Landscaping analyses of the ROC predictions of discrete-slots and signal-detection models of visual working memory

Chris Donkin, Sophia Chi Tran & Robert Nosofsky

Attention, Perception, & Psychophysics

ISSN 1943-3921

Atten Percept Psychophys
DOI 10.3758/s13414-013-0561-7

Atten Percept Psychophys
DOI 10.3758/s13414-013-0561-7

Landscaping analyses of the ROC predictions of discrete-slots and signal-detection models of visual working memory

Chris Donkin · Sophia Chi Tran · Robert Nosofsky

Abstract A fundamental issue concerning visual working memory is whether its capacity limits are better characterized in terms of a limited number of discrete slots (DSs) or a limited amount of a shared continuous resource. Rouder et al. (2008) found that a mixed-attention, fixed-capacity, DS model provided the best explanation of behavior in a change detection task, outperforming alternative continuous signal detection theory (SDT) models. Here, we extend their analysis in two ways: first, with experiments aimed at better distinguishing between the predictions of the DS and SDT models, and second, using a model-based analysis technique called landscaping, in which the functional-form complexity of the models is taken into account. We find that the balance of evidence supports a DS account of behavior in change detection tasks but that the SDT model is best when the visual displays always consist of the same number of items. In our General Discussion section, we outline, but ultimately reject, a number of potential explanations for the observed pattern of results. We finish by describing future research that is needed to pinpoint the basis for this observed pattern of results.

Keywords Visual working memory · Math modeling · Model selection

Visual working memory (WM) is the short-term memory system that maintains visual representations of stimulus inputs. It plays a fundamental role in a wide variety of ways in visual perception and cognition, including the ability to detect changes in scenes and to reason about visual displays (Luck & Hollingworth, 2008).

There is widespread consensus that visual WM is limited in its capacity (e.g., Luck & Vogel, 1997; Matsukura & Hollingworth, 2011). However, there is considerable ongoing debate concerning the basis for those capacity limits and the manner in which visual WM operates.

According to discrete-slots (DS) views, visual WM makes available some limited number of DSs for storing objects in memory (e.g., Awh, Barton, & Vogel, 2007; Barton, Ester, & Awh, 2009; Cowan, 2001; Luck & Vogel, 1997; Rouder et al., 2008; Zhang & Luck, 2008). Each slot is presumed to store a single to-be-remembered object. Furthermore, if an object is stored in one of the slots, it is stored with the maximum resolution that the system allows, regardless of the number of other objects in the to-be-remembered set. By contrast, if an object is not stored in one of the DSs, there is a complete loss of resolution for the object. Thus, the DS models are all-or-none in character and posit that a mixture of cognitive states (memory vs. no memory) governs performance in visual WM tasks.

A dramatically different view of visual WM is provided by continuous shared-resources models (e.g., Bays, Catalao, & Husain, 2009; van den Berg, Shin, Chou, George, & Ma, 2012; Wilken & Ma, 2004). According to these models, visual WM makes available some total pool of resources that is shared in continuous fashion across the members of a to-be-remembered set of objects. If the number of objects in the set is small, the observer can store high-resolution memory representations of all of them. As the number of objects in the set increases, there is a gradual and continuous decrease in the memory resolution associated with each individual object.

In this research, we pursue an approach that was used by Rouder et al. (2008) for trying to disentangle the predictions from DS versus continuous shared-resource models. These researchers used a well-known version of a visual WM task based on change detection (Luck & Vogel, 1997). At study,

C. Donkin (✉) · S. C. Tran
School of Psychology, University of New South Wales, Kensington, NSW 2052, Australia
e-mail: christopher.donkin@gmail.com

R. Nosofsky
Indiana University Bloomington, Bloomington, USA

Published online: 12 October 2013

observers were presented with a visual array of highly discriminable colored squares. At test, a single probe square was presented at one of the locations of the original array. The observer judged whether the color of the square at the probed location changed or stayed the same. Rouder et al. manipulated two independent variables in their task. The first variable was memory set size (M): the study array consisted of two, five, or eight squares. The second variable was change probability (π): In each block of trials, the color of the probed square changed with a probability of .3, .5, or .7.

Rouder et al. (2008) developed families of mathematical models to represent the DS and shared-resources views. We describe these families of models and their precise predictions in the Mathematical Models section of this article. The important conceptual point is that the two families of models made very different predictions regarding the structure of the probability of "change" judgments that one should observe in the task. In particular, following Rouder et al., we define the probability of a "hit" [p(h)] as the probability that an observer responds "change" when the color of a square does indeed change, whereas the probability of a "false alarm" [p(f)] is the probability that the observer responds "change" when the color of the square remains the same. As is explained in detail in the Mathematical Models section, the DS models predict that, for each set size condition (M), if one plots a receiver operating characteristic (ROC) curve by plotting p(h) against p(f) in each change-probability condition (π), the resulting ROC curves should be straight lines with slope equal to one. By contrast, the continuous shared-resources models predict that these ROC curves should have a curvilinear, bow-shaped form.

To assess these predictions, Rouder et al. (2008) fitted the members of the DS and shared-resources families to the individual-subject ROC data. Because individual models had differing numbers of free parameters, the researchers used a variety of model evaluation statistics as criteria of fit, including the Akaike information criterion (AIC; Akaike, 1974) and Bayesian information criterion (BIC; Schwarz, 1978). These statistics include terms that measure the absolute goodness of fit of a model but that also penalize a model for its number of free parameters. The model that yields the smallest AIC or BIC is considered to provide the most parsimonious account of the data. Rouder et al. found that a member from the DS family provided the best overall AIC and BIC fits to their individual-subject ROC curves, thereby providing support for the DS view of visual WM.

Although Rouder et al.'s (2008) results are intriguing and provide a considerable advance, there are some limitations associated with their study. One limitation is that it may be difficult to assess the linearity of the ROC curves in their study. On the basis of inspection of the averaged ROC curves reported in the study (Rouder et al., 2008, Fig. 2A), it seems likely that for many individual subjects, the three points on each ROC curve may have been located close together.[1] Under such conditions, it is exceedingly difficult to tell apart models that predict linear versus bow-shaped ROCs. Because the favored model from the DS family used fewer free parameters than did models from the shared-resources family, the statistical support for the DS model may have had more to do with its simplicity than with its ability to account for highly diagnostic data. In addition, although statistics such as AIC and BIC are extremely reasonable ones, they provide only approximations to ideal goals for model selection. The extent to which those statistics allow recovery of the "correct" model in any given situation needs to be carefully assessed.

In this research, our goal was to address these concerns by extending the original Rouder et al. (2008) investigation in two main ways. First, we tested a number of variants of the change detection task used by Rouder et al. Some of these variants included tasks with a greater number of change probability conditions (as well as more extreme change probability manipulations) than had been used by Rouder et al. The hope was to generate more challenging ROC data with greater diagnosticity than that obtained in the original study. Second, to investigate the extent to which the AIC and BIC statistics allow recovery of the correct models in these situations, we conducted what are known as *landscaping* analyses of the model-fitting results (Navarro, Pitt, & Myung, 2004). As is described in depth in the Model Analysis section, these techniques provide much deeper analysis of the relative merits of the competing models than simple listing of AIC and BIC fits alone.

Mathematical models

In this section, we describe some of the formal models that Rouder et al. (2008) used as representatives of the DS and continuous shared-resources views. In all cases, we presume that the models are applied to the type of change detection paradigm used by Rouder et al. In particular, (1) memory set size (M) is varied within blocks at different levels denoted i; (2) objective change probability (π) is manipulated between blocks at different levels denoted j; and (3) the stimuli are highly discriminable, so that when changes occur and items are stored in one of the DSs, the DS models presume that correct change responses are made with a probability of one.

