to find evidence for a tradeoff between the quantity and quality of memory representations is that they used a measure to assess memory capacity that assumes that discrete-slot models are accurate. This measure of capacity is called the K metric (Cowan, 2001; Pashler, 1988), and it is based on a class of cognitive models that assume that memory operates in an all-or-none fashion (Blackwell, 1953; Swets, 1986). Importantly, this metric does not explain performance in the same way as measures based on continuous resource models of memory. In the context of the change detection task in which a

single item is shown in the test array, the discrete-slot model assumes that a person's hit rate (i.e., correctly responding that a change occurred when it did occur) is driven by the probability that the item is remembered (with probability K/N—that is, a person's capacity [K] divided by the number of to-be-remembered items [N]) and the probability that the subject lacks knowledge but guesses correctly (with probability $[1 - \frac{K}{N}] * g$) that a change occurred. Furthermore, the model assumes that false alarms (i.e., incorrectly responding that a change occurred when it did not occur) are determined by the probability that an item is not in memory and an individuals' guessing bias (with probability $[1 - \frac{K}{N}] * g$). Solving for K requires taking the difference between hits and false alarm rates and weighting this difference by the number of to-be-remembered items (i.e.,$K\ capacity = [H - FA] * N$). It is critical to note that this model does not represent variability in the resolution of the memory representations because items are assumed to be either completely stored in memory or not stored at all. Thus, whether they realize it or not, any investigator who employs the K metric to assess VSTM capacity (and many researchers do) is implicitly adopting an all-or-none discrete slot model of VSTM.

In contrast to discrete-slot models, continuous resource models posit that representations in VSTM can be stored with variable resolution. Such models are based on the assumption that there is both external and internal noise in processing that corrupt memory representations. Noise increases as people increase the amount of visual input that they try to encode into VSTM, yielding a tradeoff between fidelity and quantity. Under this account, VSTM is not limited by the number of items that it can maintain, but rather relies on a continuous resource that can be distributed among features and items. In the change detection task, resource models have been tested with models based on signal detection theory (Green & Swets, 1988), a theory that naturally accommodates the assumption that evidence is continuous. Within the signal detection framework participants must make a decision based on some amount of evidence that a change did or did not occur. Even on no-change trials some evidence for change will be present due to noise in the system; it is both convenient and theoretically parsimonious to assume that this noise is normally distributed (Wickens, 2001). On change trials, the test stimulus is actually different from what is in memory, leading to even more evidence that a change occurred on average. This is represented by a second normal distribution (*noise* + *change*). Following standard signal detection theory, the standardized distance (d') between the two normal distributions reflects an individuals' discrimination of change from no-change trials. Furthermore, because there is always some evidence that a change occurred, individuals must set a decision criterion C for responding that a change occurred. From a psychological standpoint d' can be seen as an index of the fidelity of the VSTM representation, whereas C is participants' response bias to endorse a change versus a no change response.

Unlike discrete-slot models, continuous resource models provide a continuous measure of the resolution of memory representation and explicitly quantify a person's tendency to endorse a specific response (e.g., change versus no-change). Critically, capacity estimates based on formulae derived from discrete-slot and continuous resource models do not lead to identical conclusions regarding the effects of a given manipulation on a person's capacity. This point is illustrated in Table 1, which shows simulated data for four subjects in an experiment in which there are six to-be-remembered items (N). Comparing data for Participants 1 and 2 reveals that the same K estimate of capacity can correspond to different values of d' and C response bias. The converse is also true. Participants 3 and 4 exhibit large differences in K, but within the signal detection framework these differences are assumed to reflect differences in decision criterion rather than memory. This set of examples illustrates that determining which models best characterize VSTM storage and representation has significant implications not only for the understanding of VSTM structure, but also for interpreting the effects of a given experimental manipulation on people's capacity. We note that discrepancies between the two metrics are not limited to hypothetical data: in the General Discussion we point to and discuss similar patterns in our own results.

1.1.3. Evidence from formal model fitting

A formal approach to testing between discrete-slot and signal detection models in the change detection task involves Receiver Operating Characteristic (ROC) analysis (Swets, 1986). In ROC analysis, hit rates are plotted as a function of false alarm rates for different values of a participant's response bias, that is, his or her tendency to endorse a specific response (e.g., change versus no change). The discrete-slot model and signal detection-based resource model make distinct predictions about the shape of ROC curves. Specifically, the standard discrete-slot model predicts a straight ROC curve because hit rates increase as a linear function of false alarm rates. In contrast, for partially overlapping noise and signal distributions where the variance of noise and signal distributions is equal (set to 1), a signal-detection model predicts a symmetric curvilinear ROC function (Swets, 1986).

ROC analyses of VSTM data in the change detection task have yielded conflicting results. For instance, Wilken and Ma (2004) found that a symmetric curvilinear function provided the best fit to their data, a finding that supports a common variant of the continuous resource model called the equal-variance signal-detection model (EVSD). That finding is difficult to interpret, however, because they used ROC curves based on data aggregated across subjects. Assessing models based on pooled data may lead to average ROC curves that do not correspond to any individual's data (e.g., Estes, 1956), so it is possible that Wilken and Ma's results reflect an aggregation artifact.

Table 1
Hypothetical data that demonstrate how metrics based off the discrete-slot and signal detection models may lead to different conclusions regarding VSTM processing.

Participant	N	H	FA	$H - FA$	K	d'	C
1	6	0.9	0.41	0.49	2.94	1.51	−0.527
2	6	0.5	0.01	0.49	2.94	2.33	1.16
3	6	0.224	0.015	0.209	1.25	1.41	1.46
4	6	0.91	0.471	0.439	2.63	1.41	−0.634

1.1.4. Discrepancy in modeling results from change detection and delayed estimation tasks

Wilken and Ma (2004) reported another result with a *delayed estimation task* that is quite persuasive, however. In that task, subjects reported their memory of a feature of a probed item using a continuous scale, such as a color wheel. They found that participants' reports became more variable with increasing set size, a result straightforwardly inconsistent with the idea that items were either successfully encoded in full detail or not. This result indicates that memory precision declines with increased memory load. In a later study, Zhang and Luck (2008) found a similar pattern of results and developed a specialized hybrid model that combined processing assumptions from both the discrete-slot and resource models. This model assumes that memory performance is determined by two factors, a discrete item capacity and a continuous attention-like process that determines the resolution with which items are coded in memory. Since the Wilken and Ma (2004) study, the delayed estimation task has become the canonical VSTM task for evaluating VSTM models and all such studies have reported support either for hybrid or for resource models of VSTM.

In contrast, several studies that have used the change detection tasks and ROC analysis to evaluate these models have reported linear ROC functions and thus supported the predictions of pure discrete-slot models. Both Rouder et al. (2008) and Donkin, Tran, and Nosofsky (2014) reported that a model based on discrete representations outperformed a continuous resource model. There are, however, some important reasons to question the generality of their results. We review those concerns in the next section and then report a set of new experiments that employ what we see as procedural and analytic enhancements over prior work.

1.1.5. Confidence ratings vs. change probability manipulations of bias

One concern about the findings of Rouder et al. and Donkin et al. lies in how these authors manipulated participants' response bias. A standard practice for collecting responses over different levels of response bias is to collect confidence ratings. This technique allows the researcher to sample a wide range of levels of response bias (or points on the ROC curve) with fewer observations (Wickens, 2001). This was the approach taken by Wilken and Ma (who, it will be recalled, reported evidence for resource models of capacity). In contrast, Rouder et al. and Donkin et al. experimentally induced bias by manipulating the probability of a change occurring across the experimental session. A known limitation of this approach is that changing true base rates in this way may produce changes in accuracy, or in response strategies—an effect that would violate assumptions of ROC analyses of this type (Balakrishnan, 1999; Dube & Rotello, 2012; Markowitz & Swets, 1967). A further concern is that, if participants are not sufficiently sensitive to the probability manipulation, this may lead to clustering of points in ROC space (Donkin et al., 2014; Healy & Jones, 1975; Healy & Kubovy, 1978). Indeed, perusal of the Rouder et al. data suggests the presence of a very restricted range on the ROC. If these authors were sampling points on a sufficiently small portion of a curvilinear ROC curve, it is nearly impossible to discriminate between a linear and a curvilinear form of the underlying latent function. Furthermore, several of the reported individual ROC curves in Donkin et al.'s study were nonmonotonic, leaving open the possibility that participants may have not been sensitive to all levels of the bias manipulation. A similar critique of manipulating base rates to construct ROC curves has been made by Dube, Rotello, and Heit (2010), who reported that researchers may erroneously infer evidence for linear ROC functions under conditions where there is a clustering of ROC points. It is worth noting, however, that the confidence-rating procedure is not without concerns as well (e.g., Benjamin, Tullis, & Lee, 2013; Van Zandt, 2000). We discuss and address these limitations in the General Discussion.

1.2. Current research

1.2.1. Addressing limitations of prior studies

The first aim of the set of experiments reported here is to address limitations of the studies reviewed above. To this end, we (1) used confidence ratings instead of a change probability manipulation in order to sample a wide range of points on the ROC curve and (2) we fit models to individual-subject data rather than to aggregate data to avoid averaging artifacts. As detailed in the next section, we also (3) tested a wider range of models, including a newly developed variant of the mixture model (Zhang & Luck, 2008) designed specifically to fit ROC data in the change detection task (Xie & Zhang, 2017). We additionally (4) fit models to data from two substantially different VSTM tasks (described below) to examine whether modeling results are robust across different experimental paradigms. Finally, we (5) applied joint modeling analysis to quantitatively test the prediction that a single latent variable of capacity, or resource, can capture performance across these different VSTM tasks.

1.2.2. Contending models

1.2.2.1. Pure slot models. The theoretical ROC curves for all contending models are shown in Fig. 1. We refer to the first class of models as *pure slot* models. This class of models includes the classic discrete-slot model for the single-probe version of the change detection task developed by Cowan (2001). The standard version of this model captures VSTM structure with the latent variable, K, which denotes an individual's VSTM capacity, or the number of object slots that comprise VSTM. As explained earlier, the model predicts that the probability of a hit (a change response on a change trial) is the union of the probabilities that the probed item is in memory or that the probed item is not in memory and a correct guess is made, with probability g, that a change occurred. Similarly, a false alarm occurs when the probed item is not in memory and an incorrect guess is made that a change occurred. Formally, these predictions are expressed by the following equations, where N denotes the number of to-be-remembered items and $\min(1, \frac{K}{N})$ denotes the model constraint that K does not exceed the number of to-be-remembered items.

$$P(Hit) = \min\left(1, \frac{K}{N}\right) + \left(1 - \min\left(1, \frac{K}{N}\right)\right) * g$$

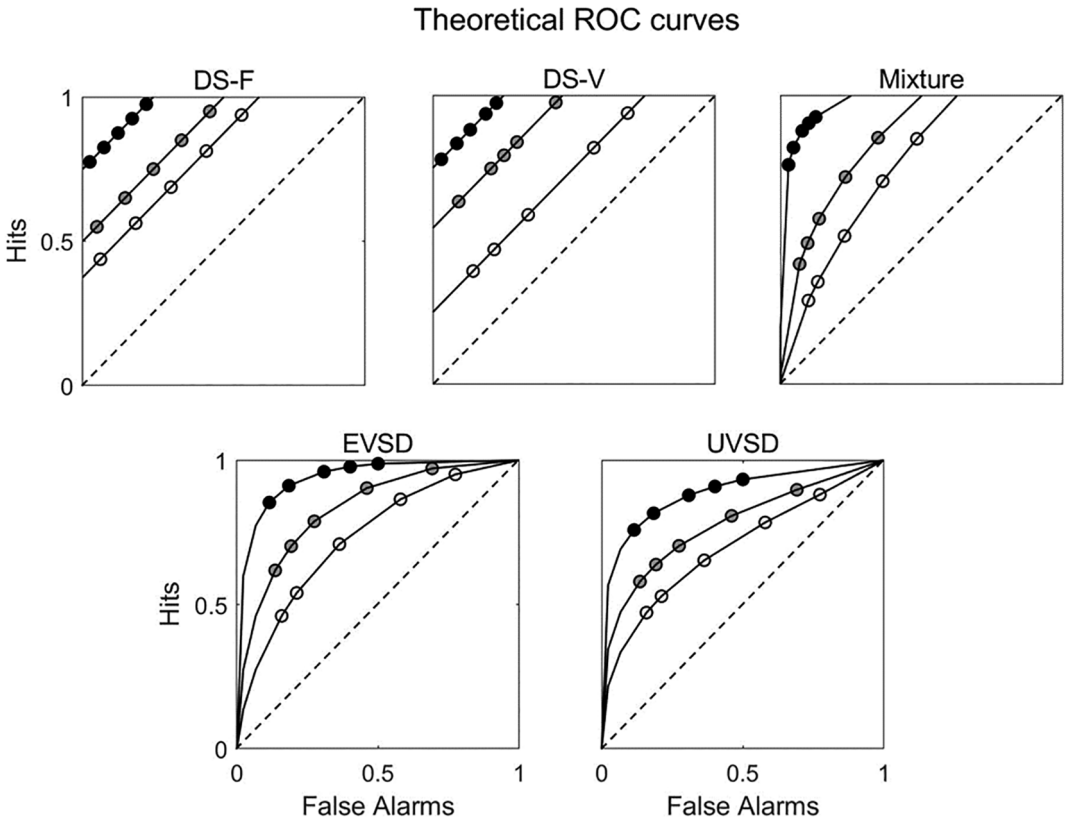

Fig. 1. The theoretical ROC curves for each candidate model. The three curves in each panel represent different set sizes (black = 4 items; grey = 6 items; white = 8 items). The first row shows predictions for models that assume an underlying item limit and the second row shows predictions for resource models. The d' and response criteria values used to generate these curves are identical for the mixture, equal and unequal variance signal detection models. Unequally spaced points on the ROC curve indicate that response criteria are free to vary across set sizes.

$$P(False\ alarm) = \left(1 - \min\left(1, \frac{K}{N}\right)\right) * g$$

In this manuscript, we consider two variants of the classic discrete-slot model. The first variant we will refer to as the *fixed discrete-slot* model (Fixed DS) because this model captures performance across different memory set sizes with a single latent variable (K); that is, K is fixed across memory set sizes. The second variant we refer to as the *free discrete-slot* model (Free DS) because K is free to vary across memory set sizes. We consider the Free DS model because some researchers have reported that estimates of K may vary as a function of the number of to-be-remembered items (e.g., Endress & Potter, 2014). This model implies that individuals bring different amounts of capacity depending on the task difficulty.

In addition to these models we consider a variant of the discrete-slot model that captures the possibility that in the change detection task people's performance on change trials can be informed both by their memory of the old item in the probed spatial location (which is no longer in the probed spatial location), as well as their memory of the lure (which is shown in the probed spatial location). The primary reason we consider this model is because in our change detection experiments on change trials we include trials in which lures were items shown in the original memory array but in a different spatial location. The requirement that participants have to remember the spatial location of the probed stimulus made our design different from the one used by Rouder et al., who always presented a new color (one not presented in the original memory array) on change trials. We opted for this design for the following two reasons. First, both Rouder et al. and we used a fixed and limited set of colors (ten). One consequence of this design choice is that the informational content of the memory arrays varies as a function of set size. In particular, participants can learn the number of possible colors (ten in total) throughout the experimental session. When eight items are shown in the memory array participants can scan the array (without attempting to memorize colors) for the two missing colors and attempt to make change/no change responses based on searching the test display for one of the missing colors. Higher memory set sizes lend themselves more to this strategy. A second reason for using recycled colors in lure stimuli is because it makes our design more comparable to the standard delayed estimation paradigm, in which participants are probed on the properties of a to-be-remembered item with a spatial probe (e.g., Zhang & Luck, 2008), therefore requiring participants to bind colors to specific spatial locations.

Nevertheless, our design introduces a different problem, which is that on change trials participants may rely not only on their memory of the probed item but also on their memory for the color of the item that switched locations to the probed item's position.

This approach is similar to the recall-to-reject strategy identified in tasks like associative recognition (Rotello & Heit, 2000) and with distractors that are highly similar to targets (Matzen, Taylor, & Benjamin, 2011). Although Donkin et al. (2014) used both versions of the change detection task (the one used by Rouder et al. and the one we use here) and reported no meaningful differences in performance between these tasks, we wanted to examine more directly whether participants may have employed such a strategy. To this end, we included a model that captures how memory for lures may affect performance on the change detection task. A version of this model was originally proposed by Cowan, Blume, and Saults (2013). We will refer to this model as the *alternative strategy discrete-slot* (Alt DS) model. Like the Fixed DS model, this model also predicts that memory is represented by a single latent variable (K). However, this model predicts that on change trials, the probability of a hit is the union of the probability that only the old item (the original item shown on the probed spatial location) is in memory, or that only the lure item (the item shown in the memory array in a different spatial location) is in memory, or that both of these items are in memory, or that neither of the two items are in memory but a correct guess is made that a change occurred. The probability of a false alarm is the same as it is for the standard Fixed DS model because it can only be informed by memory of the unchanged probed item. These predictions are given in the following equations, where c denotes the probability that an individual remembers both the probed item and the lure (first product), or only one of these items (second product).

$$c = \left(\min\left(1, \frac{K}{N}\right)\right) * \left(\min\left(1, \frac{K-1}{N-1}\right)\right) + 2 * \left(\min\left(1, \frac{K}{N}\right)\right) * \left(1 - \min\left(1, \frac{K-1}{N-1}\right)\right)$$

$$p(Hit) = \min(1, c) + (1 - \min(1, c)) * g$$

$$p(False\ \ alarm) = \left(1 - \min\left(1, \frac{K}{N}\right)\right) * g$$

1.2.2.2. Pure resource models. The second class of models we refer to as *pure resource models*. As noted, these models construe VSTM as consisting of a resource that is distributed in a continuous fashion across memory representations. From the signal detection framework, these models assume that people use continuous evidence to determine change no/change responses in the change detection task. We consider two major signal detection models. The equal-variance signal detection (EVSD) model assumes that VSTM can be represented by a single latent variable μ, which reflects the standardized distance between Gaussian noise and signal distributions and can be interpreted as the resolution, or strength, of memory representations[1]. According to this model, the probability of a hit is the probability that the amount of evidence generated from the signal distribution with mean μ exceeds the decision criterion c. The probability of a false alarm is the probability that the amount of evidence generated from the noise distribution with mean 0 exceeds the decision criterion. The predictions are given in the formulas below, where Φ denotes the cumulative normal distribution function.

$$P(Hit) = \Phi(c - \mu)$$

$$P(False\ alarm) = \Phi(c)$$

The second variant of the signal detection model we consider is the unequal variance signal detection (UVSD) model. According to this model VSTM can be represented by two latent variables, μ and σ, which reflect the standardized distance between noise and signal distribution and the variance of the signal distribution, respectively. Under some conditions, this model may capture the result that the standardized distance between the signal and noise distributions may vary over trials and across stimuli (DeCarlo, 2010). The decision rule for this model is the same as it is for the EVSD model except that the variance of the signal distribution is a free parameter. The prediction for hits and false alarms are given in the equations below.

$$P(Hit) = \Phi\left(\frac{(c - \mu)}{\sigma}\right)$$

$$P(False\ \ alarm) = \Phi(c)$$

The EVSD and UVSD model can be seen as analogues to general versions of the equal and variable precision models of VSTM, respectively. In many studies of long-term memory, the UVSD model has been shown to be superior (e.g., Wixted, 2007).

