Mem Cogn
DOI 10.3758/s13421-013-0333-6

Visual working memory declines when more features must be remembered for each object

Klaus Oberauer · Simon Eichenberger

© Psychonomic Society, Inc. 2013

Abstract The article reports three experiments investigating the limits of visual working memory capacity with a single-item probe change detection paradigm. Contrary to previous reports (e.g., Vogel, Woodman, & Luck, *Journal of Experimental Psychology. Human Perception and Performance,* 27, 92–114, 2001), increasing the number of features to be remembered for each object impaired change detection. The degree of impairment was not modulated by encoding duration, size of change, or the number of different levels on each feature dimension. Therefore, a larger number of features does not merely impair memory precision. The effect is unlikely to be due to encoding limitations, to verbal encoding of features, or to chunk learning of multifeature objects. The robust effect of number of features contradicts the view that the capacity of visual working memory can be described in terms of number of objects regardless of their characteristics. Visual working memory capacity is limited on at least three dimensions: the number of objects, the number of features per object, and the precision of memory for each feature.

Keywords Working memory · Visual working memory · Visual features · Capacity

It has been claimed that performance of visual working memory (VWM) depends on the number of objects to be retained, but not on the number of features of each object (Luck & Vogel, 1997; Vogel, Woodman, & Luck, 2001). This empirical claim is important for theories arguing that the capacity of VWM can be measured in terms of the number of objects that can be maintained, such as the slot model (Luck & Vogel, 1997; Zhang & Luck, 2008) and the hypothesis of a "magical number" of chunks that can be held in WM (Cowan, 2001, 2005). These theories assume that integrated objects, including all their features, form the units of VWM, such that the capacity of VWM can be measured as the number of objects that can be held simultaneously, regardless of the complexity of the objects or the number of features that have to be retained for each object (Awh, Barton, & Vogel, 2007; Fukuda, Awh, & Vogel, 2010). If the number of features of an object matters for performance in VWM tasks, this notion of a "magical number" of objects needs to be revised. Here, we show that the number of features that have to be remembered has a substantial impact on performance in a standard VWM task, change detection.

In the change detection paradigm (Luck & Vogel, 1997; Wheeler & Treisman, 2002; Wilken & Ma, 2004), an array of visual objects (the memory array) is displayed simultaneously, followed by a retention interval that exceeds the presumed duration of visual sensory memory. After the retention interval, a second display (the probe array) is presented that contains either the same number of objects (full display) or a single object in the location of one of the memory objects (single-object display). The task is to determine whether any object (in the full-display condition) or the re-presented object (single-object display) has changed, relative to the memory array. Vogel et al. (2001) presented memory arrays in which each object had two features (color and orientation) or, in one experiment, even four features (color, orientation, size, and the presence or absence of a gap). In the single-feature condition, they asked participants to attend to one predetermined feature dimension (e.g., orientation), and a change could occur only on that feature dimension (e.g., if there was a change, it was in the

K. Oberauer (✉) · S. Eichenberger
Department of Psychology – Cognitive Psychology,
University of Zurich, Binzmühlestrasse 14/22,
8050 Zürich, Switzerland
e-mail: k.oberauer@psychologie.uzh.ch

Published online: 29 May 2013

orientation of one object, while all other features remained constant between memory and probe array). In the conjunction condition, participants were asked to attend to all features, and a change could occur in any feature. Change detection declined with the number of objects in the array but did not differ between single-feature and conjunction conditions. This was found in one case even for the conjunction of two features on the same feature dimension (i.e., two squares of different colors embedded in each other to form a single object).

Later research has called into question the strong assumption that VWM performance is invariant to the number of features to be remembered for each object. Several researchers have meanwhile observed substantial differences between memory for single-colored and for bicolored objects (Delvenne & Bruyer, 2004; Olson & Jiang, 2002; Wheeler & Treisman, 2002; Xu, 2002). We regard the matter of conjunctions of features from the same dimension as settled and focus on the question of whether conjunctions of features from different dimensions can be remembered as well as single features.