Included among the members of the DS family in Rouder et al.'s (2008) investigation were fixed-capacity and variable-capacity DS models that assumed that performance across trials arises from a mixture of attentive and inattentive states.

[1] The data from our Experiment 1 (a replication of Rouder et al., 2008) are consistent with this notion; we found that for 65 % of participants, hit and false alarm rates differed across change proportion conditions by, at most, .3.

The representatives of the shared-resources family were equal- and unequal-variance signal detection theory (SDT) models. In this article, the main focus is on just two of the models, one from each family: the mixed-attention, fixed-capacity DS model and the equal-variance SDT model. Rouder et al. found that these models were the representatives from the DS and shared-resources families that best accounted for their data. It is worth noting that we fitted the complete battery of models to our data sets; however, like Rouder et al., we found that the fixed-capacity DS model and equal-variance SDT models performed best.

Discrete-slots model

The mixed-attention, fixed-capacity DS model assumes that participants have a fixed average number of "slots" into which items presented during study can be stored. (Although the number of slots may be variable across trials, the average number is presumed to be fixed across the different set size and change probability conditions.) This fixed average number of slots is represented by a capacity parameter k in the model. Let m_i denote the probability that a probed item is stored in one of the slots when memory set size (M) is at level i. Then, according to the fixed-capacity DS model, m_i is given by[2]

$$m_i = \min\left(1, {}^{k}/_{M}\right).$$

The model assumes that when a location is probed at test, if the item in that location was stored in memory, then a perfect *change* versus *same* classification is made (because the colored squares are highly discriminable in the Rouder et al., 2008, paradigm). However, if the item in the probed location was not stored in memory, the observer must guess. The probability that the observer guesses *change* is determined by a guessing parameter g. Adopting the assumption of selective influence, the value of g is presumed to depend only on the level of change probability (π) that operates in a block of trials. Thus, when π is at level j, the observer guesses *change* with probability g_j. Presumably, as the level of change probability that operates within a block of trials increases, observers will tend to guess *change* with higher probability.

Finally, Rouder et al. (2008) found that it was necessary to include an *attention* parameter, a, into the model to allow for imperfect performance when $M < k$ (people reliably make a small number of errors on even the simplest of trials). It is assumed that on some proportion of trials ($1-a$), the observer fails to encode the visual display and so must guess. The guess process due to inattention is assumed to be identical to that used when an item is not in memory (i.e., the observer guesses *change* with probability g_j).

Combining the assumptions outlined above, the DS model predicts the following for hit rates $p(\mathrm{h})$ and false alarm rates $p(\mathrm{f})$ for the ith study set size and jth change probability condition:

$$p\left(h_{ij}\right) = am_i + a(1-m_i)g_j + (1-a)g_j \tag{1a}$$

$$p\left(f_{ij}\right) = a(1-m_i)g_j + (1-a)g_j. \tag{1b}$$

From inspection of Eqs. 1a and 1b, it is straightforward to see that, for any given set size condition i, the DS model predicts linear ROC (isosensitivity) curves with slope equal to one and y-intercept equal to am_i. Also, solving Eqs. 1a and 1b for m_i and then eliminating that term reveals that the DS model also predicts linear isobias curves of the form $p\left(h_{ij}\right) = 1 + \left(1 - {}^{1}/_{g_j}\right)p\left(f_{ij}\right)$. (Example predictions from the model are shown as gray lines in the top panels of Fig. 1.) This DS model estimates a capacity parameter k, an attention parameter a, and G guessing parameters g_j, where G is the number of change probability conditions.

Signal detection model

An alternative to the fixed-capacity DS model is the continuous SDT model, which assumes that the study display is remembered with varying degrees of precision as the number of items in the display changes. When a location is probed with a test item, the item evokes a certain level of familiarity in the observer, x, who then uses a criterion β to determine whether the item has changed or is the same. It is assumed that items that do not change from study to test evoke a distribution of familiarities represented by a standard normal. When an item does change from study to test, the familiarity is represented by a normal distribution with mean d' (and variance 1 in the equal-variance version we focus on here). As such, when a test item evokes a particular level of familiarity, x, the participant compares the likelihood of the familiarity under the two possible situations:

$$LR(x) = \frac{\phi(x-d')}{\phi(x)}, \tag{2}$$

where φ is the density of the standard normal distribution. If the likelihood ratio is above the criterion β, then the observer responds *change* and otherwise responds *same*.

Rouder et al. (2008) presented the equations for calculating hit and false alarm rates from the model. If one assumes that manipulation of study set size influences only the memory strength d' and the change probability manipulation affects only the criterion β, the hit and false alarm rate predictions for

[2] The model presumes that when memory set size (M) is less than the average number of slots (k), all items are stored with a probability of 1.

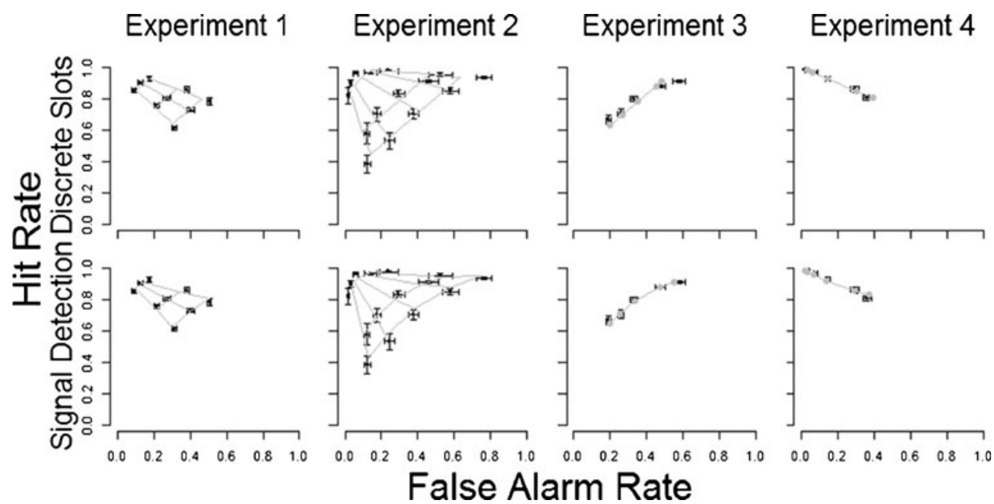

Fig. 1 ROC curves averaged over participants for each of the four experiments (columns). The averaged predictions of the discrete-slots model (top row) and the signal detection theory model (bottom row) are shown in gray, whereas the data (which are the same in both rows of plots) are shown in black. Error bars are standard errors across individuals

the ith study set size and jth change probability conditions are as follows:

$$p(h_{ij}) = \Phi\left(\frac{d'_i}{2} - \frac{\log\beta_j}{d'_i}\right) \quad (3a)$$

$$p(f_{ij}) = \Phi\left(\frac{-d'_i}{2} - \frac{\log\beta_j}{d'_i}\right), \quad (3b)$$

where Φ is the cumulative distribution function of the standard normal distribution.

As is well known, the equal-variance SDT model predicts curvilinear bow-shaped ROC curves that are symmetric about the negative diagonal. (Example predictions from the model are shown as gray lines in the bottom panels of Fig. 1.) The model estimates S sensitivity (d') parameters and G response bias (β) parameters, where S is the number of set size conditions and G is the number of change probability conditions.