1.2.2.3. Hybrid (mixture) model. The final model we consider is a hybrid of the pure discrete-slot and pure resource models. Several variants of the mixture model have been developed for application to the delayed estimation task. In the current studies we consider the most general version of this model (also referred to as a slot + resource model), which was recently proposed by Xie and Zhang (2017). This model has a component that derives from discrete-slot models: it allows for complete information loss—that is, the possibility that an item is not stored in memory at all and people guess whether or not a change occurred. However, unlike pure discrete-slot models, this model also assumes that if an item is in memory it can be coded with variable resolution. Within the change detection task variability in resolution is represented with a signal detection component. The predictions for hits and false alarms of

[1] In this manuscript we define signal trials as 'change' trials, similar to Rouder et al. (2008). For the modeling results, there is no meaningful consequence to this choice over choosing to define 'same' trials as the signal.

this model are given in the equations below.

$$P(Hit) = \frac{K_T}{N} * \Phi(c - \mu) + \left(1 - \frac{K_T}{N}\right) * \Phi(c)$$

$$P(False\ \ alarm) = \left(1 - \frac{K_L}{N}\right) * \Phi(c)$$

In the variant of the model proposed by Xie and Zhang there are separate K parameters (K_T and K_L) that index performance on change and no change trials, respectively. Formally, this model is a special case of the mixture model proposed by DeCarlo (2010), with the difference that the model component that reflects failure to process the signal due to capacity limits ($\frac{K_T}{N}$ and $\frac{K_L}{N}$) in the VSTM mixture model Xie & Zhang, 2017), instead reflects failure to process the signal due to attentional lapses (represented with parameter λ) in the long-term memory version of the mixture model (DeCarlo, 2010). Furthermore, the mixture model (DeCarlo, 2010) is also formally equivalent to the UVSD under conditions in which the variance parameter in the UVSD model is larger on signal than noise trials. We emphasize these aspects of the mixture model to reinforce the point that, contrary to mainstream opinion (e.g., Zhang & Luck, 2008; Luck & Vogel, 2013), this model differs substantially from pure discrete-slot models both in its mathematical properties and its ability to model empirical data.

1.2.3. VSTM tasks

As noted, a central aim of our joint modeling analysis was to test the prediction that a single estimate of capacity, such as K or d', can capture performance across experimental paradigms that differ from one another. With this aim in mind, the tasks we chose differed in the type of stimuli and presentation format used. The first task we employed was a standard *change detection* task, similar to the one used by Wilken and Ma (2004), Rouder et al. (2008), and Donkin et al. (2014). In this task, participants were presented with two sequential displays of simple stimuli (e.g., colored squares) and then probed on whether any of the items changed across displays. The second task was one used by Endress and Potter (2014) to study VSTM (see also, Gorgoraptis, Catalao, Bays, & Husain, 2011; Nosofsky & Donkin, 2016; Berry, Waterman, Baddeley, Hitch, & Allen, 2018). In this *rapid serial visual presentation* (RSVP) task, participants were briefly presented with a sequence of photos of real-world objects, followed with a single probe stimulus that may or may not have appeared in the trial sequence.

Importantly, Endress and Potter manipulated whether images in the sequence were presented only a single time throughout the experimental session (unique condition) or repeatedly throughout the experimental session (repeated condition). The authors found that estimates of capacity increased without apparent bound in the unique condition, presumably because participants could rely on long-term memory to recognize images at test. In contrast, the authors found that estimates of capacity ranged from 2.3 to 4.8 in the repeated condition, similar to the capacity estimates typically found in change detection experiments (e.g., Luck & Vogel, 1997). The authors reasoned that participants could no longer rely on long-term memory because images presented on prior trials interfered with their ability to recognize images on the current trial (via proactive interference). In the current set of experiments, we used the repeated presentation version of the task because of its greater apparent reliance on VSTM. Thus, the two tasks used in our experiments differ both in how to-be-remembered information is presented (simultaneously versus sequentially) and the type of information that is presented (colored squares versus photos of real-world objects).

1.2.4. Model generalizability: Joint modelling of different VSTM tasks

A central aim of our research is to address the fundamental theoretical question of how well each of the candidate models generalizes across different paradigms (Experiments 2 and 3) and samples of data. The way in which we assessed model generalizability was by fitting candidate models *jointly* to the two different VSTM tasks to evaluate the degree to which a common model and set of parameters could accommodate performance across two procedures that differ considerably. As a computational approach, joint modeling of data from different tasks has been advocated as a powerful tool in the cognitive sciences because of its potential to build strong theories by bridging data across diverse experimental paradigms (Lee, 2011). Despite these benefits, it has never been used to evaluate models of VSTM.

Importantly, this type of joint modeling analysis goes beyond examining correlations across tasks. Evidence for correlations between tasks, where individuals who perform better on one task also perform better on a different task, does not indicate that the same processes are engaged across the two tasks, or that they are engaged in the same way across the two tasks. In contrast, the joint modeling approach we employ permits us to directly test the prediction that a single parameter corresponding to an individual's underlying capacity or resource can capture performance across different VSTM tasks (for related approaches, see Turner et al., 2013). Similarly, this approach differs from latent variable techniques (e.g., factor analysis), which help elucidate the question of *whether a latent construct is shared across tasks* (e.g., Unsworth et al., 2014). In contrast, the joint modeling approach employed here enables us to examine *how to best represent the latent construct across tasks* by testing theories that explicitly define key hypothetical constructs via computational models.

A similar joint modelling approach was recently applied to understanding relationships among long-term memory tasks (Cox et al., 2018). Those authors used a single measurement model to assess correlations between parameters estimated across different tasks. They used hierarchical Bayesian modeling to assess correlations at the level of individuals and stimuli to evaluate the degree to which the same memory processes are engaged across tasks, and the degree to which the same informational content is extracted from items across tasks.

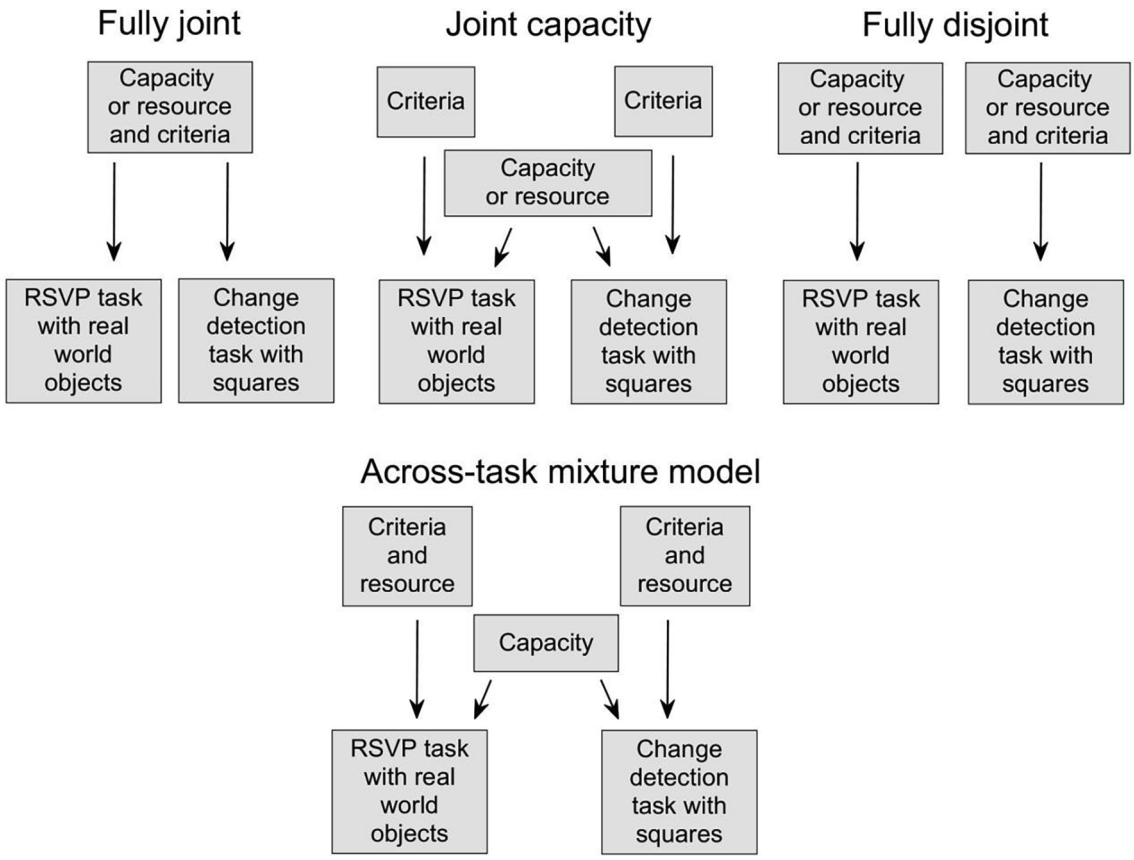

Fig. 2. Task models tested in the joint modeling analysis.

Our approach differs from that one because we directly test the much stronger prediction that a single latent variable, such as *K* capacity, can capture performance across the two different VSTM tasks. Specifically, for each type of model, we evaluate a 'task' model, which either assumes that capacity and criteria are shared across tasks (fully joint model); that only capacity is shared across tasks (joint capacity model); or that neither of these variables is shared across tasks (fully disjoint model). These models are represented in Fig. 2. Rather than using a single measurement model (Cox et al., 2018), we also compared different models to assess whether there is a model that best describes performance across tasks, either with a single set of parameters, or with distinct parameters.

1.2.5. Across-task fixed capacity + resource model

We leveraged this joint modeling analysis to evaluate a new 'across-task' variant of the mixture model and also to test a central prediction of models that assume a fixed item limit in VSTM. Our prediction is based on the finding that estimates of capacity based on pure resource models may differ for different stimuli. For instance, although Endress and Potter did find that capacity was limited in the repeated condition, overall capacity estimates for the real-world images that they used were somewhat higher than those observed for the simple stimuli (e.g., colored shapes) that are typically used in standard change detection experiments. Higher estimates of capacity for objects were also reported by Brady, Störmer, and Alvarez (2016), who compared capacity for objects and simple stimuli under conditions where both were presented simultaneously. Using discrete-slot measures of capacity, these authors found that participants could remember approximately 1.5 more items when these were real-world objects than when they were simple stimuli. In short, the finding that individual capacity for objects and simple stimuli may differ has been well replicated (and we found similar differences in our own data with measures based on discrete-slot and signal detection models; see Appendix A). This finding poses a conceptual problem for pure discrete-slot models of visual short-term memory.

More specifically, it is unclear how the notion of a *fixed object capacity* can be reconciled with the finding that individuals' estimates of capacity vary as a function of the informational content of the to-be-remembered items. One possibility is that there is a fixed item capacity but that it is fundamentally different for meaningful and meaningless stimuli. This explanation has the unique quality of denying the very quality it presumes—if capacity differs depending on the material, can it really be considered capacity at all? Certainly, capacity must have some domain-generality to make sense.

We considered another possibility, which is that item capacity is fixed but that the resolution with which to-be-remembered items are coded and maintained varies as a function of task demands, which depend on the type of information that is remembered

(meaningful or meaningless) and the manner in which it is presented.[2] For instance, it is possible that VSTM consists of a fixed number of slots but that top-down feedback based on conceptual knowledge in long-term memory can boost the fidelity of VSTM representations (and thereby inflate estimates of K). To test this prediction, we developed and evaluated a model that we will refer to as the *across-task fixed capacity + resource* model. This model follows from the previously discussed mixture model of capacity. The across-task model presented here makes the theoretical prediction that item capacity (i.e., maximum number of 'slots' in which memory representations are stored) is fixed across tasks, but the ability to bolster the resolution of representations for each type of stimulus (for instance, by way of attentional processing) differs depending on the task demands (e.g., the characteristics of to-be-remembered stimuli and the way in which they are presented). To our knowledge, this variant of the mixture model has not been formally tested in prior research. We evaluated the ability of this model (as well as the standard discrete and continuous models described above) to capture performance across the change detection and RSVP tasks in Experiments 2 and 3.

1.2.6. Model comparison

The second way in which we evaluated model generalizability was by testing the predictive power of each model. To this end, rather than guiding model selection solely by model fit, we also applied within-sample k-fold cross-validation analysis, in which part of the data are used to estimate best fitting parameters for a given model and these best fitting parameters are then used to predict the remaining part of the data (for a related approach to model comparison see Donkin et al., 2014). Cross-validation analysis is a commonly used remedy in statistics and machine learning to the problem that overly flexible models may overfit data in a sample – or, rather, that they fit error variance instead of variance that is generated by the underlying process of interest (Holbert & Stephenson, 2002; Roberts & Pashler, 2000). Standard approaches to model selection, such as the use of fit statistics that penalize a model based on its number of parameters, are inadequate because model complexity is not exclusively determined by the number of parameters, but also by how these parameters combine in the model equation (e.g., Pitt & Myung, 2002). Therefore, in the context of selecting nonlinear models that are highly similar, such as the UVSD model and the mixture model, it is desirable to guide model selection based on how well models predict unseen data, rather than by how well a model fits a single sample of data. Overly flexible models that fit error variance within a subsample of the data typically perform poorly in cross-validation analysis in which these parameters are used to predict a different subset of the data (Browne, 2000).

Cross-validation also has limitations, however (e.g., Lei, 2017). Most notably, the application of cross-validation with small samples may still lead to overfitting (Varoquaux, 2018). In addition, separating the data set into components dedicated to training and others to evaluation mean that both components do not enjoy the full power of the experimental design. In the current set of studies, we evaluated models with both measures of fit and with predictive capacity. Our choice of specific indices was based on model recovery simulations that compared how well each measure recovered models that generated the data (Jang, Wixted, & Hubner, 2009; van den Berg, Awh, Ma, 2014). These simulations are discussed below and in more detail in Appendix B. We found that Akaike Information Criteria (AIC) and cross-validation performed best at recovering models that generated the underlying set of data. Cross-validation and AIC performed best at discriminating between the broad classes of slot and pure resource models, and AIC did a better job of discriminating specific instances of models within each class. As will become evident upon presentation of these statistics, the indices agree in suggesting that continuous-resource, and not discrete, models best account for our data.

2. Experiment 1

2.1. Materials and methods

2.1.1. Experiment 1a: change detection task with simple stimuli

Experiment 1a was designed to test models of VSTM capacity in a standard change detection paradigm. As mentioned, our design and analysis addressed several limitations of prior work. First, rather than inducing bias experimentally we plotted levels of response bias on the ROC curve by collecting participants' confidence ratings. This approach circumvents problems with applying base rate manipulations with naïve subjects, such as the possibility that participants are not sensitive to all levels of the experimentally induced bias. Furthermore, we analyzed individual participants' data rather than aggregated data to avoid potential averaging artifacts at the level of participants. We tested a total of six[3] mainstream capacity/resource models: fixed discrete-slot (DS-F); variable discrete-slot (DS-V); alternative discrete-slot (DS-A); mixture; and equal (EVSD) and unequal variance (UVSD) signal detection models.

2.1.1.1. Participants. Sample size in this and all other experiments was based on the sample size used by Rouder et al. (2008). Those authors collected data from 23 participants. To be conservative, we collected data from 30 participants from the University of Illinois community in exchange for course credit or $8. All participants reported normal or corrected-to-normal vision. The study was

[2] It is also possible that differences in the comparison process when memory is probed may lead to such differences in capacity estimates (e.g., Awh et al., 2007). For instance, Awh et al. reported that an instructional manipulation—whether subjects are required to make fine-grained versus coarse discriminations when memory is probed—affects capacity estimates. The authors interpreted this effect as indicative of a cost of comparison errors for difficult discriminations (e.g., two shades of blue), rather than a true difference in how many items are maintained in VSTM. We used colors in the VSTM task that are categorically different, making this interpretation difficult to apply to our data.

[3] We did not focus on a variant of the pure discrete-slot model developed by Rouder et al. (2008) because its parameters may not be recoverable in popular versions of the change detection task (Feuerstahler, Luck, MacDonald III, & Waller, 2019).