One recent series of experiments has cast doubt on this invariance as well: Fougnie, Asplund, and Marois (2010) used a reproduction paradigm (Wilken & Ma, 2004; Zhang & Luck, 2008) to test VWM for colors, orientations, or both. After encoding a memory array of three objects differing in color and orientation, participants were asked to reproduce the color of one object (probed by a location cue) or to reproduce the object's orientation. In the single-feature condition, participants were told in advance which feature they would have to reproduce, whereas in the two-feature condition, they were not informed and probed for one of the two features at random. Fougnie et al. analyzed the response distributions with a model incorporating the assumption that VWM performance depends on two factors: the probability of having any memory for a tested feature and the precision with which that feature is remembered, given that it is remembered at all (Awh et al., 2007; Zhang & Luck, 2008). For instance, a person could have no clue about the tested color and, therefore, could only guess at random for a response. Alternatively, the person could have some memory of the tested color (e.g., that the tested color was reddish) and would then reproduce a color that is more or less similar to the correct color, depending on the precision of memory. The mixture model (Zhang & Luck, 2008) returns two parameter estimates: the probability of remembering the probed object's feature and the precision of memory for that feature, given that it was remembered at all. In the Fougnie et al. study, there was no difference between the two-feature condition and the single-feature condition in the probability of remembering the probed feature, but in the two-feature condition, the feature was reproduced with lower precision.

Consistent with this result, in a change detection experiment, Fougnie et al. (2010) found that the two-feature condition was more difficult than the single-feature condition when the changes were small (i.e., 20° in color or orientation space), but not when the changes were large (i.e., 90°). Because previous change detection experiments comparing single-feature and multifeature conditions (e.g., Vogel et al., 2001) used large changes, this finding might explain why those previous experiments did not detect an effect of the number of to-be-remembered features: The changes were so large that they could be reliably detected even with reduced memorial precision.

The results of Fougnie et al. (2010) imply that retaining more features in VWM is not cost-free, but they do not question the assumption that the capacity of VWM is a fixed number of objects, as specified in the slot model of Zhang and Luck (2008): The number of objects that can be held in VWM determines the probability that one randomly selected object probed for recall is held in VWM. If the object is remembered (i.e., is held in a slot), its feature can be reproduced with some precision, whereas if it is not remembered, people need to guess. In light of the Fougnie et al. results, we would need to add the assumption that the more features need to be remembered for any given object held in a slot, the less precisely each feature is retained. We could think of each slot as a discrete quantity of a resource that needs to be divided among the features of the object held in that slot.

In the present experiments, we investigated change detection with a single feature and multiple features per object. In Experiment 1, we additionally varied the extent of a change (small or large) to test whether small changes, but not large changes, would become more difficult to detect as the number of features was increased (Fougnie et al., 2010). In Experiment 2, we varied the presentation duration to investigate whether limitations of encoding (likely to diminish with longer presentation), or encoding of features through verbal codes, or encoding into long-term memory (more likely to occur at longer presentation durations) would modulate the effect of the number of features on VWM performance. In Experiment 3, we investigated whether an effect of number of features depends on how many different values people need to discriminate on each feature dimension, testing the idea that the number of features ceases to matter when features need to be retained with only a minimal level of precision.

Data analysis

We analyzed the results through a Bayesian analysis of variance (BANOVA). Bayesian statistics provides a sounder foundation for probabilistic inference than does null-

hypothesis significance testing (Kruschke, 2011; Raftery, 1995; Wagenmakers, 2007). We used the *BayesFactor* 0.9.0 package (R. D. Morey & Rouder, 2012) in R (R Development Core Team, 2012), which implements the Jeffreys–Zellner–Siow (JZS) default prior on effect sizes (Rouder, Morey, Speckman, & Province, 2012). We used the *anovaBF* function, which is part of the *BayesFactor* package, with its default settings, with one exception: We changed the scaling factor of the effect size for fixed effects from 0.5 to 1/sqrt(2). The latter value is equivalent to the scaling used for the default prior of the Bayesian *t*-test developed by Rouder, Speckman, Sun, and Morey (2009). We chose the larger scaling factor because it shifts the prior for the effect size toward larger values, thereby making it slightly easier to obtain evidence in favor of the null hypothesis and raising the bar for evidence for the alternative hypothesis.

The *anovaBF* function compares a range of linear models either with the null model (M_0, assuming no effect of the independent variables) or with the full model (M_f, assuming all main effects and interactions). For each comparison, the function returns the Bayes factor of the given model M_1 relative to the comparison model (null model or full model). The Bayes factor quantifies the strength of evidence in favor of M_1 relative to the comparison model (Berger, 2005). Specifically, the Bayes factor is the Bayesian likelihood ratio of M_1 and the comparison model. It specifies to what extent the ratio of the prior probabilities for the two models should be updated in light of the data to obtain the ratio of posterior probabilities. For instance, if we assume equal prior probabilities for two models, M_1 and M_0, and the data imply a Bayes factor of $BF_{10} = 20$, then the ratio of posterior probabilities is 20:1, implying that we should assign M_1 a posterior probability 20 times larger than that of M_0.