Experiments

We conducted four experiments that were variants of the change detection task conducted by Rouder et al. (2008). Experiment 1 was a minor variant of the Rouder et al. task, with three different memory set size conditions (M = 3, 5, 8) crossed with three different change probability conditions (π = .3, .5, .7). In addition, although the manipulation turned out to have little effect, within that experiment we conducted both an external-change and an internal-change condition. In the external-change condition, on change trials, the probe square was different in color from any of the colors in the study array. In the internal-change condition, on change trials, the probe square was the same color as another square from the study array. In Experiment 2, the values of M were {2, 5, 8}, and in Experiment 3, the value of M was held fixed at 6. In these experiments, five different change probability values were tested: π = .15, .3, .5, .7, and .85 for Experiment 2, and π = .14, .3, .5, .7, and .86 for Experiment 3. The goal was to construct isosensitivity curves with a greater number of points and, perhaps, a greater span in hit and false alarm rates than in the previous experiment. Finally, in Experiment 4, we tested six different set size conditions (M = 1, 2, 3, 4, 6, 8), with change probability held fixed at π = .5. The goal was to produce an isobias curve with a greater number of points than obtained in the previous study. In Experiments 2–4, change trials always involved external changes. A summary of the design of the four experiments is provided in Table 1.

Method

Participants

Ninety-nine, 20, 44, and 30 participants took part in Experiments 1–4, respectively. Participants in Experiment 1 were from Indiana University, while participants from the remaining experiments were tested at the University of New South Wales. Participants in Experiment 2 were reimbursed $15 per session for four sessions, while the remaining participants received course credit for a single session. An additional participant completed just two sessions of Experiment 2 and was therefore excluded from analysis.

Stimuli

Stimuli were a set of 10 color squares (white, black, red, blue, green, yellow, orange, cyan, purple, and dark blue-green) presented on a gray background. Experiment 1 utilized 17-in. CRT monitors, while Experiments 2–4 used 24-in. LCD

Table 1 Designs for the four experiments

Experiment	Study Set Sizes	Change Probabilities	Number of Participants	Trials per Cell	Number of Parameters	
					DS	SDT
1	3,5,8	.3,.5,.7	99	18–42	5	6
2	2,5,8	.15,.3,.5,.7,.85	20	24–136	7	8
3	6	.14,.3,.5,.7,.86	44	14–86	6	6
4	1,2,3,4,6,8	0.5	30	84	3	7

Note. DS, discrete slot; SDT, signal detection theory

monitors. Each color square was 0.75° × 0.75° in size. Stimuli were presented within an array whose visual angle was within a 9.8° × 7.3° rectangle. Items were presented in randomly chosen locations, with the restriction that they had to be at least 2° away from any other item and from the center of the viewing area. The cue indicating the position of the probe color square was a black, 1 pixel thick, 1.5° diameter circle that surrounded the probe color square.

Procedure

Trials began with a fixation cross for 1,000 ms. A study array of *M* color squares was then presented for 500 ms. After a 500-ms blank screen, a multicolored pattern mask was presented at each study location for 500 ms. A single test color was then presented in one of the locations of the study array (this location was also marked by the circular cue that we described earlier). The test color was either the same as the item in that location during study (a *no-change* trial) or different from the item in that particular location of the study array (a *change* trial). The participant was asked to indicate whether the test item was the same as the study item or had changed by pressing "F" or "J" on the keyboard, respectively. The participants then received feedback on their response for 1,000 ms, and the next trial then began after a 1,000-ms blank screen.

At the start of each block, participants were told explicitly how often the study item would change at test in the upcoming block of trials. The change probability conditions were presented in random order, with the constraint that all conditions were experienced before one was repeated. An equal number of each study set size condition, for both change and no-change trials, was presented in random order within each block of trials. Due to differences in design, each experiment yielded a distinct number of trials per cell of the experiment (i.e., where a cell corresponds to a combination of study set size and change probability condition). Note that because different change probability conditions necessarily produce more or fewer change or no-change trials, we report a range of trials per cell for each experiment. Participants in Experiment 1 completed 9 blocks of 60 trials, yielding between 18 and 42 trials per cell per participant. Those in Experiment 2 completed four sessions of 10 blocks of 60 trials, yielding between 24 and 136 trials per cell per participant. In Experiment 3, there were 10 blocks of 50 trials, yielding between 14 and 86 trials per cell per participant. Finally, Experiment 4 utilized 6 blocks of 84 trials, yielding 84 trials per cell per participant.

Results

Data censoring

We used the following procedure to censor data from each experiment. First, in Experiment 1, we excluded 2 participants whose accuracy was below 60 % and mean response time (RT) was less than 500 ms (mean RT for all other participants was close to 1 s). Second, we removed the first 50 trials of the first block of each experiment, since participants were still learning to do the task. Third, we removed trials that were unrealistically fast or slow by excluding trials on which RT was shorter than 180 ms or longer than 4 s. This exclusion led to 1.8 %, 4.3 %, 2.9 %, and 0.2 % of trials being removed from Experiments 1–4, respectively. Finally, for each participant, we excluded any trial on which the RT was longer than 2.5 standard deviations above the mean RT. This procedure led to a further exclusion of 2.7 %, 1.8 %, 2.5 %, and 2.8 % of trials in Experiments 1–4, respectively.

Hit rates and false alarms

Before turning to a model-based analysis of the data, we briefly discuss the ROC data from each experiment. The averaged ROC data from each experiment are displayed in Fig. 1 (black symbols). As can be seen from the figure, across all four experiments, manipulations of memory set size and objective change probability had the expected effects on the general patterns of performance. Increases in memory set size led to decreases in overall performance (decreasing hit rates and increasing false alarm rates). Increases in objective change probability led to increasing proportions of *change* responses (increasing hit rates and increasing false alarm rates).