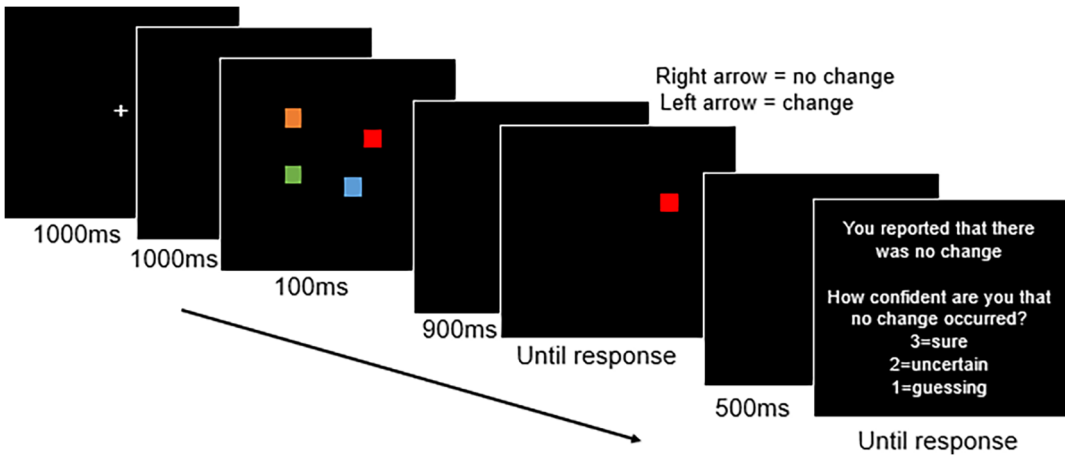

Fig. 3. Example sequence of events for a no-change trial with four to-be-remembered items (Experiment 1a).

approved by the university's Institutional Review Board.

2.1.1.2. Stimuli and procedure. The experimental program was written in Matlab, using Psychophysics Toolbox extensions (Brainard, 1997). Stimuli were presented on a Samsung 2013 LED monitor with screen resolution of 1920 by 1200 pixels and a refresh rate of 60 Hz. Stimuli were presented on a black background.

The procedure is shown in Fig. 3. At the beginning of each trial a white fixation cross (1° by 1°) was presented for 1000 ms and was followed by a blank interval for 1,000 ms. Next, participants were shown a memory array that contained 4, 6 or 8 colored squares for 100 ms. We chose 4 as the minimum set size in this and all of our other experiments because these were the set sizes used by Rouder et al. and Donkin et al. Furthermore, previous research that has used measures of VSTM based on discrete-slot models reported that average VSTM capacity is limited to 3–4 items (Luck & Vogel, 2013). Accordingly, using a smaller set size is likely to yield extremely high levels of performance, which may lead to a clustering of ROC points, making it difficult to evaluate the contending models (Pazzaglia, Dube, & Rotello, 2013). Finally, discrete-slot measures of VSTM are valid only under conditions when the set size is equal to or exceeds individuals' capacity (e.g., Morey, 2011).

Following standard procedures for this task (see Luck & Vogel, 1997), each square (size: 0.75° by 0.75°) was shown in a pseudo-random location within a 9.8° by 7.3° visual angle view with the constraint that every square (center to center) was at least 2° from every other square and no square was located within 2° of the center. Each square in the memory array was of a different color. There were 8 possible colors; white (255, 255, 255), orange (255, 128, 0), grey (96, 96, 96), purple (128, 0, 128), blue (0, 0, 255), red (255, 0, 0), yellow (255, 255, 0), and green (0, 128, 0). The memory array was followed by a blank screen retention interval (900 ms). Following the retention interval, participants were shown a test stimulus, which was a square that appeared in the same location as one of the squares in the memory array. On half of the trials (change trials), the square was a different color than the square presented in that location in the memory array. On such change trials, the square always was the color of a different square that appeared in the original memory array, thus forcing participants to remember the location of each colored square. On the other half of the trials (no change trials), the square was the same color as the square presented in that location in the memory array. Participants were instructed to report with a button press, as accurately as possible, whether the square changed color (left arrow key) or did not (right arrow key). After their response, the memory array offset. A brief interval (500 ms) after the memory array offset, participants were shown a confidence rating self-report screen. Participants were instructed to respond using a three-point confidence scale how confident they were in their change or no change response (i.e., 1 = *guessing*, 2 = *uncertain*, 3 = *sure*). Participants entered responses using the number keypad and pressed the Enter key to advance to the next trial. There were 6 practice trials in the beginning of the experiment and 360 experimental trials with an equal number of observations (60) in each cell. Participants were given the opportunity to take an optional break every 72 trials. Each experimental session lasted approximately one hour.

2.1.2. Experiment 1b: change detection task with simple stimuli and masking

Experiment 1b was identical to Experiment 1a with the exception that a visual mask was presented between the presentation of the memory and test arrays. The mask was designed to eliminate any potential contributions of sensory or iconic memory to visual short-term memory performance (Cowan, 2001), and closely replicates a design used by Rouder et al. (2008).

2.1.2.1. Participants. Thirty new individuals from the University of Illinois community participated in exchange for course credit or $8. All participants reported normal or corrected-to-normal vision.

2.1.2.2. Stimuli and procedure. The stimuli and procedure in Experiment 1b (see Fig. 4) were identical to those used in Experiment 1a except for the following changes. After the presentation of a fixation cross and a blank interval, the memory array was presented for 500 ms. Stimuli in the memory array (colored squares) were presented in the same way as in Experiment 1a. After a 500 ms blank

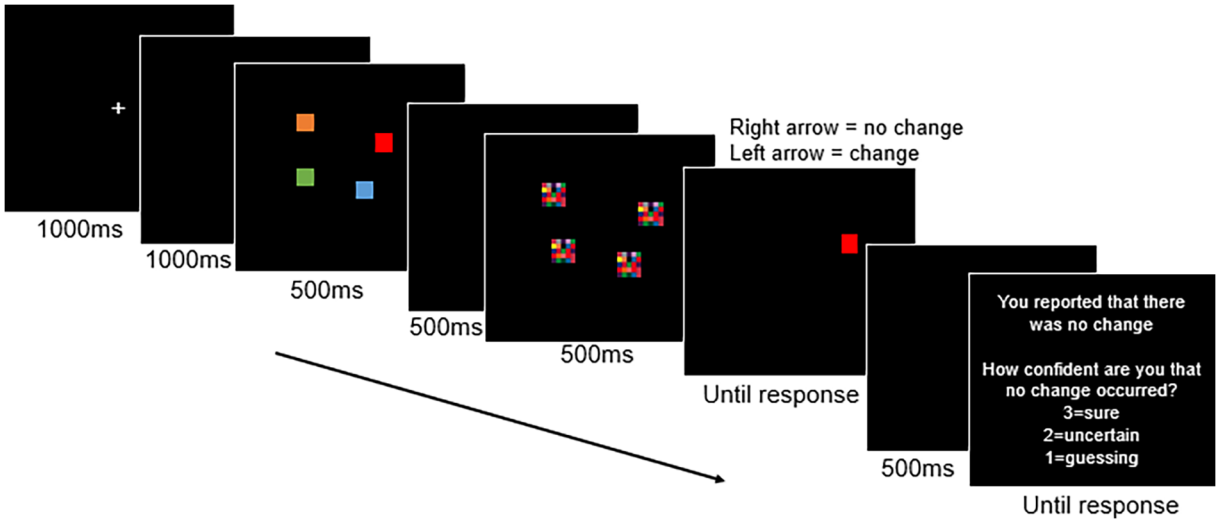

Fig. 4. Example sequence of events on a no-change trial with four to-be-remembered items (Experiment 1b).

interval, masks were presented in the locations of the presented squares for 500 ms. Mask stimuli consisted of a small (1.25° by 1.25°) 5 by 5 matrix with squares of different colors (each color was chosen randomly out of the 8 colors used in the memory array).

2.1.3. Experiment 1c: rapid serial presentation stream task with images of real-world objects

In Experiment 1c we tested models of VSTM capacity using a task that differs from the change detection task in both presentation format and the type of stimuli used. Specifically, participants were presented with stimuli sequentially, rather than simultaneously, and stimuli were pictures of real-world objects rather than simple features bound to spatial locations. We chose this task precisely because it is meaningfully different from the task used in Experiments 1a and 1b but is also used in the literature to measure VSTM capacity (e.g., Berry et al., 2018; Endress & Potter, 2014, Gorgoraptis et al., 2011; Nosofsky & Donkin, 2016). Experiment 1c allowed us to examine whether the modelling results found using the change detection task replicate under substantially different conditions.

2.1.3.1. Participants. Thirty new individuals from the University of Illinois community participated in exchange for course credit or $8. All participants reported normal or corrected-to-normal vision.

2.1.3.2. Stimuli and procedures. The procedure used in Experiment 1c followed the procedure for the repeated condition used by Endress and Potter (2014). A schematic of the procedure is shown in Fig. 5. Stimuli were presented on a white background. Each trial began with the presentation of a black fixation cross (500 ms). Following a blank interval (500 ms), participants were shown a sequence of rapidly presented photos of objects. Each image was presented for 250 ms with no interval between the images. Images

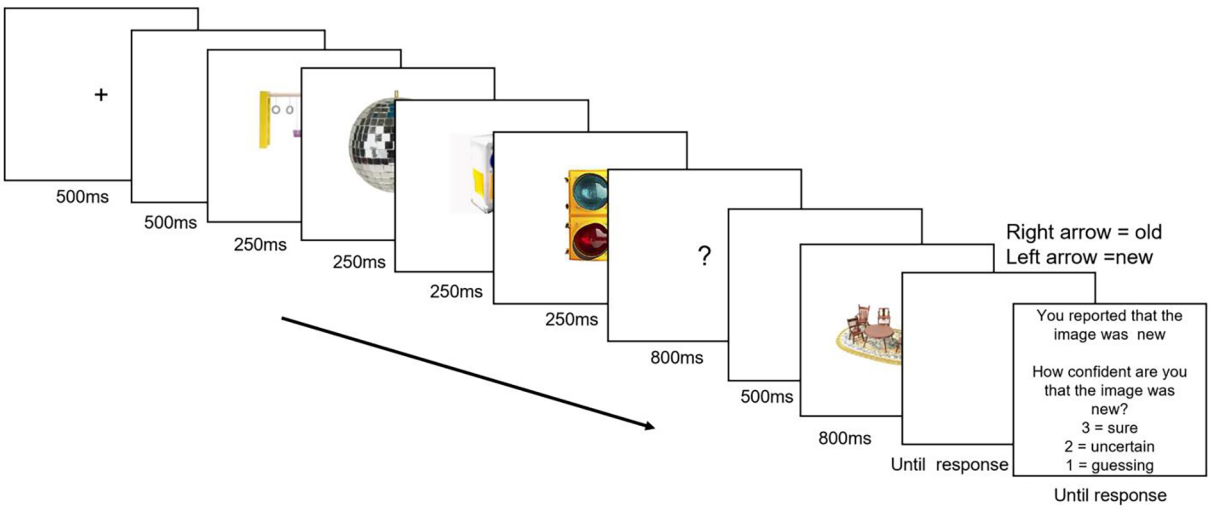

Fig. 5. Example sequence of events on a trial where a new item was probed and there were four to-be-remembered items (Experiment 1c).

were color photographs of real-world objects used by Brady, Konkle, Alvarez, and Oliva (2008). Each participant was presented with a fixed set of 13 randomly selected images throughout the experimental session. Four, six or eight images were shown in a sequence on a given trial. Following the image sequence, a question mark was shown for 800 ms followed by a blank screen for 500 ms. Next, a probe image was shown for 800 ms, followed by a blank screen that remained until participants made a response. Participants were instructed to respond "old" if the probe image was an image that had appeared in the sequence of images shown on that trial, and to respond "new" if the probe image was an image that had not appeared in the sequence of images shown on the trial. On lure trials, an image was randomly selected from the 13 images that were not shown in the sequence on that trial. On trials when the image had been shown in the memory sequence, one of the images was selected at random from the sequence of images shown on that trial. 50% of the trials contained a target, and 50% a lure. After providing an old/new judgment, participants were asked to report their confidence that the test image was new or old. The confidence scale was identical to that used in all other experiments.

2.2. Results

The raw data are included in the supplemental materials for this and all experiments reported in this paper. Appendix A summarizes the results of transformations of the data from each experiment based on the discrete-slot (K capacity estimates) and equal-variance (sensitivity (d') and response bias (C)) signal detection model as well as a measure of performance based on the unequal variance model (d_a). Here we focus on an evaluation of the competing models. Model evaluation used AIC, an information-theoretic measure of model fit, and sum of squared error of prediction (SSEP), a measure based on cross-validation analysis that reflects the predictive ability of each model. Model curves were fit using the Maximum Likelihood Estimation procedure in Matlab and the ROC toolbox (Koen, Barrett, Harlow, & Yonelinas, 2016). Code was modified to fit each type of capacity model reported here. Optimization was performed using the Matlab `fmincon` optimization algorithm. The maximum number of iterations for the optimization routine was set to 1,000,000. We fit models to the entire set of data to obtain AIC values.

For the cross-validation analysis, the outcome measure in our experiments is the residual error obtained for each model fitted via cross-validation to an out-of-sample dataset (SSEP). For each trial, the participant's change/no-change response and confidence rating were converted to a value of 1–6 (i.e., 1 = no-change/sure, 2 = no-change/uncertain, 3 = no-change/guessing, 4 = change/guessing, 5 = change/uncertain, 6 = change/sure). To implement the 5-fold cross-validation analysis, a vector was created that contained the frequency of responses of each type. For instance, if a participant had 15 observations in the 1st confidence bin (i.e., no-change/sure), 17 observations in the 2nd confidence bin, etc., this would generate a vector of fifteen 1 s, seventeen 2 s, and etc. A random 4/5 of the data in this vector were selected and used to construct an empirical ROC. This function was used as the training set, to which model curves were fit using the maximum likelihood estimation. This model fitting was done separately for each condition (set size) and participant. The remaining 1/5 of the data were used as the test set, and best-fitting parameters from the training set were used to predict points on the ROC function based on the data in the test set. This process was repeated five times, using each fold as the test set, and further repeated 20 times with 20 different random splits of the data. Model selection was guided by comparing the average sum of squared errors of prediction across the 100 cross-validation iterations of each model. The average residual error of prediction provides an index of the predictive ability of a model (e.g., Picard & Cook, 1983). Note that smaller values of sum of squared errors of prediction (SSEP) indicate that a given model captured more variance of the underlying process of interest rather than error variance in the training set.

We chose these two measures because our model recovery simulations (Appendix B) revealed that they performed best in recovering the true data-generating models, given the number of observations used in these data and the best-fitting parameters of each model that best described the data. We used a permutation test designed for within subject comparisons with 1,000,000 iterations to test differences in AIC and SSEP across participants between each model. Fig. 6 shows empirical ROCs and model fits of individual and aggregated data from participants for which a given model was the best fitting model. For each model comparison, we report the p-value of the permutation test. Table 2 summarizes average AIC and SSEP values for each experiment and model. Fig. 7 shows heatmaps for Δ AIC. By definition, the best fitting model will have Δ AIC = 0, whereas the other models will have positive values of Δ AIC. We interpret Δ AIC in the standard way (Burnham & Anderson, 2004), with Δ AIC ≤ 2 indicating that the best-fitting model and a competing model perform comparably well, 4 ≤ Δ AIC ≤ 7 indicating weak support for the best-fitting model over a competing model, and Δ AIC > 10 indicating functionally no support for the competing model relative to the best-fitting model. In cases where two models have substantial support, the more parsimonious model, i.e., the model with fewer free parameters, is favored.

We first examined which model had the lowest average AIC and SSEP, separately for each experiment[4]. Average AIC was lowest for the EVSD model in Experiment 1a and 1b and lowest for the UVSD model in Experiment 1c. In all cases, Δ AIC values indicated substantial support only for pure resource models. Permutation tests revealed that AIC was lower for the EVSD model than the standard DS-F model ($p < .001$; 24 and 25 of 30 participants in Experiments 1a and 1b, respectively), the DS-V model ($p < .001$; 26 and 24 of 30 participants, respectively), the DS-A model ($p = 0.047$ and $p = 0.014$ for 21 and 18 of 30 participants, respectively), the mixture model ($p = 0.005$ and $p = 0.02$ for 22 and 21 of 30 participants, respectively) and the UVSD model ($p = .03$ and $p < 0.001$ for 22 and 23 of 30 participants, respectively). In Experiment 1c, AIC was lower for the UVSD model than the DS-F model ($p < 0.001$; 24 of 30 participants), the DS-V model ($p < 0.001$; 23 of 30 participants), the DS-A model ($p < 0.001$; 25 of 30 participants), and the mixture model ($p < .001$; 21 of 30 participants). AIC did not differ significantly between the EVSD and UVSD model ($p = .49$; 13 of

[4] We did not make any assumptions about how response bias should vary as a function of set size, and so allowed response criteria to vary freely across set sizes. In Appendix C we also report results for each model in which criteria were fixed across set sizes.