Experiment 1

Method

Participants

Thirty students at the University of Zurich (25 female; age range, 19–35 years) took part in a single session in exchange for course credit or 15 CHF (about 15 USD).

Materials

All memory arrays consisted of three rectangular objects. The objects were placed in three out of four possible locations, centered in the four quadrants of the screen. The three occupied locations were selected at random for each trial. The rectangles were displayed on a gray background (RGB = [150, 150, 150]). All rectangles had horizontal–vertical orientation. Each rectangle had a thin black outline and was filled with a pattern consisting of thick black stripes and thin black stripes, with the thin stripes oriented orthogonally to the thick stripes. Objects could vary along six feature dimensions: background color, orientation of the thick stripes, shape (i.e., ratio of width to height of the rectangle), size, thickness of the thick stripes, and spatial frequency of the thin stripes. There were 8 different feature values on each dimension, except for color, for which there were 12 different values (see Table 1). Example arrays are shown in Fig. 1.

There were six single-feature conditions, one for each feature dimension. In single-feature conditions, only the relevant feature varied across objects; all other features were held constant at their neutral value (see Table 1). There were 2 three-feature conditions, one with variation in color, shape, and size, and the other with variation in orientation of thick stripes, thickness of thick stripes, and frequency of thin stripes. The three irrelevant features were held constant at their neutral values. Finally, in the six-feature condition, all six features varied across objects. The feature values for each object on all varied feature dimensions were selected at random without replacement for each memory array.

The probe display consisted of a single object in the location of one of the memory-array objects, selected at random. On repetition trials (50 % of all trials), the probe object was identical to the corresponding object in the memory array. On change trials, the probe object was changed with regard to one feature on one of the relevant (i.e., varied) dimensions. In the three-feature and the six-feature conditions, the feature dimension on which the change occurred was selected at random from the relevant feature dimensions. Half the change trials had a small change, and the other half had a large change. Small changes were changes of two steps on the feature dimension, and large changes were changes of four steps. One step is a change from one feature value to the next among the 8 (or 12, in the case of color) predefined values (see Table 1). The new feature value in change trials never matched a feature value of one of the other two objects in the memory array. No-change trials, small-change trials, and large-change trials were presented in random order.

Procedure

Each trial started with an empty gray screen for 1,000 ms, followed by the memory array for 1,000 ms. During the following 1,000-ms retention interval, the screen went all gray again. The probe display following the retention interval stayed on until a response was given. Participants pressed the right arrow key for a change and the left arrow key for no change. The instruction emphasized accuracy, not speed. An error was signaled by a click sound.

Table 1 Feature values for the feature dimensions in Experiments 1–3

Feature Dimension	Feature Values E1	Feature Values E2 and E3
Color [RGB]	red [255, 0, 0]	**red** [**255**, **0**, **0**]
	orange [255, 106, 0]	orange [255, 140, 0]
	yellow [255, 216, 0]	yellow [255, 255, 0]
	light green [182, 255, 0]	light green [76, 255, 0]
	green [76, 255, 0]	**green** [**0**, **205**, **0**]
	turquoise [0, 255, 144]	cyan [0, 255, 255]
	cyan [0, 255, 255]	blue [28, 134, 238]
	blue [0, 148, 255]	magenta [255, 0, 255]
	dark blue [0, 38, 255]	
	purple [178, 0, 255]	
	violet [255, 0, 220]	
	pink [255, 0, 110]	
	(white [255, 255, 255])	
Orientation of bars: degrees	11, 29, 47, 65, 83, 101, 119, 137, (90)	11.25, 33.75, 56.25, 78.75, 101.25, 123.75, 146.25, 168.75
		Binary features in E3: **0**, **90**
Shape: width-to-height ratio: pixels	72 × 300	72 × 300
	101 × 270	**101** × **270**
	131 × 241	131 × 241
	160 × 212	160 × 212
	190 × 182	190 × 182
	219 × 153	**219** × **153**
	248 × 123	248 × 123
	278 × 94	278 × 94
	(199 × 199)	
Size: height = width at standard shape	160, 172, 185, 199, 214, 230, 247, 265, (199)	160, **172**, 185, 199, 214, **230**,247, 265
Thickness of bars: pixels	5, 15, 35, 45, 55, 65, 75, 85, (45)	30
Spatial frequency: distance between lines: pixels	4, 5, 8, 12, 19, 28, 43, 66 (12)	no lines

Note. Values in parentheses indicate the neutral feature values in Experiment 1. Values in bold were used in the binary-feature condition of Experiment 3

The three conditions (single-feature, three-feature, and six-feature) were administered in separate blocks, the order of which was determined at random for each participant. The single-feature block was subdivided into six subblocks, one for each feature dimension. Each subblock started with an instruction about the relevant feature, followed by 8 practice trials with that feature and 24 test trials. The three-feature block consisted of two subblocks, one in which color, shape, and size were the relevant features, and another in which orientation, thickness, and frequency were the relevant features. Each subblock consisted of 24 practice trials followed by 72 test trials. The six-feature block consisted of 48 practice trials, followed by 144 test trials.