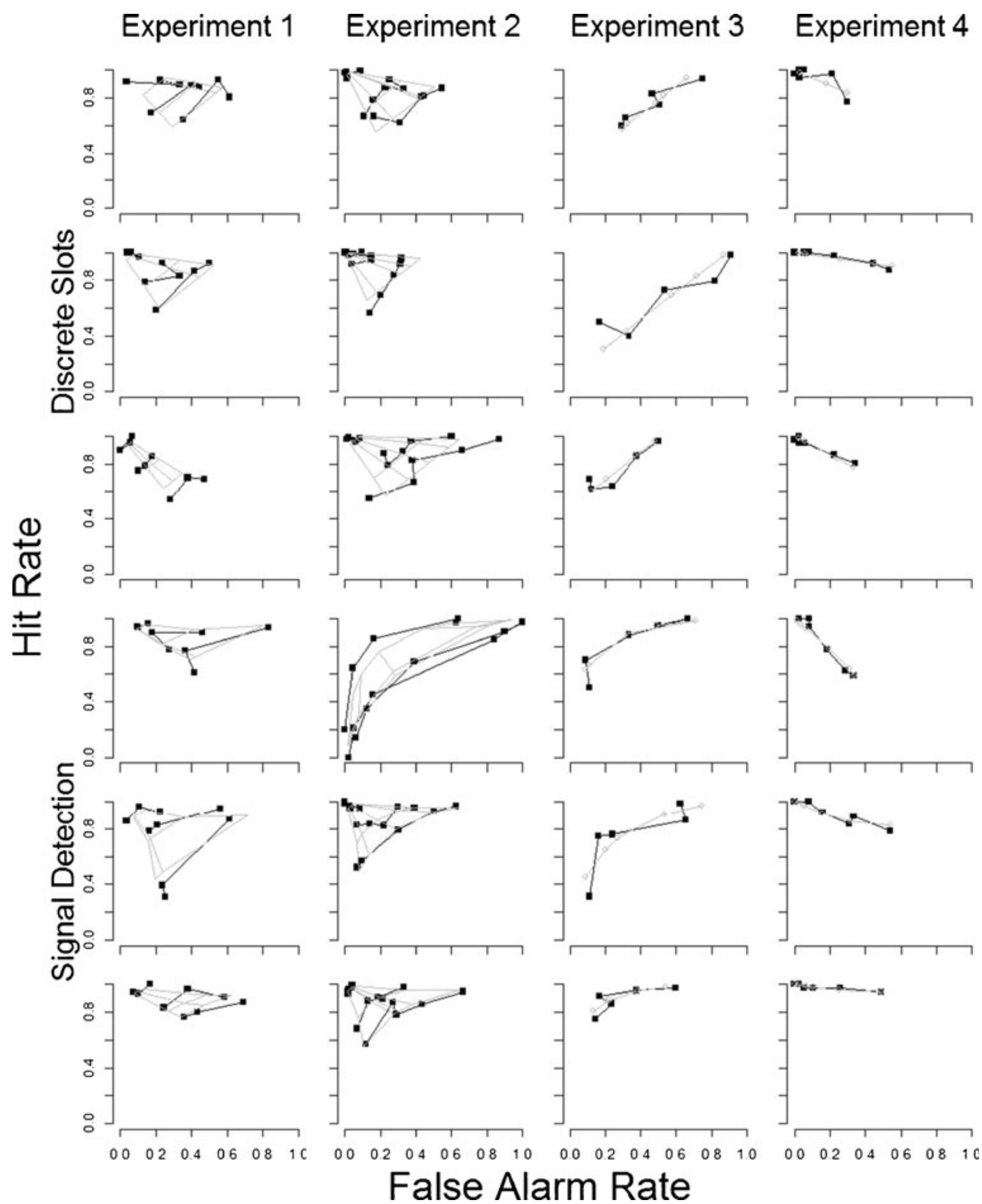

Fig. 2 ROC curves for 6 individual participants (rows) in each of the four experiments (columns). Data are shown in black, and model predictions are shown in gray. The top three rows show predictions from the discrete-slots model, and the bottom three rows show predictions from the signal detection theory model

Our purpose in displaying these averaged ROC curves is simply to document the overall trends in performance. However, presentations of averaged ROC curves can be misleading. For example, a single curvilinear ROC participant will make the average ROC curvilinear even if all other ROC curves are linear. In an attempt to highlight individual differences, Fig. 2 contains the ROC results for 6 participants from each of the four experiments (the remaining participants' data and model fits are available upon request). Three of these representative participants' data were better fit by the DS model, and three were better fit by the SDT model (see the Model Analysis section). As can be seen, the data patterns at the level of individual subjects tend to be noisy, and there can be large individual differences.

Modeling analysis

Parameter estimates We fitted the fixed-capacity DS model (Eqs. 1a and 1b) and the equal-variance SDT model (Eqs. 3a and 3b) to the data of each individual participant across all four experiments. As has been outlined earlier, the DS model has an attention parameter, a, a capacity parameter, k, and a separate guess parameter, g_j, for each of the change probability conditions in the experiment. The SDT model has a d'

parameter for each of the study set size conditions and a β parameter for each of the change probability conditions. The number of parameters used by each model in each experiment is reported in Table 1.

We fitted the models to each individual's data using maximum likelihood estimation (for details, see Rouder et al., 2008). Best-fitting parameters were estimated using a SIMPLEX algorithm started from 100 different random starting points. Within each of the 100 searches, we also restarted the SIMPLEX with the current best-fitting parameters multiple times, since this broadens the search and thus avoids local maxima.

Table 2 reports the best-fitting parameters for the two models averaged across participants in each experiment. According to the DS model, participants have a capacity of around three items and attend to the visual display on about 85 % of trials. Furthermore, their probability of guessing *change* increases systematically with increases in objective change probability. Overall, participants are biased toward guessing *change* rather than *no change*. The parameter estimates from the SDT model have similar implications for behavior: Participants become less able to discriminate between *change* versus *no-change* trials (i.e., d' decreases) as the number of study items increases. In addition, their criterion settings for making *change* responses become increasingly lax as objective change probability increases.

As was mentioned earlier, there were considerable individual differences in performance and, therefore, wide ranges of individual-participant parameter estimates. Table 3 reports the range of parameter values across individuals within each experiment.

Table 2 Best-fitting parameters, averaged over participants, for discrete-slots (DS) and signal detection theory (SDT) models for the four experiments

	DS					SDT			
Experiment	1	2	3	4		1	2	3	4
k	3.36	2.80	2.64[1]	3.52	d'_1	–	–	–	4.50[2]
a	0.80	0.88	–	0.95	d'_2	–	3.25	–	4.50[2]
$g_{0.15}$	–	0.22	0.36	–	d'_3	2.64	-	-	3.76
$g_{0.3}$	0.47	0.36	0.48	–	d'_4	–	–	–	2.74
$g_{0.5}$	0.59	0.60	0.62	0.68	d'_5	1.53	1.57	–	–
$g_{0.7}$	0.70	0.80	0.79	–	d'_6	–	–	1.33	1.75
$g_{0.85}$	–	0.90	0.84	–	d'_8	0.92	1.00	–	1.36
					$\beta_{0.15}$	–	2.05	1.46	–
					$\beta_{0.3}$	1.12	1.34	1.08	–
					$\beta_{0.5}$	0.92	0.82	0.79	0.66
					$\beta_{0.7}$	0.85	0.53	0.53	–
					$\beta_{0.85}$	–	0.33	0.42	–

[1] Note that since the a parameter is not identifiable with just one set size, this capacity is likely to be an underestimate of the true capacity value.

[2] This value is the ceiling placed on the d' parameter.

Table 3 The minimum and maximum best-fitting parameters across participants for discrete-slots (DS) and signal detection theory (SDT) models for the four experiments, rounded to the nearest .05

	DS					SDT			
Experiment	1	2	3	4		1	2	3	4
k	1.20	1.20	0.60	2.00	d'_1	–	–	–	3.00
	6.80	5.60	5.40	5.60					4.50
a	0.45	0.60	–	0.80	d'_2	–	1.50	–	2.60
	1.00	1.00		1.00			4.50		4.50
$g_{0.15}$	–	0.00	0.00	–	d'_3	1.00	–	–	1.90
		0.45	0.65			4.50			4.50
$g_{0.3}$	0.00	0.10	0.10	–	d'_4	–	–	–	1.50
	0.80	0.60	0.90						4.50
$g_{0.5}$	0.25	0.40	0.35	0.35	d'_5	0.50	0.60	–	–
	0.85	0.80	0.96	0.95		3.00	3.10		
$g_{0.7}$	0.40	0.55	0.45	–	d'_6	–	–	0.10	0.75
	1.00	1.00	1.00					3.00	3.10
$g_{0.85}$	–	0.70	0.55	–	d'_8	0.30	0.50	–	0.60
		1.00	1.00			2.00	1.90		2.20
					$\beta_{0.15}$	–	0.95	0.50	–
							4.50	2.50	
					$\beta_{0.3}$	0.40	0.95	0.45	–
						2.50	2.50	2.10	
					$\beta_{0.5}$	0.40	0.40	0.30	0.25
						1.60	1.20	1.30	1.35
					$\beta_{0.7}$	0.30	0.20	0.10	–
						1.20	0.95	1.10	
					$\beta_{0.85}$	–	0.10	0.05	–
							0.55	0.90	

Model fits We assessed the parsimony of the fit of each model by using the AIC and BIC statistics. Both model selection criteria become smaller as the quality of the agreement between model and data improves but get larger as the number of parameters in the model increases. As such, the model with the smaller criteria gives a more parsimonious account of the data. The criteria are formalized as: $AIC = 2p - 2l$ and $BIC = p\log(n) - 2l$, where l is the (maximum) log- likelihood of the parameters given the data, p is the number of free parameters in the model, and n is the number of data points being fit by the model. Note that BIC is more penalizing of models with extra complexity than is AIC.