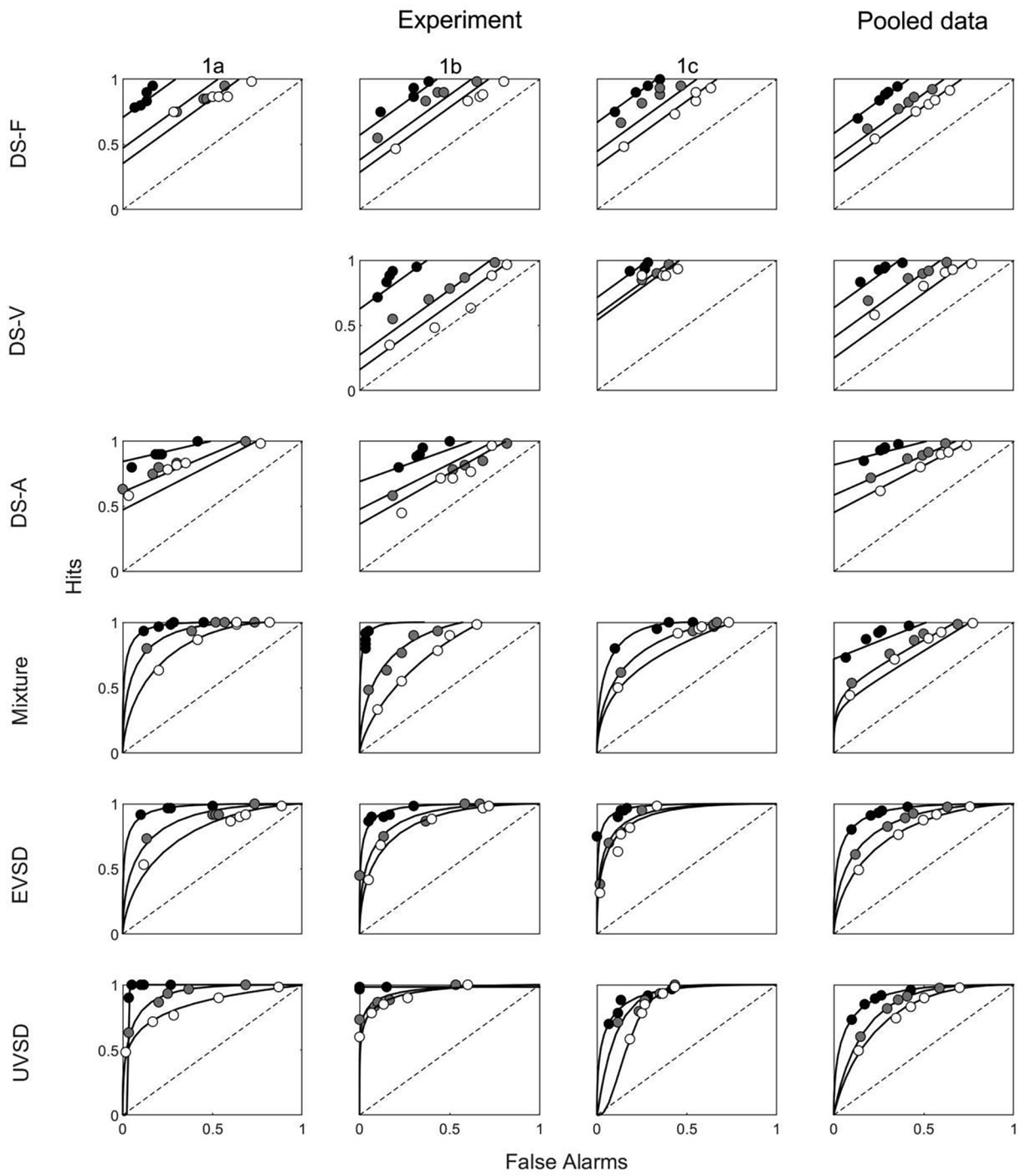

Fig. 6. The first three columns show ROC plots for individual participants. We show data for those participants for which a given model was the best performing model. The fourth column (pooled data) shows data aggregated across participants and across experiments for whom the given model was the best performing model. Note that we did not fit models to aggregate data in our analyses.

30 participants) in this experiment.

We repeated these permutation tests using SSEP as the outcome measure. SSEP was consistently lowest for the UVSD model, which outperformed all other models in all three experiments. Average SSEP was lower for the UVSD model than the standard DS-F model ($p < .001$; all 30 participants in Experiments 1a-c), the DS-V model ($p < .001$; all 30 participants in Experiments 1a, and 29 of 30 participants in Experiments 1b-c), the DS-A model ($p < .001$; all 30 participants in Experiments 1a-c), the mixture model ($p < .001$; 26 of 30 participants in Experiment 1a, 21 of 30 participants in Experiment 1b and 28 of 30 participants in Experiment 1c)

Table 2
Average AIC and SSEP vales for each model in Experiment 1a-c. Underlined values in bold denote best performing models.

		Model and number of parameters					
Experiment	Measure of model performance	DS-F (16)	DS-V (18)	DS-A (16)	Mix (20)	EVSD (18)	UVSD (21)
1a	AIC	884	885	871	871	**863**	865
	SSEP	0.1840	0.1608	0.1677	0.0866	0.0741	**0.0678**
1b	AIC	915	914	902	899	**895**	897
	SSEP	0.1803	0.1533	0.1940	0.0850	0.0758	**0.0719**
1c	AIC	932	933	941	928	**923**	**922**
	SSEP	0.148	0.1254	0.2463	0.0866	0.0761	**0.0679**

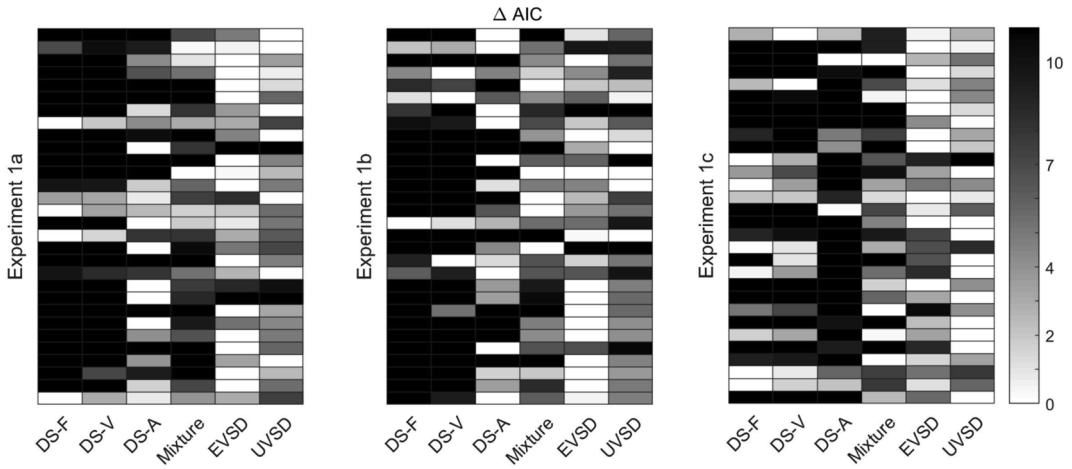

Fig. 7. Heatmaps of Δ AIC (the difference between AIC of each model and the minimum AIC value of the best performing model) for Experiment 1a-c. Rows denote individual participants (30 total in each experiment).

and the EVSD model ($p < .001$; 26 of 30 participants in Experiment 1a and 22 of 30 participants in Experiments 1b-c).

2.3. Discussion

Both model fitting and cross-validation results revealed that pure resource models consistently outperformed discrete-slot models as well as the hybrid model. These results were impressively consistent across participants. Thus, we failed to replicate the findings of Rouder et al. and Donkin et al. that discrete-slot models outperform signal detection models. In addition to standard discrete-slot models, we tested a mixture model (Xie & Zhang, 2017), which is a hybrid of the signal detection and discrete-slot models. In our analyses we fixed the item capacity parameters across set sizes and allowed the resolution parameter to vary freely across set sizes. This allowed us to test a central prediction of this model, namely that there is a fixed item capacity limit and an independent resource that can lead to variability in item encoding/maintenance that varies as a function of memory load. This model also performed more poorly than pure resource models, however. Our results suggest that pure resource models provide the best description of how information is represented in VSTM, at least out of the set of candidate models we considered in this analysis. We found evidence for this in the change detection task under conditions where to-be-remembered items were and were not masked after presentation, indicating that these results are not due to any contributions of perceptual or iconic memory. Furthermore, we found evidence for the pure resource models under very different experimental conditions, in which the to-be-remembered items were photos of real-world objects and were presented sequentially instead of simultaneously.

3. Experiment 2

In the first set of experiments we tested models of capacity within different VSTM tasks. Our results provided support for continuous-resource over discrete-slot and mixture models of capacity in both tasks. The aim of Experiment 2 was to extend these findings by evaluating the ability of these models to jointly account for performance in the two different tasks. To our knowledge, this is the first time that these models have been evaluated using more than one task, despite the general agreement that what is needed is a model of VSTM, not a model of one specific task or another. This approach also allows us to test varying levels of constraint across

the fits of an individual model to the two tasks.

We had a new group of participants complete both the change detection and RSVP tasks in separate experimental sessions and we fit different variants of the discrete-slot, mixture and signal detection models to their data from both tasks. Specifically, we first fit a variant of each model in which capacity and response criteria were fixed across tasks (fully joint model). These models make the strong prediction that the same capacity or resource (K or d') and response bias (g or C) estimates can account for performance across the two tasks. We also fit a variant of each model in which capacity was fixed across tasks, but response bias was free to vary (joint capacity/disjoint criteria model). These models make the prediction that both tasks place similar demands on VSTM capacity processing, but that participants sometimes apply different response policies in the two tasks. As part of this analysis, we also tested the across-task capacity + resource model, which predicts that both tasks place demands on a single underlying item capacity but may place different demands on an independent attention-like resource that determines the resolution of item representations. In addition, this model assumes that participants may place different response criteria as a function of memory load and task. Finally, we fit a variant of each model where both capacity and response bias were free to vary across the two tasks (fully disjoint model). These variants assume that the two tasks place different demands on VSTM, both at the level of how capacity or resources are allotted, and how response bias is set. Note that this set of analyses formally tests the strong prediction that follows from discrete-slot models, which is that the limits of visual short-term memory are set in terms of objects, independently of object complexity and their informational content.

3.1. Materials and methods

3.1.1. Participants

We collected data until our sample size reached $n = 30$. In total, we collected data from 36 participants. Six participants did not show up for the second session of the study; these participants were replaced, and their data were not analyzed. All participants were from the University of Illinois community and participated in exchange for course credit or $8. All participants reported normal or corrected-to-normal vision.

3.1.2. Stimuli and Procedure

In this experiment participants completed both the change detection and the RSVP task. The tasks were completed in separate sessions on different days. The average duration between sessions was approximately 24 h. The stimuli and procedure in the change detection task were identical to those used in Experiment 1a and the stimuli and procedure in the RSVP task were identical to those used in Experiment 1c. The order of tasks was counterbalanced across participants.

3.2. Results

Table 3 shows average AIC and SSEP values for each model. Fig. 8 shows Δ AIC values. We began by finding the model that had the lowest average AIC and SSEP. The fully disjoint EVSD model had the smallest average AIC. We found that the fully disjoint EVSD model outperformed all three variants of the DS-F model (maximum $p < .001$; minimum 24 of 30 participants) and all three variants of the DS-V model ($p < .001$; minimum of 25 of 30 participants). Although average AIC was quantitatively lower for the fully disjoint EVSD model than for all variants of the mixture model we found that the difference between the fully disjoint EVSD and the fully disjoint mixture model was not significant with a permutation test ($p = 0.09$). We used a Bayes Factor for a paired t-test to quantify evidence for the null (Rouder, Speckman, Sun, Morey, & Iverson, 2009). This test revealed weak evidence for the null ($BF_{NULL} = 1.74$; Cauchy prior on the effect size [scaling factor $r = 1.0$] and a non-informative Jeffreys prior on variance). Based on the principle of parsimony these modeling results favor the EVSD model because it has two fewer parameters than the mixture model. AIC was significantly lower for the fully disjoint EVSD model than for the across-task resource + capacity model ($p < .0015$; 23 of 30

Table 3
Average AIC and SSEP values for each model in Experiment 2.

		VSTM model					
Across task model		DS-F	DS-V	AC-Mix	Mix	EVSD	UVSD
Fully joint	# pars	16	18		20	18	21
	AIC	1907	1908	N/A	1887	1881	1883
	SSEP	0.543	0.471		0.378	0.372	0.367
Joint capacity	# pars	31	33	38	35	33	36
	AIC	1890	1891	1861	1867	1856	1858
	SSEP	0.375	0.367	0.250	0.210	0.212	0.209
Fully disjoint	# pars	32	36		40	36	42
	AIC	1883	1884	N/A	1853	**1851**	1854
	SSEP	0.332	0.316		0.193	0.183	**0.172**

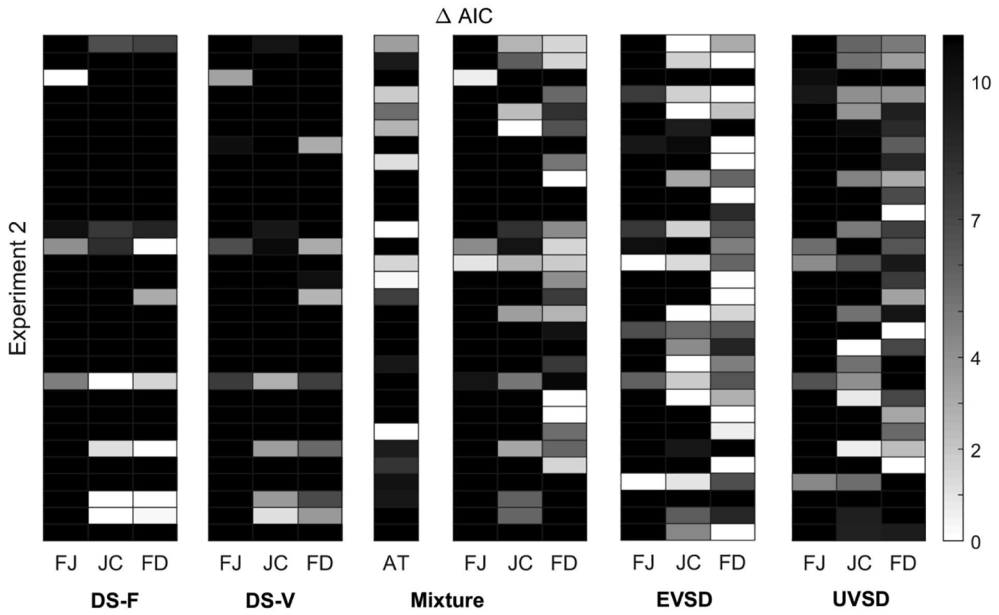

Fig. 8. Heatmap of Δ AIC values for each model in Experiment 2. Rows denote individual participants ($n = 30$).

participants). Finally, AIC for the fully disjoint EVSD model was significantly lower than AIC for other variants of the EVSD model (maximum $p = 0.04$; minimum 13 of 30 participants), and significantly lower than for all variants of the UVSD model (maximum $p = 0.005$; minimum of 23 of 30 participants).

We repeated this analysis using SSEP as the outcome measure. The model with the lowest average SSEP was the disjoint UVSD model. We found that the fully disjoint UVSD model outperformed the fully disjoint DS-F model ($p < .001$; 29 of 30 participants), the fully disjoint DS-V model ($p < .001$; 27 of 30 participants), the fully disjoint mixture model ($p < .001$; 23 of 30 participants), the across-task capacity + resource model ($p < .001$; 26 of 30 participants) and the fully disjoint EVSD model ($p < .001$; 26 of 30 participants). Permutation tests for repeated measures indicated that the fully disjoint model outperformed both the joint capacity/disjoint criteria and the fully joint model for all variants of the discrete-slot and signal detection models (all $p < .001$). In addition, the joint capacity/disjoint criteria model always outperformed the fully joint model (all $p < .001$).

3.3. Discussion

The modeling results of Experiment 2 generally replicated the results of Experiments 1a, 1b, and 1c. Pure resource models outperformed all other models, including the mainstream discrete-slot models of capacity. Importantly, we found that the fully disjoint EVSD model outperformed the across-task capacity + resource model, according to which capacity is fixed across tasks but the resolution with which items are represented varies as a function of task demands. We did find that AIC across participants was not significantly different between the fully disjoint mixture and the fully disjoint EVSD model. Given that the EVSD model has overall lower AIC and is more parsimonious than the mixture model (because it assumes a single resource rather than a capacity and resource), we interpret these results as providing converging support for pure resource models of VSTM. The cross-validation results provide converging support for this interpretation because they also favored pure resource models of VSTM. In addition, we found that the fully disjoint EVSD model, in which resource and response criteria parameters were allowed to vary freely across tasks, outperformed both the fully joint variant of this model, in which all parameters were fixed across tasks, and the joint capacity/disjoint criteria variant of this model, in which the resource parameter was fixed across tasks, but response criteria were free to vary. Although the fully disjoint UVSD model outperformed the fully disjoint EVSD model in the cross-validation analysis, our model recovery simulations (Appendix B) showed that cross-validation is somewhat biased towards more flexible models (note that the fully disjoint UVSD model has 6 more parameters than the fully disjoint EVSD model) and favored the UVSD model over the EVSD model even when simulated data were generated from an EVSD model. In either case, these results provide further support for pure resource models over discrete-slot models of VSTM and further indicate that resources are continuously distributed across items but that the distribution varies depending on the demands of the VSTM task. We elaborate on this interpretation in the General Discussion.

4. Experiment 3

Modelling results from Experiment 2 provided support for a fully disjoint resource model of performance for the change detection and RSVP tasks. It is unclear, however, whether these results were obtained because the two VSTM tasks place different demands on VSTM, or instead were due to a more general effect of completing VSTM tasks multiple times (e.g., via practice or fatigue effects).

Previous work indicates that VSTM performance can change with practice (Xu, Adam, Fang, & Vogel, 2018), so it is possible that we would obtain a similar pattern of results if participants completed multiple sessions of the same VSTM task. To examine this possibility, we had two separate groups of participants perform the same VSTM task (either RSVP or change detection) in two separate sessions and tested the fully disjoint, joint capacity/disjoint criteria, and fully joint models from Experiment 2. Because participants performed the same task across sessions we would expect the joint models to fit the data better in this experiment than in Experiment 2, in which participants performed different tasks across sessions. If the disjoint model is favored to the same extent as is apparent in Experiment 2, even when fitted to two tasks of the same type, then this would suggest that practice and/or fatigue were responsible for the results of Experiment 2. In contrast, if the difference between the fully joint model and the joint capacity/disjoint criteria model is lower in this experiment than in Experiment 2, then this would support the conclusion that the results of Experiment 2 were not due merely to practice or fatigue from participating in multiple experimental sessions.