Results

Accuracy of change detection declined from the single-feature to the three-feature condition and further declined in the six-feature condition. This was the case for no-change trials, for small-change trials, and for large-change trials. The mean accuracies are shown in Fig. 2.

The results of a BANOVA with number of features (one, three, or six) and size of change (zero, two, or four steps) are shown in Table 2. The initial rows of the table report the Bayes factors of linear models, each including one main effect, as compared with the null model omitting that main effect. These Bayes factors assess the strength of evidence in favor of each main effect in the experiment. The third row contains the Bayes factor for a model with both main effects and their interaction, as compared with a model omitting the interaction but keeping the two main effects. These Bayes factors reflect evidence for the interaction.

For Experiment 1, the data provide compelling evidence for both main effects. In particular, the main effect for the number of features had a Bayes factor of 5.5^{20}. This means that, if we started from equal prior probabilities for the null model and the model with this main effect, the posterior probability of the main effect model would be about 5^{20}

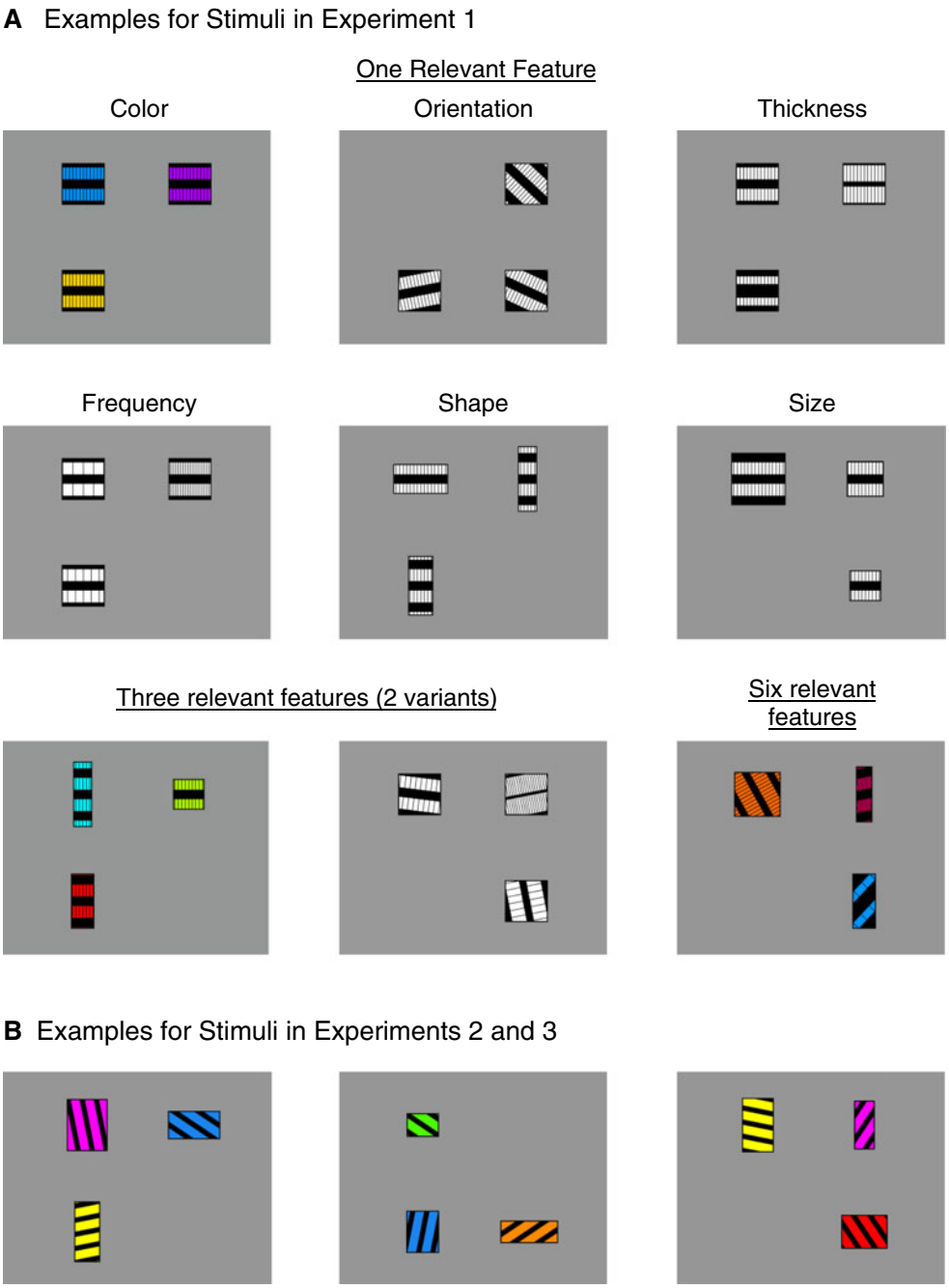

Fig. 1 Example memory displays for the experiments

times higher than that of the null model. The data provide evidence against the interaction: Starting from equal priors for the model with and the model without interaction, the posterior probability of the model without interaction is 1/0.07 = 14 times higher than that of the model including the interaction.

A BANOVA limited to the change trials (size of change = 2 or 4) returned equivalent results: The models with the two main effects had large Bayes factors, relative to the null model, for number of features BF = $3.4 * 10^{14}$, and for size of change BF = $2.9 * 10^{10}$. The model including the interaction had a Bayes factor of 0.15 relative to the model excluding the interaction. The reciprocal value expresses the support for the model without interaction: BF = 6.62.