In Table 4, we report the number of participants better fit by the DS and the SDT models using AIC and BIC. Using BIC, the majority of participants in Experiments 1, 2, and 4 are best fit by the DS model. The sum of BIC across participants in those experiments also favors the DS model. The opposite is true for Experiment 3, where the SDT model provides a better fit to the

Table 4 The number of participants, *N*, in each experiment selected using either the Akaike information criterion (AIC) or the Bayesian information criterion (BIC), the proportions of participants better fit by each model (in parentheses), and the sums of AIC and BIC values (Σ)

		AIC				BIC			
		N		Σ		N		Σ	
		DS	SDT	DS	SDT	DS	SDT	DS	SDT
Experiment	1	58 (.60)	38 (.40)	41,923	41,899	85 (.89)	11 (.11)	43,915	44,289
	2	9 (.45)	11 (.55)	29,367	29,330	13 (.65)	7 (.35)	30,159	30,235
	3	13 (.29)	31 (.71)	17,906	17,840	13 (.29)	31 (.71)	17,906	17,840
	4	22 (.73)	8 (.27)	7,735	7,806	30 (1.00)	0 (0.00)	8,104	8,666
	Total	102 (.54)	88 (.46)	96,931	96,875	141 (.74)	49 (.26)	100,084	101,030

Note. DS, discrete slot; SDT, signal detection theory

majority of participants (this is true for both AIC and BIC, which yield identical results in this experiment, since the models have the same number of free parameters). With the exception of Experiment 4, for which the DS model is still favored, the results are less clear when using AIC to do model selection.

Figure 3 contains a histogram of the differences between (negative) 2× log-likelihood, AIC, and BIC for the two models for each participant in the four experiments. A negative value in this plot corresponds to a participant for which the DS model has a better fit than the SDT model. The histograms are generally reflective of the pattern observed in Table 4 but show that the evidence for one model over the other lies on a continuum, even within a single experiment. That is, the evidence never appears to be overwhelmingly supportive of one model over the other (except in the case of BIC in Experiment 4). Taken together, the results of the four experiments provide a rather mixed message regarding whether one should prefer the DS or the SDT model.

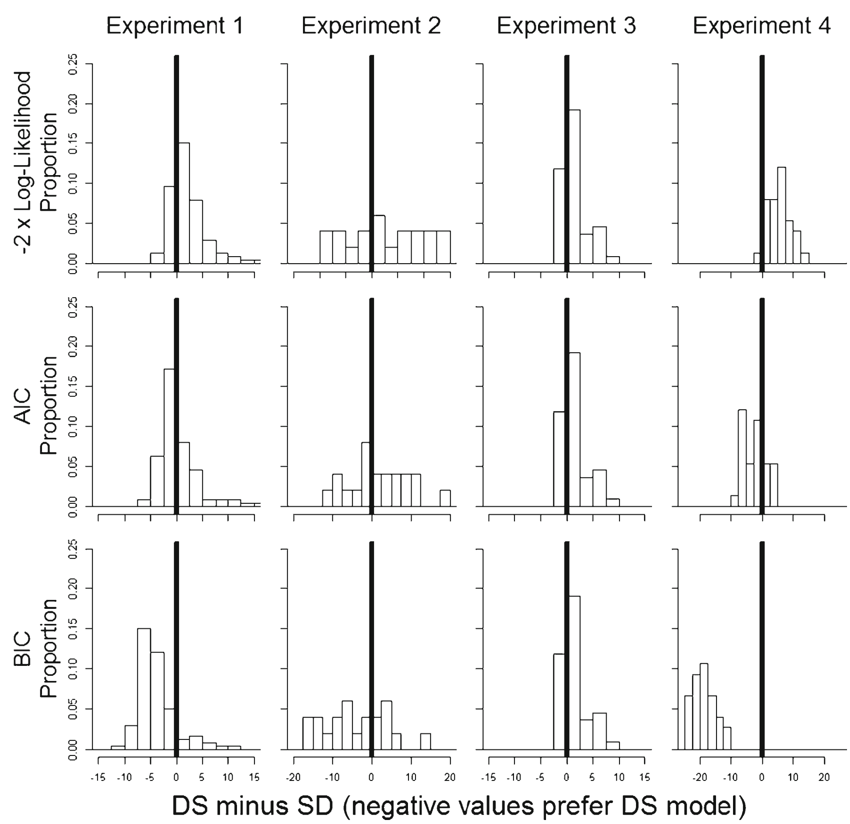

Fig. 3 Histograms of the difference between participants' (negative) 2× log-likelihood, Akaike information criterion (AIC) and Bayesian information criterion (BIC) (rows) values for each of the four experiments (columns). Negative values indicate that the discrete-slots (DS) model outperforms the signal detection theory (SDT) model. The vertical line at 0 on each plot indicates the point at which both models perform equally well

Landscaping analysis

Although AIC and BIC are often-used model selection tools, they rely solely on the number of free parameters as a proxy for complexity. However, a simple parameter count does not take into account the fact that not all parameters in a model have the same influence on the number of data patterns that the model can produce—a concept known as functional form complexity (Myung, 2000; Pitt, Kim, Navarro, & Myung, 2006; Pitt & Myung, 2002; Pitt, Myung, & Zhang, 2002). We now use the landscaping technique to compare the DS and SDT models, taking into account functional form complexity (Navarro et al., 2004; for a related method, see Wagenmakers, Ratcliff, Gomez, & Iverson, 2004).

In short, landscaping asks whether the fits of Model A and Model B to observed data are more likely under one model than under the other. The technique requires that one simulate a large number of data sets from Model A and then fit those data sets with both Model A and Model B. The same process is repeated, but this time the simulated data sets come from Model B. For each data-generating model, one creates a "landscape" by plotting the fits to the simulated data from Model A against fits from Model B. Upon each of the landscapes from Model A and Model B, one then plots the fits of the two models to the empirical data. If the empirical fits are more likely to have been produced under one landscape than under the other, it follows that the model that generated that landscape is more likely to be the true model.

We began by simulating 10,000 data sets from each of the DS and the SDT models for each of the four experiments (for a total of 80,000 data sets). Each simulated data set was generated to have the same design (i.e., trial numbers, study set size conditions, and change probability conditions) as the original experiment. For example, the data sets intended to mimic Experiment 1 had 60 trials per study set size condition, with 0.3, 0.5, or 0.7 of those trials being *change* trials. The k, a, and d' parameters used to simulate data sets were random draws from the respective range of parameters reported in Table 3. The g and β parameters for each data set were constructed by sampling parameters from the smallest and largest change probability conditions (from Table 3) and then creating a set of G equally spaced g or β parameters, where G was the number of change probability conditions in the given experiment. That is, the guessing and bias parameters of the models were constrained to increase with the change probability condition. The simulated number of hits and false alarms in each data set were a draw from a binomial distribution with probability governed by the predictions from the model given the sampled parameters.

We then fit each of the 80,000 simulated data sets with the DS and SDT models. The method for fitting the simulated data sets was identical to the way that the models were fit to the empirical data (i.e., the same objective functions, search algorithms, and start points).

Before inspecting the landscapes, we first performed a brief model recovery analysis, asking how often the model that provided the best fit to the simulated data was the data-generating model. Table 5 reports the proportion of data sets for which each model was the best-fitting model in cases in which either the DS or the SDT model was the true data-generating model. When the DS model is the true model, BIC generally chooses it with high probability, and AIC chooses it with somewhat lower probability. When the SDT model is the true model, AIC generally chooses it with fairly high probability, but BIC often fails to choose the correct model. In general, BIC tends to be biased toward selecting the DS model over the SDT model. On the other hand, AIC tends to select more equivocally between the DS and SDT models.

Our recovery analysis suggests that using just one of the selection criteria to decide between the two models is not necessarily reliable. This result is particularly true for BIC, which seems to have punished the SDT model too harshly for its extra parameter(s). The analysis also suggests that when AIC is used, it will be difficult to achieve unequivocal evidence for either model in all but Experiment 2. These model recovery results provide a strong rationale for using the landscape technique that we now apply.