4.1. Materials and methods

4.1.1. Participants

We collected data until our sample size reached $n = 60$. In total, we collected data from 63 participants. Three participants did not show up for the second session of the study and their data were not analyzed. Participants were from the University of Illinois community and participated in exchange for course credit or $8. All participants reported normal or corrected-to-normal vision.

4.1.2. Stimuli and procedure

Participants were assigned to one of two groups, the change detection group or the RSVP group. The *change detection group* completed the change detection task in two separate sessions on two separate days. The stimuli and procedure were identical to those used in Experiment 1a. The *RSVP group* completed the RSVP task in two separate sessions on two separate days. The stimuli and procedure were identical to those used in Experiment 2.

4.2. Results

4.2.1. Two session change detection task.

Table 4 summarizes average AIC and SSEP values for each model. Fig. 9 shows Δ AIC values. Average AIC was again smallest for the fully joint EVSD model. A permutation test for repeated measures revealed that this version of the model outperformed all three variants of the DS-F model (maximum $p < .001$; minimum 24 of 30 participants) and all three variants of the DS-V model (maximum $p < .001$; minimum 24 of 30 participants). AIC was also significantly lower for the fully joint EVSD model than the fully joint and joint capacity versions of the mixture model (maximum $p = 0.02$; minimum 22 of 30 participants). It was not statistically different from the AIC of the fully disjoint mixture model ($p = 0.61$). Furthermore, there was no statistical difference in AIC between the fully joint EVSD model and the joint capacity EVSD ($p = .78$) and fully disjoint EVSD model ($p = .54$). Finally, AIC was significantly lower for the fully joint EVSD model than the fully joint UVSD model ($p = 0.003$, minimum 22 of 30 participants), but not than the joint capacity UVSD ($p = 0.40$) or the fully disjoint UVSD model ($p = 0.08$). Together, average AIC was overall lowest for the fully joint EVSD model than all other models, but it did not convincingly outperform all other individual models in head-to-head comparison.

Table 4
Average AIC and SSEP values for all models and the two tasks in Experiment 3.

			VSTM model				
Task	Across task model		DS-F	DS-V	Mix	EVSD	UVSD
Change detection	Fully joint	# pars	16	18	20	18	21
		AIC	1744	1743	1716	**1705**	1707
		SSEP	0.42	0.39	0.27	0.24	0.24
	Joint capacity	# pars	31	33	35	33	36
		AIC	1744	1743	1719	1706	1708
		SSEP	0.35	0.33	0.22	0.17	0.17
	Fully disjoint	# pars	32	36	40	36	42
		AIC	1743	1745	1707	1707	1712
		SSEP	0.33	0.31	0.19	0.17	**0.16**
RSVP	Fully joint	# pars	16	18	20	18	21
		AIC	1849	1849	1814	1791	**1782**
		SSEP	0.56	0.54	0.28	0.25	0.24
	Joint capacity	# pars	31	33	35	33	36
		AIC	1853	1853	1818	1794	1797
		SSEP	0.52	0.50	0.24	0.25	0.18
	Fully disjoint	# pars	32	36	40	36	42
		AIC	1851	1814	1806	1795	1803
		SSEP	0.51	0.48	0.23	0.19	**0.17**

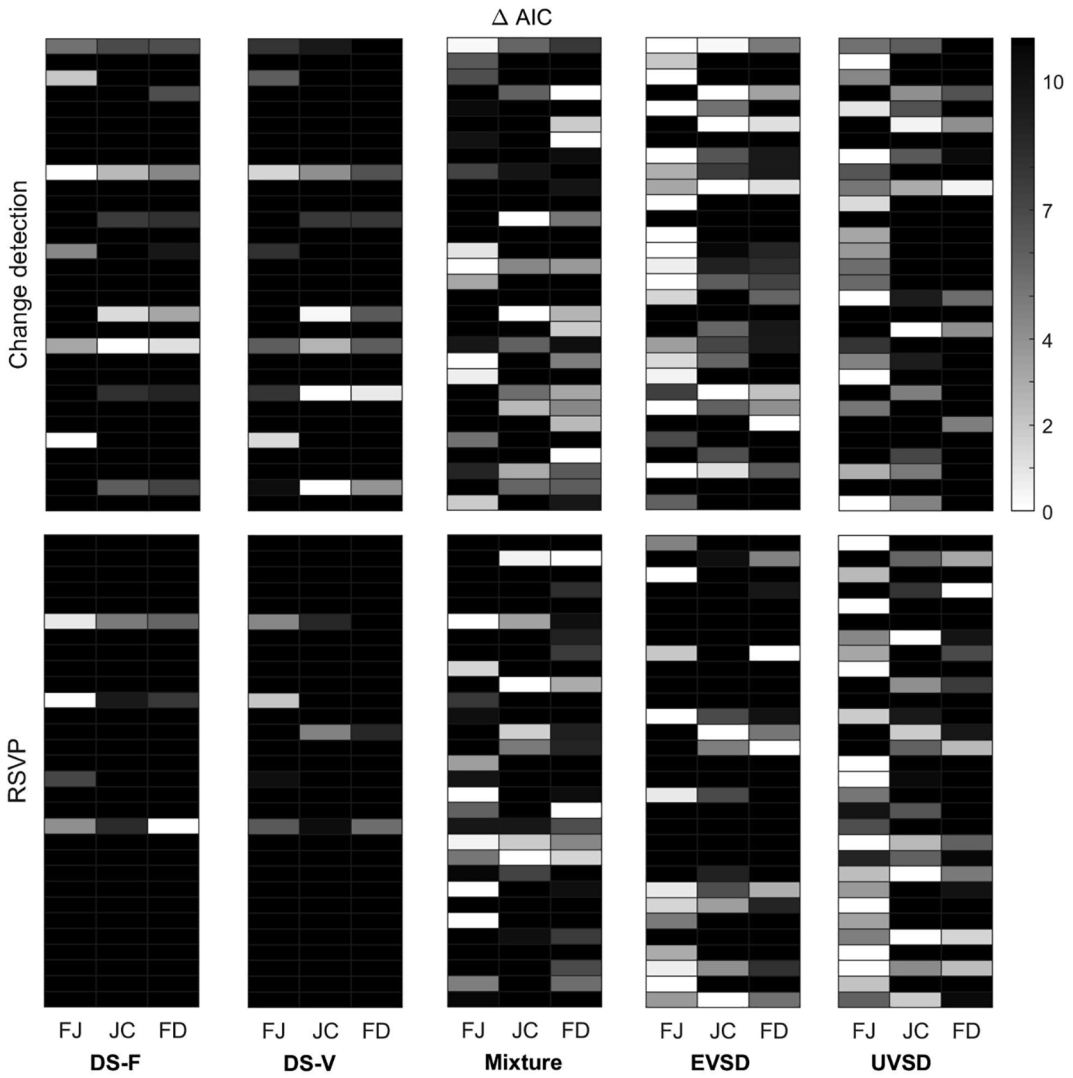

Fig. 9. Heatmaps of ΔAIC for all models in the two tasks of Experiment 3. Rows denote individual participants (n = 30).

Comparisons of AIC based on the permutation tests for repeated measures reveal that the fully joint EVSD model outperformed all other fully joint models, but performed comparably to more flexible models that had a signal detection component. Taking all of these results into account, analysis of AIC scores favors the fully joint EVSD model, which performs at least as well as all other models with a signal detection component but with fewer parameters.

If fatigue or task familiarity was responsible for the finding in Experiment 2 that the fully disjoint model outperformed the joint capacity model and fully joint models, we would expect to find that the fully disjoint model would outperform the joint capacity or fully joint model to the same degree when participants performed two sessions of the same task. The AIC comparisons indicate that the results from Experiment 2 are not merely due to effects of completing a VSTM task in two different experimental sessions.

The same comparisons based on SSEP revealed that the fully disjoint variant of the UVSD model outperformed all models (all $p < .01$; results found for at least 19 of 30 participants) *except* for the joint capacity/disjoint criteria UVSD model ($p > .1$). A Bayes Factor for a paired t-test with a Cauchy prior on the effect size (scaling factor $r = 1.0$) and a non-informative Jeffreys prior on variance revealed weak evidence for the null ($BF_{NULL} = 2.11$; Rouder et al., 2009). To examine whether these differences in SSEP models between experiments was significant, we compared average SSEP between the fully disjoint model and the joint capacity/disjoint criteria variants of the UVSD model in the change detection group of this experiment and the across-task change detection session of Experiment 2. The results of a permutation test for repeated measures revealed a significant difference between SSEP between Experiment 2 and the change detection group of Experiment 3 ($p < .001$; 26 of 30 participants). The difference in SSEP between the fully joint model and the joint capacity/disjoint criteria model was much higher in Experiment 2 (0.037) than in the change detection group of Experiment 3 (0.007). Again, this finding suggests that the results of Experiment 2 were not due merely to practice or fatigue from participating in multiple experimental sessions.

4.2.2. Two session RSVP task

Average AIC was smallest for the fully joint UVSD model. Permutation tests for repeated measures revealed that this model outperformed all three variants of the DS-F and DS-V models (maximum $p < .001$; minimum 25 of 30 participants). AIC for this model was also significantly lower than it was for the fully joint and joint mixture capacity models (maximum $p < 0.001$; minimum 21 of 30 participants); it was not statistically different from AIC in the fully disjoint mixture model ($p = 0.07$). The fully joint UVSD model outperformed all variants of the EVSD model (maximum $p = 0.002$; minimum 22 of 30 participants). It performed comparably to the joint capacity UVSD model ($p = 0.16$) and outperformed the fully disjoint UVSD model ($p = 0.03$; 24 of 30 participants). Together, AIC results favor the fully joint UVSD model. Overall AIC was quantitatively smaller for this model than any other model. As with the change detection task, there are cases in which this resource model did not significantly outperform other individual models with a signal detection component (viz., the fully disjoint mixture model and the joint capacity UVSD model.) However, the fully disjoint mixture model has 19 more parameters and the joint capacity UVSD model has 15 more parameters than the fully joint UVSD model. Despite these advantages, the UVSD model still yielded a lower AIC score.

Using SSEP as the outcome measure, we found that the fully disjoint UVSD model outperformed all other models. Once again, the critical comparison was of the difference in average SSEP between the fully disjoint and joint capacity/disjoint criteria UVSD models in Experiment 2 and the RSVP group of Experiment 3 ($p < .001$; 21 of 30 participants). A permutation test for repeated measures revealed a significant difference in SSEP between the two models in the two experiments ($p = .006$; 20 of 30 participants), with the difference in SSEP between the fully disjoint and joint capacity/disjoint criteria UVSD models being lower in the RSVP (0.015) group of Experiment 3 than in the group that completed two separate tasks in Experiment 2 (0.037).

4.3. Discussion

Experiment 3 was designed to evaluate whether the superior performance of the fully disjoint model (relative to the joint capacity/disjoint criteria model) in Experiment 2 was due to the RSVP and change detection tasks placing different demands on VSTM processing, or rather was a consequence of performing VSTM tasks over multiple sessions (e.g., practice or fatigue effects). As before, we found that pure resource models outperformed other capacity models. However, unlike Experiment 2, AIC results consistently favored fully joint variants of these models. That is, models in which capacity and response criteria are assumed to be shared across sessions provided a superior account when the same task when repeated. This finding enhances the interpretability of the across-the-board findings of the superiority of a disjoint model in Experiment 2.

The cross-validation results provided some converging support for these conclusions. The difference in SSEP between the joint capacity/disjoint criteria model and the fully disjoint model was much smaller in both the change detection and RSVP groups in Experiment 3 than in the group of participants who performed different VSTM tasks across the two sessions (Experiment 2). In other words, relative to the most flexible fully disjoint model, the joint capacity/disjoint criteria model generalized to subsets of data from two different sessions of the same task (Experiment 3) better than to data from two different tasks (Experiment 2). Notably, we did not find that the fully joint model performed best in cross-validation. However, our model recovery simulations suggest that cross-validation for these types of data may have particular difficulty in cases in which criteria are fixed across conditions. Together, our findings suggest that the results of Experiment 3 reveal true differences in how the change detection and RSVP tasks probe VSTM resources.

5. General discussion

The first aim of the set of studies reported here was to apply ROC analysis to formally evaluate candidate models of VSTM capacity while addressing several limitations of prior research. To this end, we analyzed data at the level of individual participants rather than at the aggregate level to avoid potential averaging artifacts. We ensured measurement over a wide range of response criteria by collecting participants' confidence ratings rather than by manipulating base rates (Donkin et al., 2014; Rouder et al., 2008). We used multiple tasks, both thought to tap VSTM, but very different in their implementation. Finally, we guided model selection using cross-validation as well as fit statistics to assess how each type of model predicts unseen data and how successfully each model explains variance within a given sample of data. Together, the results across 3 experiments (with 6 independent samples) revealed consistent support for pure resource models of VSTM. We found that these models outperformed the mainstream discrete-slot model, which posits that capacity in VSTM is set in terms of objects (rather than features) that are stored with complete fidelity or not at all, as well as a mixture model, which also posits that capacity in VSTM is set in terms of objects but with an additional independent process that can bolster some representations over others.

A second novel contribution of our research is that we jointly modeled performance across substantially different VSTM tasks. This is important because researchers seek to discover fundamental characteristics of VSTM, rather than those specific to one VSTM task or another. Support for pure resource models of VSTM was consistent across tasks. Resource models provided the best account of the data regardless of whether participants had to remember simple features bound to spatial locations, or pictures of real-world objects presented sequentially. Results from the joint model analysis further reveal that performance across the two tasks cannot be described with a single estimate of capacity or resource. This finding is consistent with previous evidence that memory capacity may differ for simple features and real-world objects (Brady, Konkle, & Alvarez, 2013), and that the distribution of resources may differ for simultaneously and sequentially presented stimuli (Gorgoraptis et al., 2011). Important in the current context is that this finding supports a basic prediction of continuous resource models of VSTM, which is that varying the informational content of to-be-remembered items and their encoding conditions will lead to differences in how a memory resource is distributed to encode and/or

maintain representations.

Finally, we found that pure resource models were best at capturing performance across different sessions of the same task. More specifically, we found that fully joint signal detection models, in which all parameters are fixed across different sessions of the same task, performed better than disjoint variants of these models. In some cases, we did find that fully disjoint mixture models performed comparably to fully joint signal detection models; in those cases, average AIC was consistently lower for the latter models despite the fact that disjoint mixture models had more than twice the number of parameters of fully joint EVSD models (40 and 18 parameters, respectively). From the perspective of parsimony, these results favor EVSD models, though they may not yet be considered conclusive.

5.1. Interpretation of signal detection models of VSTM

Throughout the manuscript we have described the latent variables captured by the EVSD and UVSD models as reflecting a continuous resource that is distributed across memory representations, either in a uniform or variable fashion, respectively. Our results are consistent with such a general model of VSTM, but they do not speak to what this resource represents. This resource could reflect sustained spiking activity in neural populations (Schneegans & Bays, 2018). An alternative, not mutually exclusive interpretation recently proposed by Schurgin, Brady, and Wixted (2018) is that the signal detection models capture the fact that the decision process in VSTM tasks is based on a single continuous familiarity signal. In principle, the basis of this familiarity signal may differ considerably, depending on the type of stimuli that people are required to remember. Meaningfulness and presentation regimen likely affect the signal extracted from memory in support of recognition discrimination. Models that postulate a multidimensional evidence space and projection onto a decision axis are entirely compatible with this claim (e.g., Banks, 2000). Our finding that a single parameter of these signal detection models cannot capture performance when task demands are varied is consistent with such an interpretation. Finally, this result comports with models based on resource-rational theory according to which individuals optimally calibrate the allotment of VSTM resources to the demands of the task (van den Berg & Ma, 2018).

5.2. Why have others found evidence for the discrete-slot model of capacity?

In previous studies, other groups have applied ROC analysis to test models of VSTM capacity and reported evidence for the discrete-slot model (Donkin et al., 2014; Rouder et al., 2008), a finding that is at odds with our own and those of others (e.g., Wilken & Ma, 2004). One major difference between these studies and ours is in how response bias was manipulated or measured to plot points on the ROC curve. Specifically, we collected participants' confidence ratings whereas Rouder et al. (2008) and Donkin et al. (2014) induced bias experimentally by varying true base rates (i.e., the probability of a change versus no change occurrence). Previous research indicates that the use of base rate manipulations can lead to clustering of ROC points and/or non-monotonic ROC curves, both of which may lead researchers to erroneously infer evidence for linear ROC functions (Dube, Rotello, & Heit, 2012).

Human subjects are often quite insensitive to base rate manipulations (Benjamin, Diaz, & Wee, 2009); such manipulations may additionally be unsuitable because they may introduce differences in accuracy and/or response strategies, which would violate basic assumptions of ROC analysis (Balakrishnan, 1999; Dube & Rotello, 2012; Markowitz & Swets, 1967). This point also provides an alternative interpretation to another finding by Donkin et al. (2014). Specifically, these authors reported that the discrete-slot model outperformed resource-based models only under conditions in which set size (the number of to-be-remembered items) varied across trials, whereas a continuous resource model outperformed the discrete-slot model when set size was fixed within an experimental session. In another recent paper, Donkin, Kary, Tahir, and Taylor (2016) replicated this result and proposed that people use different encoding strategies under conditions when set sizes are variable than when set sizes are fixed—specifically, people choose to distribute resources to a wide range of items when the number of to-be-remembered items is predictable (fixed set size condition), but choose to encode a subset of items when the number of to-be-remembered items is not predictable (variable set size condition). Although this interpretation describes the pattern of results found by these authors it does not provide an explanation for why such encoding strategies would be employed. An equally reasonable (and contrary) alternative prediction would have been that people may try to encode a specific subset of items with high resolution when they *can* predict the number of items in the memory array because they can anticipate the spatial distribution of these items.