We decomposed the effect of number of features by two BANOVAs focusing on pairwise comparisons, comparing the model with only a main effect of number of features with the null model. For the comparison of one versus three features, the Bayes factor in favor of a number-of-features effect was $1.3 * 10^{11}$, providing overwhelming evidence that memory is worse when three features, rather than a single feature, must be remembered per object. For the comparison of three versus six features, the Bayes factor was 3.49 in favor of a number-of-features effect. Thus, the data provide

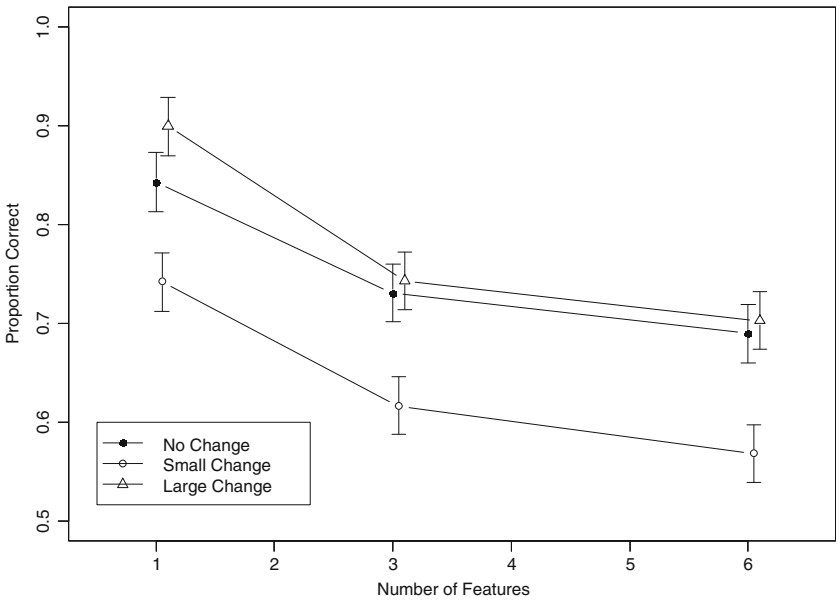

Fig. 2 Mean accuracy in Experiment 1 as a function of number of features and size of change. Error bars are 95 % Bayesian highest-density intervals (HDI95), computed using the method and the R script by Kruschke (2011). The HDI95 is the narrowest interval in which the population mean lies with a posterior probability of .95

modest evidence that remembering six features is harder than remembering three features per object.

Figure 3 shows the effect of number of features for changes on each feature dimension separately and for no-change trials. We ran six BANOVAs on change trials, one for each feature dimension, comparing the model with a main effect of number of features with the null model. The Bayes factors in Table 3 show evidence for an effect of number of features for all six feature dimensions—relatively weak for color changes, for which accuracy was close to ceiling, and strong for all other feature dimensions.

Discussion

We observed a substantial decline in change detection accuracy when the number of relevant features was increased from one to three and a further smaller decline from three to six relevant features. In contrast to the change detection experiment of Fougnie et al. (2010), this decline was not reduced for larger changes. Our large changes were hardly smaller than the large changes of Fougnie et al. (e.g., 4 steps in a color space subdivided into 12 roughly equidistant steps corresponds to the 90° changes in color space in Fougnie et al.). Therefore, the effect of number of features in our experiment cannot be explained as an effect merely on the precision of feature representations.