Landscaping results Figure 4 contains the landscape plots for the DS (top row) and SDT (bottom row) models for each experiment (columns). Each of the gray dots in the landscapes shows the log-likelihood fit of the SDT model (vertical axis) and DS model (horizontal axis) to one of the 10,000 simulated data sets. The log-likelihood values for fits to the empirical data from each experiment are shown by the black dots. It should be noted that landscapes produced by plotting AIC or BIC values would be identical in form to the log-likelihood plots shown in the figure, with all points identically translated horizontally and vertically according to the specific penalty term that is used. Although the absolute location of the landscape would shift, the relative location of the black dots (the empirical fits) within the simulated gray dots would remain the same.

It can be quickly gleaned from Fig. 4 that for Experiments 1, 2 and 4, the patterns of log-likelihood values for the empirical data look more like the landscapes produced when the DS model generated the data than when the SDT model generated the data. That is, for Experiments 1, 2, and 4, the black points in the figure look more like the gray points in the top row of plots than in the bottom row of plots. The major problem for the SDT model is that, in cases in which it is the true generating model (bottom panels), it produces many data sets for which it provides a considerably better fit than the DS model (gray points that lie on the upper-left side of the

Table 5 Proportions of 10,000 simulated participants better fit by each model in cases in which the DS model is true or the SDT model is true. Note that the left columns give model selection results using AIC and the right columns give model selection results using BIC

		DS Model Is True			
		AIC		BIC	
		DS	SDT	DS	SDT
Experiment	**1**	.76	.24	.93	.07
	2	.93	.07	.99	.01
	3	.63	.37	.63	.37
	4	.88	.12	1.00	.00
		SDT Model Is True			
		AIC		BIC	
		DS	SDT	DS	SDT
Experiment	**1**	.17	.83	.36	.64
	2	.04	.96	.14	.86
	3	.40	.60	.40	.60
	4	.24	.76	.78	.22

Note. DS, discrete slot; SDT, signal detection theory

main diagonal). The empirical data points rarely demonstrate this pattern and, instead, tend to lie close to the main diagonal where the two models provide more nearly equivalent fits to human data. By contrast, for Experiment 3, the landscapes may favor the SDT model. In particular, when the DS model is true, the empirical results (the black dots) appear to lie to the upper left of the landscape. The empirical results for Experiment 3 appear to lie more comfortably within the SDT model landscape.

To confirm our visual impression of the landscape results in Fig. 4, we calculated the *landscape-based* likelihood of the empirical data (i.e., the black dots) in each panel (cf. Gilden, 2009; Navarro et al., 2004). To apply this procedure, we first applied a kernel density estimate to the log-likelihood values in each of the landscapes, using the **ks** package in R.[3] This procedure yields a landscape-based likelihood estimate of each of the empirical data points. Figure 5 plots the log likelihood of each participant's data in the landscape where the SDT model was true (vertical axis) and the landscape where the DS model was true (horizontal axis). In Experiment 1, the vast majority of participants (79 %) fall to the lower-right of the main diagonal, indicating that the data were more likely under the DS landscape. The same is also true for Experiments 2 and 4, where 70 % of participants' data were more likely if the DS model were true. By contrast, in Experiment 3, the data are more likely under the SDT model, since the SDT model yields a higher landscape-based likelihood for 75 % of the participants.

[3] We report the results of the smoothed cross-validation method for creating the bandwidth matrix of smoothing parameters. We made the simplifying assumption that this matrix was a diagonal matrix. All other settings of the kernel density estimator were left at default. We repeated our analysis with a number of alternative settings and smoothing matrices, but our conclusions remained unchanged.

Discussion

Summary of results and potential explanations

The purpose of this research was to pursue further the ROC-based methods used by Rouder et al. (2008) for disentangling the predictions from DS versus continuous shared-resources models of visual WM change detection. Our first approach was to conduct more varied conditions than had been tested by Rouder et al. Our new conditions included ones involving more extreme manipulations of objective change probability than had been used by Rouder et al., so that the straight-line versus curvilinear ROC predictions from the DS and continuous models might be more easily discriminated. Our second approach was to use landscaping analyses of the predictions from the models. This approach was intended to provide a deeper analysis of the relative merits of the competing models' predictions than could be achieved by listing of penalty-corrected fits alone.

Considered collectively across our four experiments, our results did not point to a clear-cut winning model. On the one hand, in our Experiment 1, which was a near replication of the Rouder et al. (2008) design, the pattern of model-fitting results was extremely similar to that reported in the original study. Moreover, our landscaping analyses suggested that the DS model was indeed the preferred model for that experiment, confirming the model selection results yielded by use of the AIC and BIC statistics. Furthermore, in Experiment 2, we extended the design of the original experiment by using five change probability conditions instead of three (and by using more extreme change probability manipulations). In this extended experiment as well, the landscaping analyses pointed toward the DS model as the preferred model. In Experiment 4, we tested a single change probability condition, but a broader range of set size conditions than in the original study. The purpose was to produce an isobias curve based on more than three points to test the linear isobias curve prediction from the DS model. In this experiment as well, the landscaping analyses favored the DS model, as compared with the continuous alternative.

The "oddball" experiment turned out to be Experiment 3, in which we tested a single memory set size condition but five different change probability conditions. In this testing situation, the landscaping analyses pointed strongly toward the continuous shared-resource model rather than the DS model.

In the remainder of this section, we consider several different explanations for why the Experiment 3 results point in the opposite direction from those of the other experiments. First, in the original Rouder et al. (2008) article, as well as the

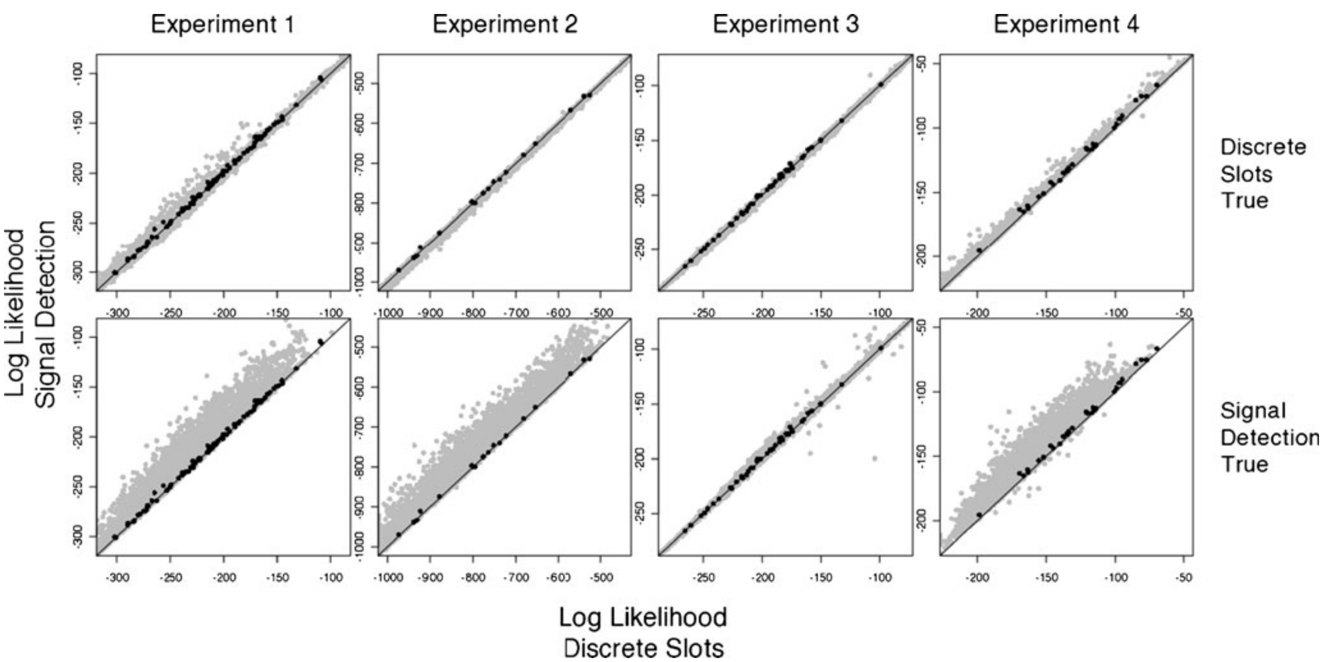

Fig. 4 Landscapes using log-likelihood for cases when the discrete-slots (DS) model is true (top row) and when the signal detection theory (SDT) model is true (bottom row), for each of the four experiments (columns). The log-likelihood fits of the SDT model (vertical axis) and the DS model (horizontal axis) to the simulated data are shown as gray points. The empirical fits are shown as black points