We propose an alternative interpretation, which is that manipulating both the number of to-be-remembered items and the true base rate (e.g., probability of a change versus no change) increases variability in criterion setting and ROCs that are consequently difficult to interpret (Benjamin et al., 2009). Our own results serve as converging evidence for this interpretation, given that we did not use a base rate manipulation, varied the number of to-be-remembered items randomly across trials, and found consistent evidence for continuous resource rather than discrete-slot models of capacity.

Finally, we note that a few recent studies have reported evidence for variants of the mixture model (e.g., Adam, Vogel, & Awh, 2017; Nosofosky & Gold, 2017) discussed here. Unlike the standard discrete-slot model (e.g., Cowan, 2001) which assumes all-or-none processing, the mixture model does posit that items in memory can be stored with variable precision. Although we found that this model outperformed standard discrete-slot models, we did not find that it outperformed a model that assumed no fixed item limit across tasks (the signal detection models). Studies that reported evidence for the mixture model used tasks that differed substantially from the two-alternative forced choice task used here. For instance, Adam et al. (2017) combined the method of adjustment and whole report methodology, Nosofsky and Donkin (2016) emphasized accuracy and response speed to test variants of linear ballistic models using reaction times in an RSVP change detection task, and Nosofosky and Gold (2017) used a multiple-alternative response task with payoffs. Therefore, one possibility is that the discrepancy in our results and the results of these researchers is due to these

methodological differences. We also do not consider all possible variants of mixture models, such as those in which item limits can vary over trials (e.g., Adam, Mance, Fukuda, & Vogel, 2015). Although such models make different architectural assumptions by assuming a discrete-item limit, in practice they can be difficult to differentiate from pure resource models that assume variable precision across items and trials (Adam et al., 2017; Ma, 2018)[5]. Furthermore, other researchers that used designs more like those used by some of these authors also reported evidence for continuous resource models of memory (e.g., Bays, 2018). As noted, we believe that a promising way of reconciling these conflicting results is by developing and testing models that can capture performance across diverse experimental conditions as well as by focusing on the predictive ability of these models.

5.3. Metrics of VSTM capacity

Based on our own and others' results, we advise researchers to move away from using (K) metrics based on the discrete-slot model of capacity to characterize VSTM performance. Although such metrics may appear to have heuristic value, they inherit the strong and untenable processing assumption of classic threshold models—specifically, that individual capacity is set in terms of objects that are stored in an all-or-none fashion with no noise in the encoding or maintenance of representations in VSTM. Furthermore, the use of K metrics may obfuscate differences in performance that are appropriately captured by other measures based on continuous resource models, such as sensitivity (d' or d_a), and response bias measures (e.g., C). In our own experiments, we found that there were important differences in how K and signal detection metrics captured performance in the change detection and RSVP task (see Appendix A). For instance, one systematic finding is that K estimates of capacity increased as a function of set size in the RSVP task but remained stable in the change detection task. In addition, K estimates of capacity were significantly higher in the RSVP task than the change detection task, but only when participants had to remember eight items (Experiment 2). The signal detection metrics revealed, however, that sensitivity was higher in the change detection task than in the RSVP task when participants had to remember four items, a pattern of results not captured by K estimates. Critical in the current context is that this pattern of results demonstrates how metrics based on discrete-slot models conflate aspects of performance that are due to the use of a different decision criterion and those that are due to true differences in sensitivity (for related criticisms of discrete-slot models see Bays (2018) and Ma (2018)).

A final demonstration of this point is the finding that response bias became more liberal as a function of set size in the change detection task but remained stable as a function of set size in the RSVP task (see Appendix A). The finding that response bias became more liberal as a function of load in the change detection task is consistent with other results showing that people may set a more conservative response criterion, or a higher standard for performance, when task demands are low (Benjamin & Bawa, 2004; Rotello & Macmillan, 2007). The fact that we did not find such an effect in the RSVP task may indicate that people employ alternative strategies, or do not have enough information to adjust their bias on a trial-by-trial basis in the RSVP task (cf. Stretch & Wixted, 1998).

5.4. Limitations and future directions

Throughout this manuscript we highlighted the potentially problematic application of base rate manipulations and proposed that these may have led researchers to incorrectly infer evidence for discrete-slot models in the change detection task. Our criticism is motivated in part by similar criticisms of base rate manipulations outside of the VSTM literature (Dube, Rotello, & Heit, 2010) as well as the fact that researchers who used confidence ratings instead of base rate manipulations to conduct ROC analysis have found consistent evidence for resource models (Wilken & Ma, 2004). It is important to acknowledge, however, that the use of confidence ratings to construct ROC curves has also been challenged. For instance, Krantz (1969) argued that participants' response strategies, or use of the confidence scale, may lead to curvilinear ROC curves even when the underlying data-generating process is threshold-based. Although we have not evaluated this possibility in our data, in practice there appears to be little evidence that curvilinear ROC functions are an artifact of individuals' response strategies (Pazzaglia et al., 2013). In addition, the response rules that must be employed in order to render curvilinear ROCs from pure threshold models are complicated and probably implausible (Wixted, 2019). The claim that pure resource models may provide a superior description of data in the change detection task is also consistent with work that has used alternative methodologies and Bayesian variants of the models we consider here (Keshvari, van den Berg, & Ma, 2013).

Another limitation of our study is that we did not consider potential effects of item variability on model fits, which may have affected parameter estimates of the signal detection model (Rouder et al., 2007). This form of aggregation can lead to parameter underestimation in the signal detection model and may have had an impact on the joint modelling results because the effects of item variability may be different in tasks that use simple features versus real-world objects. In future work, researchers may address this limitation by applying hierarchical modelling analysis to jointly model performance across tasks (Turner et al., 2013).

Relatedly, a fruitful venue for future research is to apply joint modeling analysis while more systematically varying the demands of different VSTM tasks. Our results indicate that performance cannot be captured by a single latent variable across the two tasks we employed, but it is unclear whether this is due to the differences in presentation format, encoding time, the stimuli, or some combination of all these factors. Furthermore, even though we did not find that extant models can accommodate performance across the two tasks with the same parameters, this does not mean that in principle there can be no VSTM model that can represent performance

[5] As noted earlier in the manuscript, the general version of such mixture models is formally equivalent to some variants of resource models that assume that memory signals are not processed on some proportion of trials due to attentional lapses, for instance.

across the two tasks with some shared parameters. Developing such models is an important goal for future research, though we suspect that it will require more empirical research that elucidates what processes are shared and not shared across the two tasks. More generally, we propose that joint modeling analysis of this kind may help researchers move beyond evaluating models of VSTM under highly restricted experimental conditions, such as the change detection or delayed estimation task with simple stimuli, and towards evaluating their generality.

5.5. Summary and conclusions

We applied ROC and cross-validation to evaluate candidate models of VSTM capacity within and across different VSTM tasks. Our results show consistent support for signal-detection, continuous-resource models of capacity within these tasks. Our interpretation of these models in this context is that it captures the allocation of a continuous resource that is either uniformly or unevenly distributed across memory representations, depending on the task demands. Our results also reveal that performance cannot be captured by a common estimate of a resource across these two tasks. This finding further invalidates the use of a single metric (e.g., K) as a means of summarizing individuals' VSTM capacity. Based on our collective findings, we conclude that there is no K in capacity, and that researchers should move towards signal detection metrics to capture performance in VSTM tasks.

Appendix A

In this section we summarize results of transformations of the data based on the discrete-slot and equal-variance signal detection model. We also calculated a measure of performance based on the unequal variance model (d_a). For ease of exposition, these measures as well as estimates of ROC slopes and accuracy are summarized for each experiment in Tables A1, A2 and A3, respectively. Importantly, these results did not guide model selection; we report them to provide researchers with insight into how these different metrics may lead to different conclusions regarding the effects of VSTM load on processing, particularly across different types of VSTM tasks. K capacity was calculated using the formula $K = (H - FA) * N$, where H denotes the observed hit rate, FA denotes the observed false alarm rate, and N denotes the number of to-be-remembered items on a given trial. The reported signal detection measures include estimates of sensitivity (d') and response bias (C). Sensitivity (d') was calculated using the formula $\Phi^{-1}(H) - \Phi^{-1}(FA)$, where H and FA denote hit and false alarm rates, respectively, and Φ^{-1} denotes the inverse of the cumulative normal distribution function. C response bias was calculated using the formula $-0.5 * (\Phi^{-1}(H) + \Phi^{-1}(FA))$. In cases when the

Table A1
Sensitivity metric of the UVSD model (d_a). Table contains mean d_a for each set size in each experiment. Values in parentheses denote standard deviations.

	Number of items in memory array		
	Four items	Six items	Eight items
Experiment 1a	2.08 (0.66)	1.44 (0.54)	1.07 (0.42)
Experiment 1b	2.24 (0.53)	1.48 (0.56)	1.06 (0.44)
Experiment 1c	1.57 (0.80)	1.19 (0.62)	1.01 (0.64)
Experiment 2: Change detection task	2.07 (0.79)	1.38 (0.64)	0.92 (0.47)
Experiment 2: RSVP task	1.74 (0.60)	1.30 (0.43)	1.15 (0.42)
Experiment 3: Change detection group	2.21 (0.79)	1.53 (0.63)	1.07 (0.51)
Experiment 3: RSVP group	1.70 (1.05)	1.25 (0.81)	1.11 (0.65)

Table A2
Mean regression estimates of slopes of standardized (z-score converted) ROC curves. Values in parentheses denote standard deviations.

	Number of items in memory array		
	Four items	Six items	Eight items
Experiment 1a	0.87 (0.53)	1.00 (0.35)	1.15 (0.48)
Experiment 1b	0.92 (0.59)	0.98 (0.27)	1.01 (0.24)
Experiment 1c	0.75 (0.25)	0.77 (0.25)	0.79 (0.21)
Experiment 2: Change detection task	1.05 (0.93)	0.94 (0.30)	1.03 (0.32)
Experiment 2: RSVP task	0.74 (0.16)	0.74 (0.17)	0.81 (0.22)
Experiment 3: Change detection group	0.85 (0.48)	0.94 (0.32)	1.01 (0.26)
Experiment 3: RSVP group	0.76 (0.45)	0.70 (0.27)	0.78 (0.27)

Table A3
Proportion correct aggregated across change and no-change (change detection task) and new and old (RSVP task) trials. Values in parentheses denote standard deviations.

	Number of items in memory array		
	Four items	Six items	Eight items
Experiment 1a	0.83 (0.10)	0.74 (0.09)	0.67 (0.08)
Experiment 1b	0.86 (0.07)	0.73 (0.07)	0.66 (0.07)
Experiment 1c	0.80 (0.06)	0.74 (0.06)	0.71 (0.07)
Experiment 2: Change detection task	0.83 (0.10)	0.72 (0.09)	0.65 (0.06)
Experiment 2: RSVP task	0.80 (0.08)	0.74 (0.06)	0.72 (0.06)
Experiment 3: Change detection group	0.84 (0.10)	0.74 (0.09)	0.67 (0.08)
Experiment 3: RSVP group	0.80 (0.16)	0.74 (0.14)	0.71 (0.12)

observed hit and false alarm rates were 1 or 0, respectively, we applied a correction by replacing rates of 1 with $(n - 0.5)/n$ and rates of 0 with $0.5/n$, where n denotes the number of change or no-change trials (Macmillan & Kaplan, 1985). For each non-significant difference, we report a Bayesian Information Criterion approximation of the Bayesian posterior probability that quantifies evidence for the null ($p_{BIC}[H_0|D]$; see Masson, 2011). $p_{BIC}(H_0|D)$ values of 0.50-0.75 indicate weak evidence for the null, values of 0.75-0.95 indicate positive evidence for the null, values of 0.95-0.99 indicate strong evidence for the null, and values > 0.99 indicate very strong evidence for the null.

Experiment 1a. Average metrics based off the discrete-slot and signal detection models are shown in Fig. A1. *K* capacity estimates were similar across set sizes ($F(2, 58) = 1.5$, $p = .23$, $MSe = 0.22$; $p_{BIC}(H_0|D) = 0.93$). As expected, and in contrast, *d'* estimates decreased with increasing set size ($F(2, 58) = 73.8$, $p < .001$, $MSe = 0.1$). A Scheffé 95% confidence interval based on the error term of this effect was ± 0.21. Based on this confidence interval, we found that *d'* was higher at set size four than set size six and higher at set size six than set size eight. We also found that response bias (*C*) became more liberal with increasing set size ($F(2, 58) = 19.6$, $p < .001$, $MSe = 0.04$; Scheffé 95% CI: ±0.13); *C* was higher at set size four than set size six and higher at set size six than at set size eight.

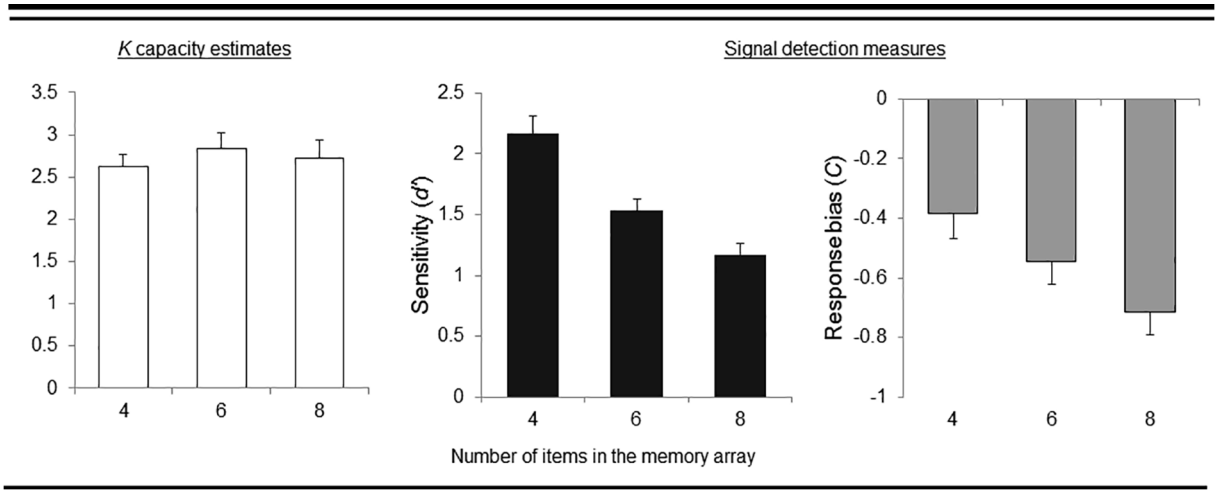

Fig. A1. Discrete-slot and signal detection model measures of performance (Experiment 1a).

Experiment 1b. Fig. A2 shows results for metrics based on the discrete-slot and signal detection models. Like in Experiment 1a, *K* capacity estimates were similar ($F(2, 58) = 2.44$, $p = 0.1$, $MSe = 0.31$; $p_{BIC}(H_0|D) = 0.84$) for set size four, set size six and set size eight. The *d'* estimates decreased with increasing set size ($F(2, 58) = 124.8$, $p < .001$, $MSe = 0.13$; Scheffé 95% CI: ±0.23). As before, *d'* was higher at set size four than set size six and higher at set size six than set size eight. Finally, we found that response bias (*C*) became more liberal with increasing set size ($F(2, 58) = 31.5$, $p < .001$, $MSe = 0.03$; Scheffé 95% CI: ±0.11). Once again, *C* was higher at set size four than set size six and higher at set size six than set size eight.

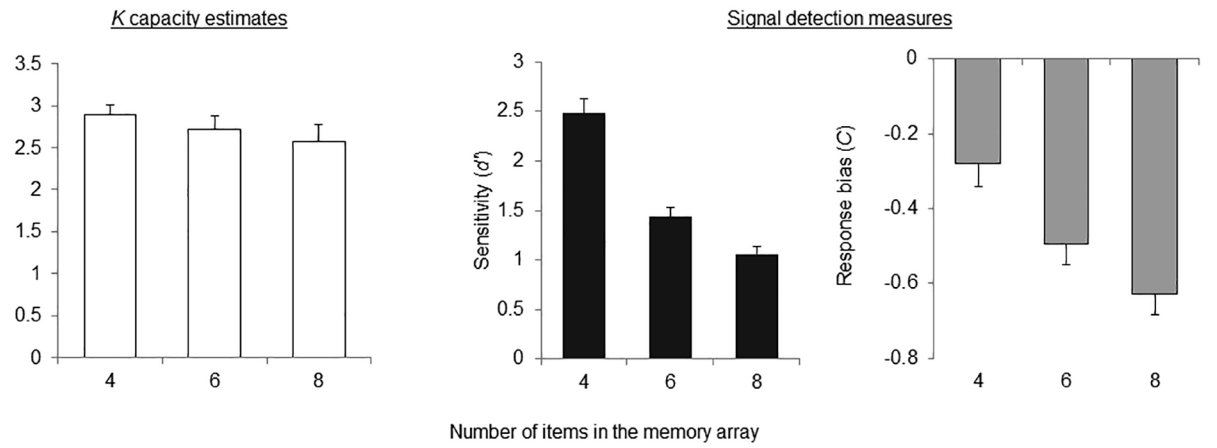

Fig. A2. Discrete-slot and signal detection model measures of performance (Experiment 1b).