One noticeable difference between our experiment and that of Fougnie and colleagues (2010) is that they used a maximum of two features per objects, whereas we used

Table 2 Bayes factors of overall BANOVA models

Model	Experiment 1	Experiment 2	Experiment 3
(1) Number of features	5.5 E+20	3.3 E+14	634.8
(2) Size of change	1.1 E+12		
(2) Presentation duration		2.8 E+03	
(2) Binary vs. eight levels			220.2
Interaction (1 × 2) vs. additive-only	0.07	0.167	0.091

Note. Bayes factors for main effects reflect the strength of evidence of the model with the main effect in the first column relative to the null model assuming no main effect. Bayes factors for the interaction reflect the strength of evidence for the model with both main effects and their interaction relative to the model with both main effects but without their interaction. Very large values are given in terms of exponents of 10 for Experiments 1 and 2

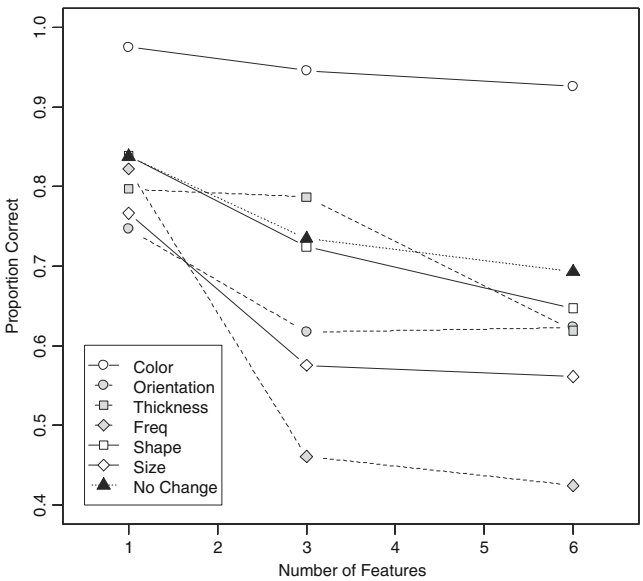

Fig. 3 Mean accuracy in Experiment 1 for change trials separately by changed feature and for no-change trials

Table 3 Bayes factors for effect of number of features for each feature dimension

Feature Dimension	Experiment 1	Experiment 2	Experiment 3
Color	2.3	1.0	.07
Orientation	40.2	1,813	.05
Thickness	6,273		
Spatial frequency	2.6 E+17		
Shape	2.7 E+04	3.4	.05
Size	3,439	2.5 E+05	.42

Note. Bayes factors are given for a model with a main effect of number of features only, as compared with the null model, using data from change trials with changes on the given feature dimension

three or six. Even in Fougnie et al.'s small-change condition, the cost of having to retain two features, relative to one, was only half as large as the cost of retaining three features, relative to one, in our experiment. It could be that increasing the number of features from one to two has a fairly benign impact on change detection accuracy, perhaps primarily affecting precision, so that it is difficult to detect with large changes. Moving beyond two features might have a more severe impact that is easier to detect experimentally even with large changes. This interpretation is consistent with the extant literature: The null effect of one versus two features per object has been replicated several times in experiments using large changes (Delvenne & Bruyer, 2004; Olson & Jiang, 2002; Riggs, Simpson, & Potts, 2011), although other experiments using equally large changes showed worse performance with two features than with one feature per object (Cowan, Blume, & Saults, 2012; Johnson, Hollingworth, & Luck, 2008; C. C. Morey & Bieler, 2012; Wheeler & Treisman, 2002; Wilson, Adamo, Barense, & Ferber, 2012). The effect of one versus two features appears to be fickle, suggesting that the effect is small on average, and probably modulated by as yet unidentified experimental details. The empirical situation is more puzzling with regard to the comparison of single-feature objects with objects with more than two features. We are aware of only one experiment (Experiment 14 in Vogel et al., 2001) including this comparison. That experiment reported a null effect for comparing a single-feature with a four-feature condition. In contrast, here we obtained a substantial effect of one versus three features and an even larger decrement of memory when each object had six features. For the next experiment, we therefore focus on this large discrepancy between our findings and Experiment 14 in Vogel et al. (2001).