present extension, the focus was on the functional form of the predicted ROC curves: The DS model predicts linear isosensitivity and isobias ROCs, whereas the continuous SDT model predicts curvilinear ones. In hindsight, however, upon considering the Experiment 3 results, it becomes salient that there are other important differences between the representatives of the DS and continuous families that were formalized by Rouder et al. One of these differences is that the favored version of the DS model makes more specific predictions than does the SDT model regarding how overall "sensitivity" will vary across the different memory set size conditions (M). Recall that the favored DS model (i.e., the fixed-capacity version) assumes that the probability of entering the "memory state" in set size condition i is given by

$$m_i = a \cdot \min\left(1, {}^{k}\!/_{M}\right)$$

where a is the attention parameter and k is the capacity parameter. Thus, the DS model uses two free parameters to predict "sensitivity" across S different set size conditions. By contrast, the versions of the SDT models that were formalized did not include these types of constraints. Instead, both the equal- and unequal-variance SDT models use S d' parameters, one for each separate set size condition. Conceivably, therefore, in the

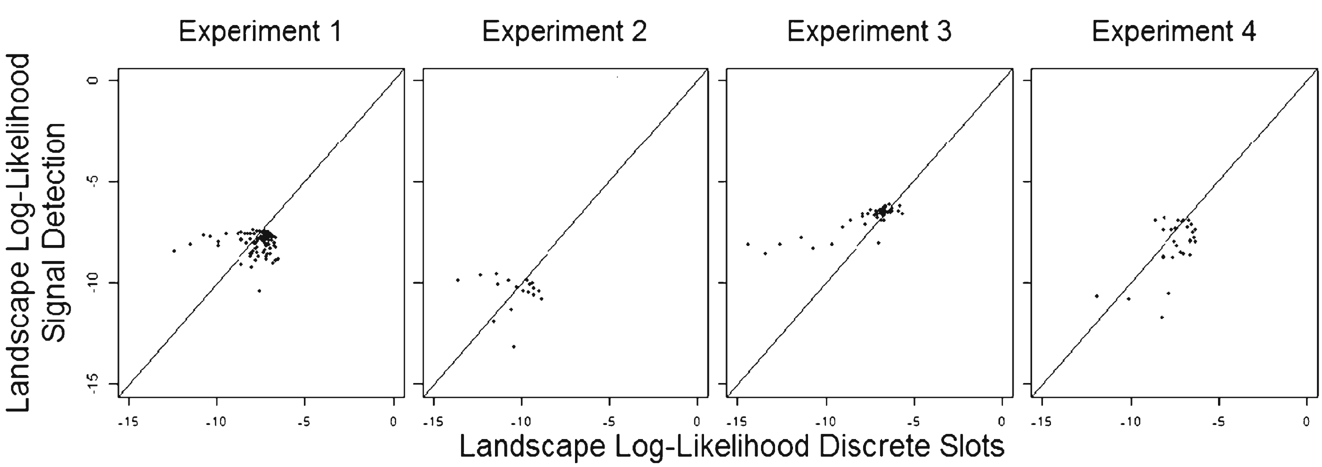

Fig. 5 The landscape-based log-likelihood of the empirical fit values when the signal detection theory model was true (vertical axis) and when the discrete-slots (DS) model was true (horizontal axis) for each of the four experiments. Points on the lower right of the main diagonal represent participants whose data are more likely under the DS model

experiments involving multiple set sizes (i.e., Experiments 1, 2, and 4), the better performance of the fixed-capacity DS model may reflect its more parsimonious predictions of how sensitivity varies with set size, rather than its predictions of the functional form of the ROCs. By contrast, in Experiment 3, there was only a single set size, so this potential source of the advantage of the fixed-capacity DS model disappears.

To evaluate this possibility, we conducted three follow-up modeling analyses, using the data from Experiment 2. (We chose Experiment 2 for analysis because it was identical to Experiment 3 but included more than one set size.) The full details of these analyses are given in the Appendix, but the aim was to compare the DS and SDT models once they were equated on the parsimony of their explanation of set size effects. The logic behind our analyses was that if the DS model was favored in Experiment 2 because it provided more parsimonious predictions of set size effects, that advantage might disappear if the models were equated on that factor. The results of the modeling analyses reported in the Appendix, however, provided no evidence to support the hypothesis.

A second possibility involves the response bias assumptions that we and Rouder et al. (2008) imposed on the SDT models. Recall that Rouder et al. adopted the *likelihood-ratio* version of SDT, in which the observer was assumed to respond *change* when the likelihood-ratio statistic (Eq. 2) exceeded a criterion β. Adopting the assumption of selective influence, the magnitude of β was assumed to depend only on the level of change probability and to be invariant with set size. However, it is a wide-open question what form of response bias may depend on only change probability and be invariant with set size. Conceivably, the relatively poor fits of the SDT model in Experiments 1, 2, and 4 could reflect inappropriate assumptions concerning the nature of response bias. Again, because only a single set size was tested in Experiment 3, this potential shortcoming of the SDT model disappears in that experiment.

In a preliminary attempt to evaluate this possibility, we fitted an SDT model to the Experiment 2 data that made alternative assumptions about response bias. The full details of this alternative SDT model are given in the Appendix, but in short, the model fitted our Experiment 2 data considerably worse than did the likelihood-ratio version used by Rouder et al. (2008) and never yielded a better BIC fit. Again, these preliminary analyses fail to provide support for the response bias explanation of our pattern of results.

A third possibility, suggested by a reviewer, is that the *diagnosticity* of the data may be the underlying cause for the atypical results in Experiment 3. The diagnosticiy of an individual's data is indexed by how often the data-generating model would be recovered on the basis of that individual's best-fitting parameters (for both models). The idea is that nondiagnostic data may be preferentially accounted for by a particular model. For example, we would observe our pattern of results if the data from Experiment 3 were nondiagnostic and if the SDT model tended to be recovered when results were nondiagnostic. Jang, Wixted, and Huber (2011) found that the diagnosticity of the data was a key factor in their comparison of SDT and dual-process models of long-term recognition memory. We repeated their analysis on our data and models but found little evidence to support this diagnosticity hypothesis.

A final potential explanation for our pattern of findings, which would need to pursued in future empirical work, is that the mode of processing in visual WM change detection tasks varies with experimental conditions. In particular, Experiment 3 was unique in that observers knew in advance on each trial the precise number of items that would be presented in the visual display. Perhaps that situation allowed them to efficiently adapt their allocation of attention to obtain at least partial information about all items in the display. By contrast, in Experiments 1, 2, and 4, set size varied randomly from trial to trial. In that situation, the observer does not know in advance the resolution with which he or she should prepare to encode items and, thus, attempts to "grab" into visual WM whatever whole items he or she can. Thus, an extremely interesting project for future research would be to conduct versions of the change detection task in which memory set size was varied *between* blocks rather than within blocks. If continuous shared-resource modes of processing are enabled when observers become adapted to sustained set size presentations, continuous SDT models might fit extremely well the complete sets of ROC data obtained under such conditions.