Experiment 1c. Metrics are shown in Fig. A3. As before, we calculated standard metrics for both types of models (Fig. 6b). Unlike in Experiment 1a and 1b, we found that *K* capacity estimates increased with increasing set size in this task ($F(2, 58) = 22.8$, $p < 0.001$, $MSe = 0.28$; Scheffé 95% CI: ±0.34). *K* was lower at set size four than set size six and *K* was lower at set size six than at set size eight. As before, *d'* estimates decreased with increasing set size ($F(2, 58) = 47.6$, $p < 0.001$, $MSe = 0.08$; Scheffé 95% CI: ±0.18). Thus, *d'* was higher at set size four than at set size six and higher at set size six than at set size eight. However, unlike in Experiments 1a and 1b we found that response bias *C* was similar ($F(2, 58) = 1.1$, $p = 0.34$, $MSe = 0.02$; $p_{BIC}(H_0|D) = 0.93$) at set size four, set size six and set size eight.

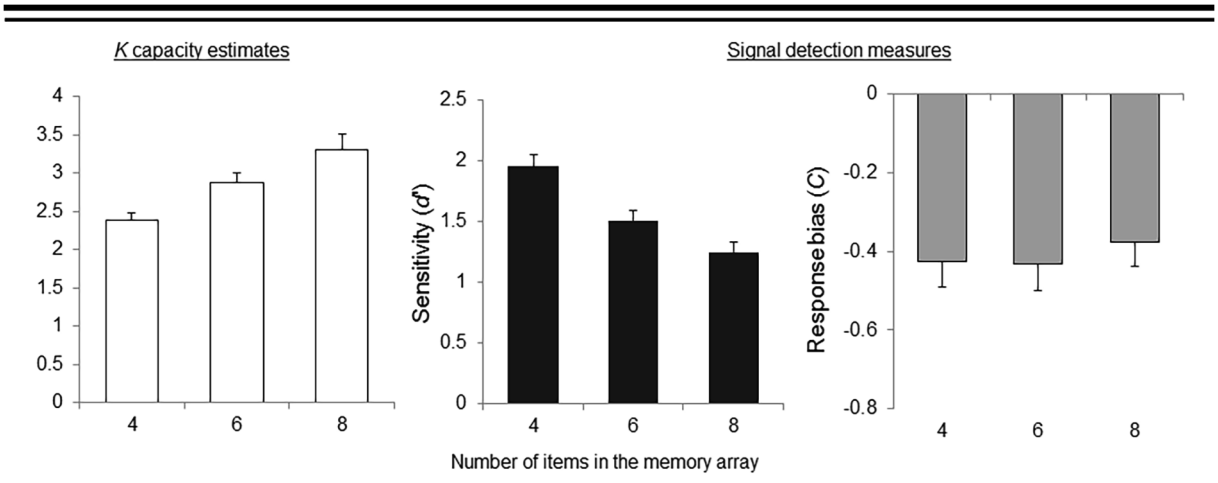

Fig. A3. Discrete-slot and signal detection model measures of performance (Experiment 1c).

Experiment 2. Metrics are shown in Fig. A4. We ran a two-way repeated measures ANOVA with set size (four, six or eight items) and task (change detection versus RSVP) as the within-subject factors to compare performance for each type of metric. There was a main effect of task ($F(1, 58) = 5.37$, $p = .028$, $MSe = 1.04$) with *K* capacity estimates being higher on the RSVP task than on the change detection task. There was also a main effect of memory set size ($F(2, 58) = 11.9$, $p < .001$, $MSe = 0.22$; Scheffé 95% CI: ±0.22). Average *K* estimates were lower at set size four than set size six and lower at set size four than at set size eight, but were not statistically different between set size six and eight. Critically, there was a significant interaction between task and set size ($F(2, 58) = 11.9$, $p < .001$, $MSe = 0.26$; Scheffé 95% CI: ±0.33). Based on this confidence interval, we found that K capacity estimates increased with increasing set size in the RSVP task, but did not change as a function of set size in the change detection task ($p_{BIC}(H_0|D) = 0.65$). We also found that *K* capacity estimates were similar in the change detection and RSVP tasks at lower set sizes (i.e., four and six items), but were significantly higher in the RSVP task than in the change detection task at the highest set size (eight items).

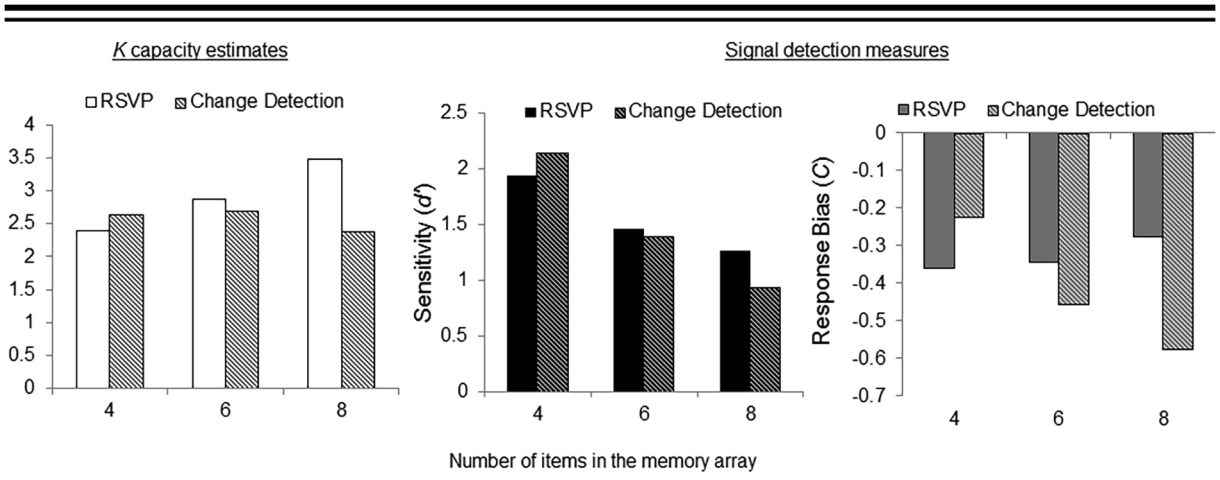

Fig. A4. Discrete-slot and signal detection model measures of performance (Experiment 2).

We ran a similar ANOVA with task and set size as the within-subject factors using the signal detection metrics as the dependent variable. An ANOVA with *d'* as the dependent variable revealed no main effect of task ($F < 1$). There was an effect of memory set size ($F(2, 58) = 92.1$, $p < .001$, $MSe = 0.15$; Scheffé 95% CI: ±0.18). As before, average *d'* was higher at set size four than set size six, and higher at set size six than set size eight. Importantly, we found a significant interaction between task and set size ($F(2, 58) = 12.3$, $p < .001$, $MSe = 0.09$; Scheffé 95% CI: ±0.19). Similar to the other experiments, we found that *d'* decreased with increasing set size in the RSVP task, and also decreased with increasing set size in the change detection task. In addition, we found that *d'* was significantly higher in the change detection task than in the RSVP task when participants had to remember 4 items; however, this time we also found that *d'* was significantly higher in the RSVP task than in the change detection task when participants had to remember 8 items. Finally, a two-way ANOVA with task and set size as the two within-subject factors and response bias (*C*) as the dependent variable revealed no main effect of task ($F(1, 29) = 2.5$, $p = .12$, $MSe = 0.15$; $p_{BIC}(H_0|D) = 0.90$). There was a main effect of set size ($F(2, 58) = 10.2$, $p < .001$, $MSe = 0.03$; Scheffé 95% CI: ±0.08). On average, response bias was more liberal at set size six than at set size four and more liberal at set size eight than at set size four, but there was no significant difference in response bias between set size six and set size eight. Importantly, there was a significant interaction between task and set size ($F(2, 58) = 21.9$, $p < .001$, $MSe = 0.03$; Scheffé 95% CI: ±0.11). Based on this confidence interval, we found that in the RSVP task response bias did not change as a function of set size ($p_{BIC}(H_0|D) = 0.91$). In contrast, response bias became more liberal with each increase in set size in the change detection task. Response bias was more liberal on the RSVP than the change detection task when participants had to remember four items, but was more liberal in the change detection than the RSVP task when participants had to remember six or eight items.

Experiment 3: Change Detection task. Metrics are shown in Fig. A5. We ran a two-way ANOVA with set size and session as the within-subject factors and *K* capacity estimates as the dependent variable. We found no main effect of session, no main effect of set size (all $F < 1$, minimum $p_{BIC}(H_0|D) = 0.84$), and no interaction between set size and session ($F < 2$; $p_{BIC}(H_0|D) = 0.93$).

Next, we ran the same ANOVA using the signal detection metrics as the dependent variable. An ANOVA with *d'* as the dependent variable revealed no main effect of session ($F < 1$; $p_{BIC}(H_0|D) = 0.84$), but a main effect of set size ($F(2,58) = 136$, $p < 0.001$, $MSe = 0.17$; Scheffé95% CI: ±0.27). Based on this difference, we found that *d'* estimates decreased with set size in the change detection task. The interaction between session and set size was not significant ($F(2,58) = 2.13$, $p = 0.13$, $MSe = 0.08$; $p_{BIC}(H_0|D) = 0.88$). We ran the same ANOVA with *C* response bias as the dependent variable. This revealed no main effect of session ($F < 1$; $p_{BIC}(H_0|D) = 0.84$), but a main effect of set size ($F(2,58) = 45.3$, $p < .001$, $MSe = 0.04$; Scheffé 95% CI: ±0.13). Based on this confidence interval, we again found that response bias became more liberal with each increase in set size in the change detection task. The interaction between session and set size was not significant ($F < 1$; $p_{BIC}(H_0|D) = 0.98$).

Experiment 3: RSVP task. Metrics are shown in Fig. A5. As before, we ran a two-way ANOVA with set size and session as the two within-subject factors and *K* capacity estimates as the dependent variable. There was no main effect of session ($F < 1$; $p_{BIC}(H_0|D) = 0.79$), but a main effect of set size ($F(2,58) = 28.4$, $p < .001$, $MSe = 0.39$; Scheffé 95% CI: ±0.41). As before, we found that capacity estimates increased with each increase in set size in the RSVP task. The interaction between session and set size was not significant ($F < 2$; $p_{BIC}(H_0|D) = 0.91$).

Next, we ran the same two-way ANOVA using the signal detection metrics as the dependent variable. An ANOVA with *d'* as the dependent variable revealed no main effect of session ($F < 2$; $p_{BIC}(H_0|D) = 0.70$), but an expected effect of set size ($F(2,58) = 42.4$, $p < .001$, $MSe = 0.24$; Scheffé 95% CI: ±0.32). Based on this difference, we found that *d'* estimates decreased with each increase in set size in the RSVP task. The interaction between session and set size was not significant ($F < 1$; $p_{BIC}(H_0|D) = 0.98$). Finally, we ran the same two-way ANOVA using response bias (*C*) as the dependent variable. There was no main effect of session ($F < 1$; $p_{BIC}(H_0|D) = 0.84$), no main effect of set size ($F(2, 58) = 2.22$, $p = 0.12$, $MSe = 0.03$; $p_{BIC}(H_0|D) = 0.89$), and no interaction between set size and session ($F < 1$; $p_{BIC}(H_0|D) = 0.98$).

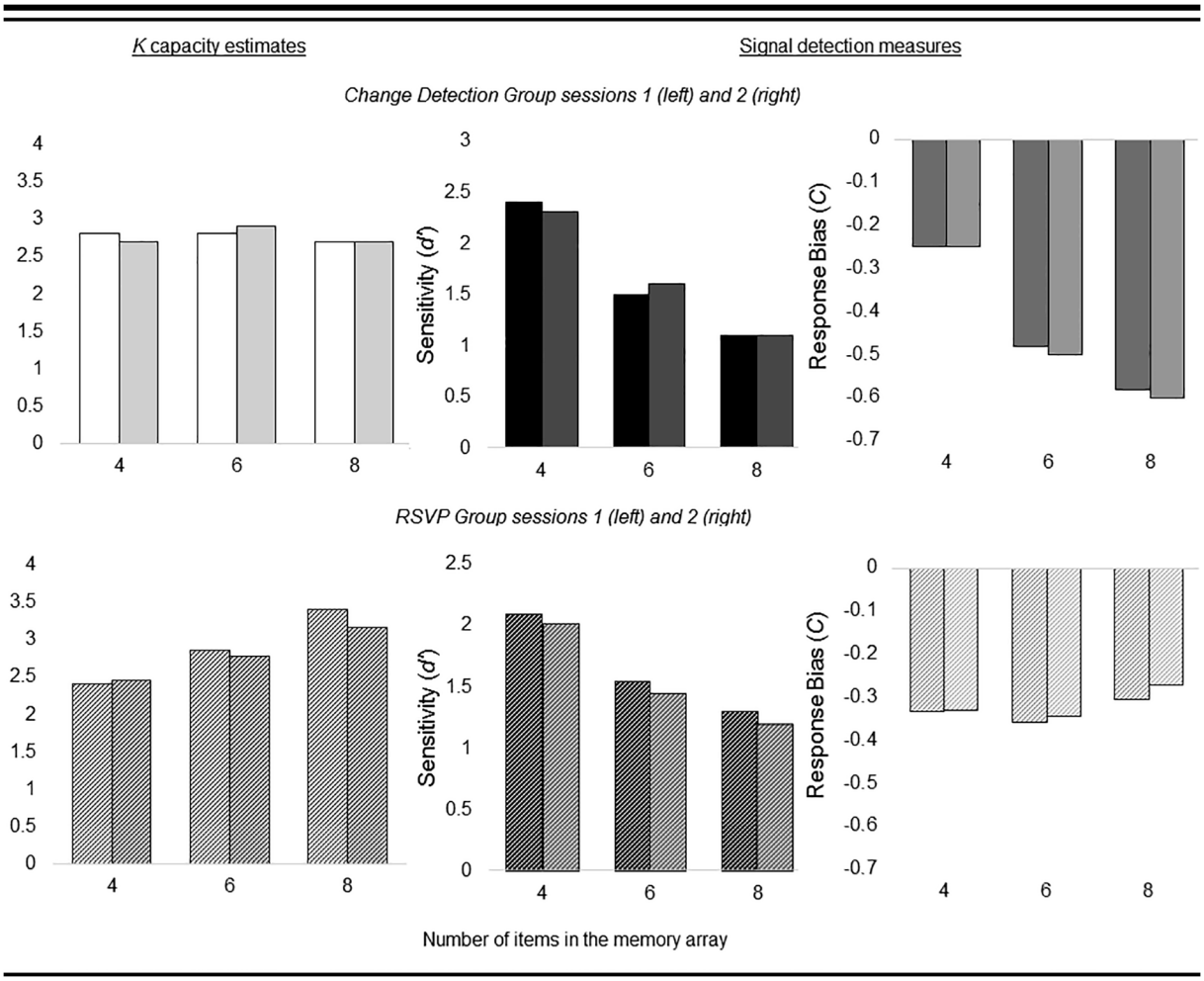

Fig. A5. Discrete-slot and signal detection model measures of performance (Experiment 3).

Appendix B

B.1. Model recovery simulations

We used model recovery simulations to compare different indices of model performance on their ability to recover the true data generating model. The model recovery analysis we use is similar to that used by van den Berg, Awh, and Ma (2014). We fit five different models to data of each subject. The five models we used were the first set of models described in the body of the paper. The model recovery analysis provides insight into the flexibility of models to account for data generated under incompatible assumptions, and so provides a tool for assessing which metrics of model quality are most valid under the design and data structure of the experiments reported here. For these simulations we fit each model separately to each participant's data. We repeated this for all participants from Experiment 1a-c, both sessions of Experiment 2, and both sessions of both tasks in Experiment 3 (a total of 270 participants). After fitting each model, we used best-fitting parameters of each model to generate synthetic data for that model using the number of observations used in our studies (360 in each experimental session). Then we fit and cross-validated each of the five models to each participant's synthetic data. We used average measures of fit and the proportion of times that the correct model was selected by each metric to assess the diagnosticity of AIC, BIC and the cross-validation residual error metric. These results are summarized in Table B1.

Table B1
Summary of model recovery analyses. Cells highlighted in gray indicate incorrect model selections. Cells highlighted in black indicate correct model selections.

BIC

Fitted model	DS-F	DS-V	Mix	EVSD	UVSD
DS-F	932 (99)	932 (97)	933 (95)	935 (75)	937 (72)
DS-V	988 (1)	971 (3)	979 (2)	979 (0)	976 (0)
Mix	995 (0)	980 (0)	981 (2)	984 (0)	981 (0)
EVSD	1006 (0)	989 (0)	991 (0)	950 (23)	949 (24)
UVSD	1018 (0)	1000 (0)	1001 (0)	963 (2)	958 (4)
	DS-F	DS-V	Mix	EVSD	UVSD

Data generating model

AIC

Fitted model	DS-F	DS-V	Mix	EVSD	UVSD
DS-F	909 (87)	909 (35)	909 (34)	914 (9)	914 (6)
DS-V	918 (10)	902 (46)	909 (23)	909 (1)	906 (3)
Mix	919 (3)	905(13)	905 (34)	908 (3)	905 (3)
EVSD	936 (0)	919 (4)	921 (6)	880 (67)	880 (41)
UVSD	936 (0)	919 (2)	920 (3)	881 (18)	876 (47)
	DS-F	DS-V	Mix	EVSD	UVSD

Data generating model

Cross-validation

Cross-validated model	DS-F	DS-V	Mix	EVSD	UVSD
DS-F	.10 (10)	.15 (1)	.13 (3)	.24 (0)	.26 (0)
DS-V	.09 (39)	.09 (67)	.10 (30)	.19 (2)	.20 (3)
Mix	.09 (31)	.14 (4)	.09 (38)	.19 (2)	.20 (1)
EVSD	.12 (3)	.11 (6)	.11 (4)	.08 (37)	.09 (12)
UVSD	.11 (18)	.11 (23)	.10 (25)	.08 (59)	.08 (84)
	DS-F	DS-V	Mix	EVSD	UVSD

Data generating model

As shown, AIC performed best in terms of recovering the true data-generating model. This was reflected in the average AIC values as well as the fact that this metric selected the true model the majority of the time across participants. Cross-validation performed second best. We found that cross-validation was somewhat biased towards more flexible models. For instance, it was biased towards the DS-V model when the DS-F model generated the data and the UVSD model when the EVSD model generated the data. Notably, however, cross-validation performed better than all other metrics at recovering the general structure of the data-generating model—i.e., whether it was discrete, continuous, or a mix of the two. Since it is this dimension that is most central to the conclusions we draw, we feature cross-validation in addition to AIC in the paper. Finally, BIC was consistently biased towards the model with the fewest number of parameters (DS-F) and failed to recover the true generating model at a higher rate than the other statistics.