We can think of four potential reasons why we found an effect of the number of features whereas Vogel et al. (2001) did not. One is that Vogel et al. (2001) presented the memory array for only 100 ms, thereby rendering verbal encoding virtually impossible, whereas we used a longer presentation interval that might have enabled participants to encode some features verbally (even though most features were difficult to verbalize in a way that discriminated them from other feature values on the same dimension).[1] Because coding features verbally is easier for sets of single-feature objects (which comprise only 3 features) than with multiple-feature objects (for a total of 9 or even 18 features), any contribution of verbal feature labels could lead to a selective benefit for the single-feature condition. Therefore, in Experiment 2, we compared single-feature and multiple-feature objects with much shorter presentation times.

The second reason could be that we set the not-relevant feature dimensions to constant, neutral values, whereas in Experiment 14 of Vogel et al. (2001), objects varied across four features even in the single-feature conditions. Thus, participants in Vogel et al. (2001) had to filter the relevant feature at encoding through selective attention, whereas our participants received perceptual assistance for selecting the relevant features in the single-feature trials. Although Vogel et al. (2001) compared the two ways of presentation for single-feature trials (with the irrelevant feature varied or held constant) across their Experiments 11 and 12 and found no difference between them, it is still possible that holding all irrelevant feature dimensions constant in stimuli with up to six feature dimensions gave the single-feature conditions in our experiment an advantage over the multiple-feature conditions. To rule out this possibility, in Experiment 2, we compared single-feature and four-feature conditions, with variation along all four feature dimensions for both conditions.

A third potential reason for the discrepancy between our results and those of Vogel et al. (2001) is that we used features that might have been more difficult to encode, because they require attention to details of the objects. This possibility pertains to the thickness of the thick stripes and the frequency of the thin stripes, which perhaps required more fine-grained visual attention because they applied to component of the objects rather than to the overall outline of the objects. Two results speak against this possibility. First, detection of changes on these two features was not harder than on any other feature, at least in the single-feature condition (see Fig. 3). Second, change detection was impaired with three features relative to a single feature for the combination of color, shape, and size, all of which are global features of the objects, for which there is no reason to believe that they were difficult to encode. Nevertheless, to rule out any contribution of encoding difficulties, we varied presentation duration and, thereby, encoding opportunity over a large range in Experiment 2.

[1] Vogel et al. (2001) also asked participants to hold two digits in mind during each change-detection trial and encouraged them to subvocally rehearse the digits. It is not clear that participants actually did rehearse the digits, and Morey and Cowan (2004) showed that reciting two digits aloud had no effect on change detection, so this methodological feature is unlikely to be of any consequence.

A fourth reason for the discrepancy might be that we used feature dimensions with 8 (or even 12, in the case of colors) different values. In contrast, Vogel et al. (2001) used binary feature dimensions in the experiment where they studied four-feature objects. With binary feature dimensions, it is easier to categorize feature values and even whole objects defined by the conjunction of multiple features. For instance, in the four-feature condition of Experiment 14 in Vogel et al. (2001), there were only $2^4 = 16$ different possible objects. It is conceivable that participants learned unified representations of these objects over the course of the experiment, such that each memory display could be coded by representations of known objects rather than by ad hoc bindings between features. Change in any feature would then be recognized as a change in object identity, not a change in an object's feature. It is possible that the unit of VWM is the largest unified representation in long-term memory that can be used to code the memoranda. For well-known objects, the unit would be the object, regardless of how many features it involves, whereas novel objects must be represented in terms of their feature values. As a consequence, the number of features would affect the accuracy of WM if and only if the stimuli are novel conjunctions of features that do not correspond to unified object representations in long-term memory. We will test this possibility in Experiment 3.

Experiment 2

In the second experiment, we again varied the number of relevant features per object. Because, in Experiment 1, both three and six features impaired change detection performance relative to a single feature, here we used only one multifeature condition with four relevant features, as in Experiment 14 of Vogel et al. (2001). We varied presentation duration over three levels: 0.1, 1.0, and 1.9 s. The shortest presentation duration served to eliminate any possibility for encoding the stimuli verbally. The longest presentation duration served to minimize difficulties of encoding, which might have contributed to the effect of the number of features in Experiment 1. Vogel, Woodman, and Luck (2006) estimated that it takes about 50 ms per object to encode single-feature objects into VWM, and Woodman and Vogel (2008) have shown that the time to encode two-feature objects is no longer than the time for encoding the slower of the two features individually, suggesting that multiple features of an object are encoded in parallel. Even on the most conservative assumption that all 12 features were encoded sequentially, 1,900 ms is three times as much as should be necessary to encode them all. Size of change did not interact with number of features in the first experiment; therefore, we chose a single, intermediate size of changes for Experiment 2.

Method

Participants

Twenty-four students at the University of Zurich (age range, 20–40 years; 19 female) took part in a single session for course credit or 15 CHF.