Alternative model selection approaches and experimental paradigms

The model selection approach used in our present research was based on landscaping, which takes into account functional form complexity, unlike simpler methods such as BIC. A major criticism of landscaping is that the data are used first to define the parameter space of the landscape and then, a second time, to assess the likelihood of the data in that landscape. However, we would argue that this procedure is a relatively minor case of using data twice, because the parameters estimated from individuals define only the range of the parameter values used to generate the landscape (the parameters are then sampled from uncorrelated, uniform distributions defined across those ranges). Nevertheless, there are other, more principled model selection methods that take into account functional form complexity, such as Bayes factors and minimum description length (Myung, 2000; Pitt et al., 2002). Indeed, such methods have recently been used to evaluate the parsimony of ROC predictions made by DS and SDT models in the domain of long-term recognition memory (e.g., Kellen, Klauer, & Bröder, 2013; Klauer & Kellen, 2011). An interesting avenue for future research is to apply such methods in the domain of visual WM change detection.

Finally, the theme of the present investigation was to pursue the nature of visual WM change detection performance solely through model-based analysis of ROC data. However, in recent work, we have begun to extend the arsenal of methods for distinguishing between DS and shared-resources views. In these extended methods, we have collected extensive RT data and have formalized the predictions that alternative versions of DS and shared-resources models make for such data (Donkin, Nosofsky, Gold, & Shiffrin, in press). Such extended tools should provide still deeper insights into the nature of the forms of information processing and memory representation that underlie visual WM change detection.

Appendix Alternative SDT and DS models

Alternative set size models

In the first two analyses, we assumed specific functional relationships between set size and the sensitivity of the SDT model. First, we fitted a constrained SDT model in which d' was assumed to be a decreasing linear function of memory set size. This constrained SDT model yielded a better BIC fit than did the unconstrained version for only 6 of the 20 data sets. Furthermore, it yielded a better BIC fit than did the fixed-capacity DS model for only 2 of the 20 data sets.

In a second attempt to find a better SDT model, we fitted a *sample size* model variant of the SDT model (Palmer, 1990; Taylor, Lindsay, & Forbes, 1967). Smith and Sewell (2013) showed that the sample size model provides an excellent account of the influence of changing set size from 1 to 4. In this model, the observer is assumed to divide attention evenly among all items in the display. As such, the discriminability of any one item in a display containing M items is given by $d'_M = d'_1/\sqrt{M}$. That is, the sum of the squared d's is constant for all display sizes. We assumed the same relationship between d' and set size, estimating a single d' parameter for all set sizes. The sample size SDT model provided a better account of the data than did the unconstrained SDT model for only 5 of the 20 participants according to BIC and fit better than the DS model for only 4 of the 20 participants.

Finally, we compared the unconstrained SDT model with an equivalently unconstrained DS model. In this variable-capacity DS model, we drop the assumption of a fixed capacity for all set sizes and, instead, freely estimate the probability that an item is in memory for each of the S set size conditions. The variable-capacity DS model and the SDT model both make no a priori predictions for the effect of set size on sensitivity, and so if the set size factor is the reason for the discrepancy between Experiment 3 and the other experiments, we should expect to see the SDT model provide a much better account of the data than the variable-capacity DS model. We did not observe such a pattern, since 10 of 20 participants were still better fit by the DS model. An exact binomial test suggests that it is unlikely ($p = .048$) to have observed so many DS participants if the true proportion of participants better fit by the DS model is the same as for Experiment 3 (.29). It is also worth noting that a replication of the earlier simulation studies using the variable-capacity DS model suggests that neither model is more flexible than the other and that the true model should be recovered approximately 92 % of the time.

Clearly, the first two analyses do not rule out the idea that some other version of an SDT model that places constraints on how sensitivity varies with set size could outperform the fixed-capacity DS model. Our initial analyses, however, fail to provide support for the idea. Furthermore, our third analysis suggests that the difference between the conclusions for Experiment 3 and our other experiments is not simply due to relative parsimony in predicting how sensitivity varies with set size.

Alternative response bias model

In our alternative response bias model, we assumed that the observer sets the criterion at location c_{ij} in set size condition i and change probability condition j, and we defined *bias* (β) as the distance of the criterion from the midpoint of the noise and signal distributions, relative to overall d' (Macmillan & Creelman, 1991, p. 37):

$$\beta = \left(c_{ij} - \frac{d'_i}{2}\right)\Big/d'_i$$

We then assumed that this alternatively defined form of response bias depended only on change probability level j and was invariant with set size i.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19,* 716–723.

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

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

Bays, P. M., Catalao, R. F. G., & Husain, M. (2009). The precision of visual working memory is set by allocation of a shared resource. *Journal of Vision, 9,* 1–11.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences, 24,* 87–185.

Donkin, C., Nosofsky, R. M., Gold, J., & Shiffrin, R. M. (in press). Discrete-slots models of visual working-memory response times. *Psychological Review*.

Gilden, D. L. (2009). Global model analysis of cognitive variability. *Cognitive Science, 33,* 1141–1467.

Jang, Y., Wixted, J., & Huber, D. E. (2011). The diagnosticity of individual data for model selection: Comparing signal-detection models of recognition memory. *Psychonomic Bulletin & Review, 18,* 751–757.

Kellen, D., Klauer, K. C., & Bröder, A. (2013). Recognition memory models and binary-response ROCs: A comparison by minimum description length. *Psychonomic Bulletin & Review, 119,* 457–479.

Klauer, K. C., & Kellen, D. (2011). The flexibility of models of recognition memory: An analysis by the minimum-description length principle. *Journal of Mathematical Psychology, 55,* 430–450.

Luck, S. J., & Hollingworth, A. (2008). *Visual Memory.* New York: Oxford University Press.

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature, 342,* 543–545.

Macmillan, N. A., & Creelman, C. D. (1991). *Detection Theory: A User's Guide.* New York: Cambridge University Press.

Matsukura, M., & Hollingworth, A. (2011). Does visual short-term memory have a high-capacity stage? *Psychonomic Bulletin & Review, 18,* 1098–1104.

Myung, I. J. (2000). The importance of complexity in model selection. *Journal of Mathematical Psychology, 44,* 190–204.

Navarro, D. J., Pitt, M. A., & Myung, I. J. (2004). Assessing the distinguishability of models and the informativeness of data. *Cognitive Psychology, 49,* 47–84.

Palmer, J. (1990). Attentional limits on the perception and memory of visual information. *Journal of Experimental Psychology: Human Perception and Performance, 16,* 332–350.

Pitt, M. A., Kim, W., Navarro, D. J., & Myung, J. I. (2006). Global model analysis by parameter space partitioning. *Psychological Review, 113,* 57–83.

Pitt, M. A., & Myung, I. J. (2002). When a good fit can be bad. *Trends in Cognitive Sciences, 6,* 421–425.

Pitt, M. A., Myung, I. J., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review, 109,* 472–491.

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

Schwarz, G. (1978). Estimating the dimensions of a model. *Annals of Statistics, 6,* 461–464.

Smith, P. L., & Sewell, D. K. (2013). A competitive interaction theory of attentional selection and decision-making in brief, multielement displays. *Psychological Review, 120,* 589–627.

Taylor, M. M., Lindsay, P. H., & Forbes, S. M. (1967). Quantification of shared capacity processing in auditory and visual discrimination. *Acta Psychologica, 27,* 223–229.

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

Wagenmakers, E.-J., Ratcliff, R., Gomez, P., & Iverson, G. J. (2004). Assessing model mimicry using the parametric bootstrap. *Journal of Mathematical Psychology, 48,* 28–50.

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

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