Appendix C

Tables C1–C3.

Table C1
AIC values and Akaike weights for each of the contending models in Experiments 1 a-c. Akaike weights can be interpreted as the conditional probability that a given model provides the best description of the data given the data and the set of models considered here. Underlined values in bold denote the best performing models based on AIC and Akaike weight values.

Criteria across set sizes	Model	Experiment					
		1a		1b		1c	
		AIC	$w(AIC)$	AIC	$w(AIC)$	AIC	$w(AIC)$
Free	DS-F	884	0.02	915	< 0.01	932	0.04
	DS-V	885	0.01	914	0.02	933	0.02
	DS-A	871	0.14	902	0.14	941	< 0.01
	Mixture	871	0.03	899	0.14	928	0.04
	EVSD	**863**	**0.26**	**895**	**0.27**	923	0.06
	UVSD	865	0.19	897	0.12	922	0.06
Fixed	DS-F	882	0.06	919	0.02	928	0.14
	DS-V	883	0.03	916	0.10	927	0.10
	DS-A	873	0.14	911	0.18	934	0.09
	Mixture	872	0.06	907	0.01	920	0.09
	EVSD	884	0.04	929	< 0.01	921	0.10
	UVSD	881	0.02	925	< 0.01	**918**	**0.26**

Table C2
AIC values and Akaike weights for each of the contending models in Experiment 2.

Task model	Model	Criteria across set sizes							
		Free in both tasks		Fixed in both tasks		Free in RSVP/fixed in change detection		Free in change detection/fixed in RSVP	
		AIC	$w(AIC)$	AIC	$w(AIC)$	AIC	$w(AIC)$	AIC	$w(AIC)$
Fully joint	DS-F	1907	< 0.01	1902	< 0.01				
	DS-V	1908	< 0.01	1902	< 0.01				
	Mix	1887	< 0.01	1899	< 0.01				
	AT Mix						N/A		
	EVSD	1881	0.01	1901	< 0.01				
	UVSD	1883	< 0.01	1897	< 0.01				
Joint capacity	DS-F	1890	< 0.01	1884	0.02	1951	< 0.01	1909	0.01
	DS-V	1891	< 0.01	1886	< 0.01	1955	< 0.01	1913	< 0.01
	Mix	1867	< 0.01	1864	< 0.01	1892	< 0.01	1860	0.07
	AT Mix	1861	< 0.01	1857	0.04	1884	< 0.01	1855	0.07
	EVSD	1856	0.02	1876	< 0.01	1902	< 0.01	1852	0.10
	UVSD	1858	0.02	1872	< 0.01	1898	< 0.01	1854	0.02
Fully disjoint	DS-F	1883	0.01	1874	0.03	1936	< 0.01	1903	< 0.01
	DS-V	1884	< 0.01	1877	0.04	1901	< 0.01	1878	< 0.01
	Mix	1853	0.02	1851	0.09	1877	< 0.01	1864	0.05
	AT Mix						N/A		
	EVSD	1851	0.06	1871	< 0.01	1897	< 0.01	**1847**	**0.15**
	UVSD	1854	0.07	1868	< 0.01	1894	< 0.01	1848	0.04

Table C3
AIC values and Akaike weights for each of the contending models in Experiment 3.

Criteria across set sizes	Task model	Model	Task			
			Change detection		RSVP	
			AIC	$w(AIC)$	AIC	$w(AIC)$
Free	Fully joint	DS-F	1744	0.03	1849	<0.01
		DS-V	1743	0.01	1849	<0.01
		Mixture	1716	0.05	1814	0.03
		EVSD	**1705**	**0.21**	1791	0.05
		UVSD	1707	0.13	**1782**	**0.17**
	Joint capacity	DS-F	1744	0.02	1853	<0.01
		DS-V	1743	0.02	1853	<0.01
		Mixture	1719	0.05	1818	<0.01
		EVSD	1706	0.05	1794	<0.01
		UVSD	1708	0.04	1797	0.07
	Fully disjoint	DS-F	1743	0.01	1851	<0.01
		DS-V	1745	<0.01	1854	<0.01
		Mixture	1707	0.07	1806	<0.01
		EVSD	1707	0.05	1795	0.03
		UVSD	1712	0.02	1803	0.05
Fixed	Fully joint	DS-F	1814	<0.01	1875	<0.01
		DS-V	1752	<0.01	1844	0.02
		Mixture	1735	<0.01	1812	0.02
		EVSD	1764	<0.01	1806	<0.01
		UVSD	1756	<0.01	1796	<0.01
	Joint capacity	DS-F	1806	<0.01	1872	<0.01
		DS-V	1743	0.09	1838	0.06
		Mixture	1725	0.04	1809	0.12
		EVSD	1756	<0.01	1799	0.03
		UVSD	1748	<0.01	1788	0.02
	Fully disjoint	DS-F	1805	<0.01	1872	<0.01
		DS-V	1745	0.03	1839	0.06
		Mixture	1713	0.09	1797	**0.17**
		EVSD	1757	<0.01	1800	0.05
		UVSD	1752	<0.01	1793	0.01

References

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

Adam, K. C. S., Mance, I., Fukuda, K., & Vogel, E. K. (2015). The contribution of attentional lapses to individual differences in visual working memory capacity. *Journal of Cognitive Neuroscience, 27*, 1601–1616.

Adam, K. C. S., Vogel, E. K., & Awh, E. (2017). Clear evidence for item limits in visual working memory. *Cognitive Psychology, 97*, 79–97.

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.

Balakrishnan, J. D. (1999). Decision processes in discrimination: Fundamental misrepresentations of signal detection theory. *Journal of Experimental Psychology: Human Perception & Performance, 25*, 1189–1206.

Banks, W. P. (2000). Recognition and source memory as multivariate decision processes. *Psychological Science, 11*, 267–273.

Bays, P. M., & Husain, M. (2008). Dynamic shifts of limited working memory resources in human vision. *Science, 321*(5890), 851–854.

Bays, P. M. (2018). Failure of self-consistency in the discrete resource model of visual working memory. *Cognitive Psychology, 105*, 1–8.

Benjamin, A. S., & Bawa, S. (2004). Distractor plausibility and criterion placement in recognition. *Journal of Memory & Language, 51*, 159–172.

Benjamin, A. S., Diaz, M. L., & Wee, S. (2009). Signal detection with criterion noise: Applications to recognition memory. *Psychological Review, 116*, 84–115.

Benjamin, A. S., Tullis, J. G., & Lee, J. H. (2013). Criterion noise in ratings-based recognition: Evidence from the effects of response scale length on recognition accuracy. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*, 1601–1608.

Berry, E. D. J., Waterman, A. H., Baddeley, A. D., Hitch, G. J., & Allen, R. J. (2018). The limits of visual working memory in children: Exploring prioritization and recency effects with sequential presentation. *Developmental Psychology, 54*, 240–253.

Blackwell, H. R. (1953). *Psychological thresholds: Experimental studies of methods of measurement.* Bulletin of the Engineering Research Institute of the University of Michigan36.

Brady, T. F., Störmer, V., & Alvarez, G. A. (2016). Working memory is not fixed capacity: More active storage capacity for real-world objects than simple stimuli. *Proceedings of the National Academy of Sciences, 113*, 7459–7464.

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

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision, 10*, 433–436.

Browne, M. W. (2000). Cross-validation methods. *Journal of Mathematical Psychology, 44*, 108–132.

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods and Research, 33*, 261–304.

Cox, G. E., Hemmer, P., Aue, W. R., & Criss, A. H. (2018). Information and processes underlying semantic and episodic memory across tasks, items and individuals. *Journal of Experimental Psychology: General, 147*, 545–590.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences, 24*, 87–114.

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

DeCarlo, L. (2010). On the statistical and theoretical basis of signal detection theory and extensions: Unequal variance, random coefficient, and mixture models. *Journal of Mathematical Psychology, 54*, 304–313.
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., Tran, S. C., & 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.
Dube, C., & Rotello, C. M. (2012). Binary ROCs in perception and recognition memory are curved. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*, 130–151.
Dube, C., Rotello, C. M., & Heit, E. (2010). Assessing the belief bias effect with ROCs: It's a response bias effect. *Psychological Review, 117*, 831–863.
Endress, A. D., & Potter, M. C. (2014). Large capacity temporary visual memory. *Journal of Experimental Psychology: General, 143*, 548–565.
Estes, W. K. (1956). The problem of inference from curves based on group data. *Psychological Bulletin, 53*, 134–140.
Feuerstahler, L. M., Luck, S. J., MacDonald, A., III, & Waller, N. G. (2019). A note on the identification of change detection task models to measure storage capacity and attention in visual working memory. *Behavior Research Methods, 51*, 1360–1370.
Fougnie, D., Cormiea, S. M., Kanabar, A., & Alvarez, G. A. (2016). Strategic trade-offs between quantity and quality in working memory. *Journal of Experimental Psychology: Human Perception and Performance, 42*, 1231–1240.
Fukuda, K., Vogel, E. K., Mayr, U., & Awh, E. (2010). Quantity not quality: The relationship between fluid intelligence and working memory capacity. *Psychonomic Bulletin and Review, 17*, 673–679.
Gorgoraptis, N., Catalao, R. F. G., Bays, P. M., & Husain, M. (2011). Dynamic updating of working memory resources for visual objects. *Journal of Neuroscience, 31*, 8502–8511.
Green, D. M., & Swets, J. A. (1988). *Signal detection theory and psychophysics* (reprint ed.). Los Altos, California: Peninsula Publishing.
Healy, A. F., & Jones, C. (1975). Can subjects maintain a constant criterion in a memory task. *Memory and Cognition, 3*, 233–238.
Healy, A. F., & Kubovy, M. (1978). The effects of payoffs and prior probabilities on indices of performance and cutoff location in recognition memory. *Memory and Cognition, 6*, 544–553.
Holbert, R. L., & Stephenson, M. T. (2002). Structural equation modeling in the communication sciences, 1995–2000. *Human Communication Research, 28*, 531–551.
Jang, Y., Wixted, J. T., & Hubner, D. E. (2009). Testing signal-detection models of yes/no and two-alternative forced choice recognition memory. *Journal of Experimental Psychology: General, 138*, 291–306.
Keshvari, S., van den Berg, R., & Ma, W. J. (2013). No evidence for an item limit in change detection. *PLoS Computational Biology, 9*. https://doi.org/10.1371/journal.pcbi.1002927.
Koen, J. D., Barrett, F. S., Harlow, I. M., & Yonelinas, A. P. (2016). The ROC Toolbox: A toolbox for analyzing receiver-operating characteristics derived from confidence ratings. *Behavior Research Methods, 49*, 1399–1406.
Krantz, D. H. (1969). Threshold theories of signal detection. *Psychological Review, 76*, 308–324.
Lee, M. D. (2011). How cognitive modeling can benefit from hierarchical Bayesian models. *Journal of Mathematical Psychology, 55*, 1–7.
Luck, S. J., & Vogel, E. K. (2013). Visual working memory capacity: From psychophysics and neurobiology to individual differences in cognitive ability. *Trends in Cognitive Sciences, 17*, 391–400.
Lei, J. (2017). Cross-validation with confidence. *Journal of the American Statistical Association*. https://doi.org/10.1080/01621459.2019.1672556.
Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature, 390*, 279–281.
Luria, R., Balaban, H., Awh, E., & Vogel, E. K. (2016). The contralateral delay activity as a neural measure of visual working memory. *Neuroscience & Biobehavioral Reviews, 62*, 100–108.
Ma, W. J. (2018). Problematic usage of the Zhang and Luck mixture model. *bioRxiv*.
Masson, M. E. J. (2011). A tutorial on a practical Bayesian alternative to null-hypothesis significance testing. *Behavior Research Methods, 43*, 679–690.
Matzen, L. E., Taylor, E. G., & Benjamin, A. S. (2011). Contributions of familiarity and recollection rejection to recognition: Evidence from the time course of false recognition for semantic and conjunction lures. *Memory, 19*, 1–16.
Markowitz, J., & Swets, J. A. (1967). Factors affecting the slope of empirical ROC curves: Comparison of binary and rating responses. *Perception and Psychophysics, 2*, 91–100.
Morey, R. D. (2011). A Bayesian hierarchical model for the measurement of working memory capacity. *Journal of Mathematical Psychology, 55*, 8–24.
Murphy, K. P. (2012). *Machine learning: A probabilistic perspective.* Cambridge, Massachusetts: MIT Press.
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.
Nosofosky, R. M., & Gold, J. M. (2017). Biased guessing in a complete-identification visual-working memory task: Further evidence for mixed-state models. *Journal of Experimental Psychology: Human Perception and Performance, 44*, 603–625.
Pashler, H. (1988). Familiarity and visual change detection. *Perception and Psychophysics, 44*, 369–378.
Pazzaglia, A. M., Dube, C., & Rotello, C. M. (2013). A critical comparison of discrete-state and continuous models of recognition memory: Implications for recognition and beyond. *Psychological Bulletin, 139*, 1173–1203.
Picard, R. R., & Cook, R. D. (1983). Cross-validation of regression models. *Journal of the American Statistical Association, 79*, 575–583.
Pitt, M. A., & Myung, I. J. (2002). When a good fit can be bad. *Trends in Cognitive Sciences, 6*, 421–425.
Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review, 107*, 358–367.
Rotello, C. M., & Macmillan, N. A. (2007). Remember-know models as decision strategies in two experimental paradigms. *Journal of Memory & Language, 55*, 479–494.
Rouder, J. N., Lu, J., Sun, D., Speckman, P., Morey, R., & Naveh-Benjamin, M. (2007). Signal detection models with random participant and item effects. *Psychometrika, 72*, 621–642.
Rouder, J. N., Morey, R. D., Cowan, N., Zwilling, C. E., Morey, C. C., & Pratte, M. C. (2008). An assessment of fixed-capacity models of visual working memory. *Proceedings of the National Academy of Sciences, 105*, 5975–5979.
Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin and Review, 16*, 225–237.
Schneegans, S., & Bays, P. M. (2018). Drift in neural population activity causes working memory to deteriorate over time. *Journal of Neuroscience, 38*, 4859–4869.
Schurgin, M. W., Wixted, J. T., & Brady, T. F. (2018). Psychophysical scaling reveals a unified theory of visual memory strength. *bioRxiv*.
Stretch, V., & Wixted, J. T. (1998). On the difference between strength-based and frequency-based mirror effects in recognition memory. *Journal of Experimental Psychology: Learning, Memory and Cognition, 24*, 1379–1396.
Swets, J. A. (1986). Form of empirical ROCs in discrimination and diagnostic tasks: Implications for theory and measurement of performance. *Psychological Bulletin, 99*, 181–198.
Turner, B. M., Forstmann, B. U., Wagenmakers, E. J., Brown, S. D., Sederberg, P. B., & Steyvers, M. (2013). A Bayesian framework for simultaneously modeling neural and behavioral data. *NeuroImage, 72*, 193–206.
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.
Wheeler, M. E., & Treisman, A. M. (2002). Binding in short-term visual memory. *Journal of Experimental Psychology: General, 131*, 48–64.
Wickens, T. D. (2001). *Elementary signal detection theory.* New York, NY, US: Oxford University Press.
Wilken, P., & Ma, W. J. (2004). A detection theory account of change detection. *Journal of Vision, 29*, 1120–1135.
Wixted, J. T. (2007). Dual-process theory and signal-detection theory of recognition memory. *Psychological Review, 114*, 152–176.
Wixted, J. T. (2019). The forgotten history of signal detection theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition.*. https://doi.org/10.1037/xlm0000732.
Unsworth, N., Fukuda, K., Awh, E., & Vogel, E. K. (2014). Working memory and fluid intelligence: Capacity, attention control, and secondary memory retrieval.

Cognitive Psychology, 71, 1–26.
van den Berg, R., Awh, & Ma, W. J. (2014). Factorial comparison of working memory models. *Psychological Review, 121*, 124–149.
van den Berg, R., & Ma, W. J. (2018). A resource-rational theory of set size effects in human visual working memory. *eLife,* 7. https://doi.org/10.7554/eLife.34963.
Van Zandt, T. (2000). How to fit a response time distribution. *Psychonomic Bulletin and Review, 7*, 424–465.
Varoquaux, G. (2018). Cross-validation failure: Small sample sizes lead to large error bars. *Neuroimage, 180*, 68–77.
Xie, W., & Zhang, W. (2017). Dissociations of the number and precision of visual short-term memory representations in change detection. *Memory & Cognition, 45*, 1423–1437.
Xu, Z., Adam, K. C. S., Fang, X., & Vogel, E. K. (2018). The reliability and stability of visual working memory capacity. *Behavioral Research Methods, 50*, 576–588.
Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Sciences, 12*, 1100–1122.
Zhang, W., & Luck, S. J. (2008). Discrete fixed-resolution representations in visual working memory. *Nature, 453*, 233–235.