Materials

The materials were the same as for Experiment 1, apart from the following changes. The number of feature dimensions was reduced to four: color, shape, size, and orientation of the thick stripes. The thin stripes were omitted altogether. The number of colors was reduced to eight, so that there was an equal number of values on each dimension.

Memory arrays consisted of three objects that always varied on all four feature dimensions, regardless of condition. For the 50 % change trials, a randomly selected object was changed with regard to one of its relevant features by three steps on the feature dimension. In the four-feature trials, each feature dimension was changed equally often.

Procedure

The procedure on each trial was as in Experiment 1, with the exception that presentation duration of the memory arrays varied by condition. The session was subdivided into two blocks, one for the single-feature and one for the four-feature conditions; their order was determined at random for each participant. The single-feature block was subdivided into four subblocks, one for each feature dimension, the order of which was counterbalanced across participants. At the beginning of each subblock, participants were instructed which feature dimension was relevant for the following subblock. The multifeature block, as well as each single-feature subblock, was subdivided into three mini-blocks for the three presentation durations (0.1, 1.0, and 1.9 s). The order of presentation durations was determined at random for each participant and then held constant for that participant's subblocks. Each mini-block of the single-feature condition consisted of 2 practice trials, followed by 20 test trials, for a total of 240 test trials. Each mini-block of the four-feature condition consisted of 8 practice trials, followed by 80 test trials, so that there were again 240 test trials in total.

Results

The results are displayed in Fig. 4. Change detection accuracy again decreased when four rather than one feature had

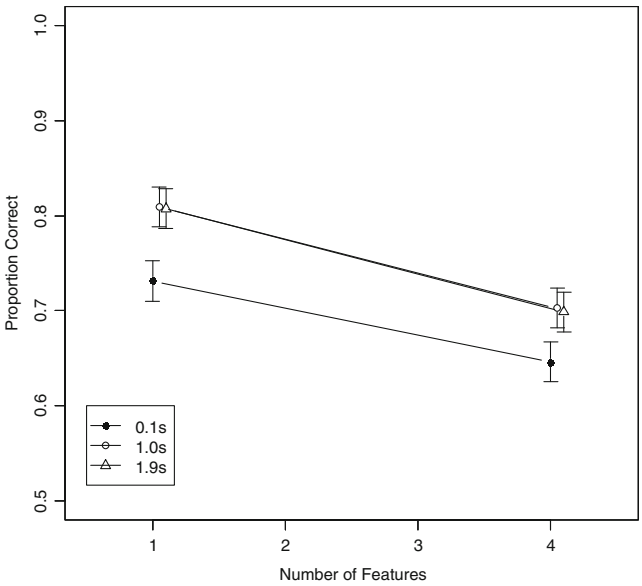

Fig. 4 Mean accuracy in Experiment 2 as a function of number of features and presentation duration. Error bars are 95 % Bayesian highest-density intervals (HDI95), computed using the method and the R script by Kruschke (2011)

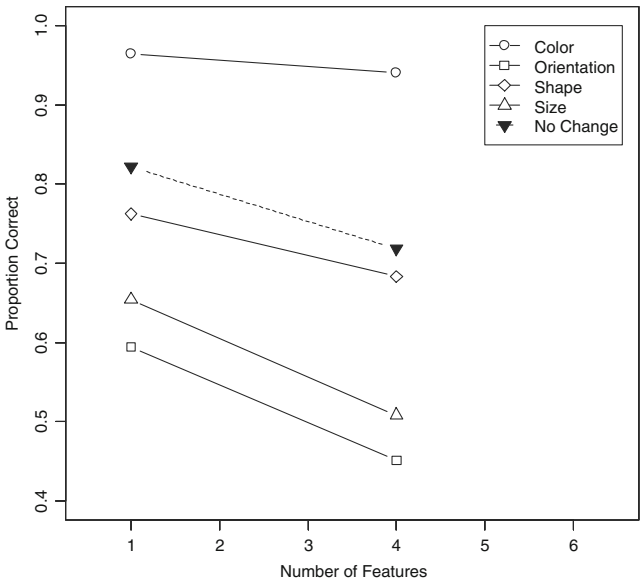

Fig. 5 Mean accuracy in Experiment 2 for change trials separately by changed feature and for no-change trials

to be remembered. This effect was present regardless of presentation duration. Increasing the presentation duration had a beneficial effect up to 1 s, but there was no further benefit from 1.0 to 1.9 s.

The Bayes factors from a BANOVA with number of features (1 vs. 4) and presentation duration (0.1, 1.0, or 1.9 s) as independent variables are presented in Table 2. There was again strong evidence for a main effect of number of features and, also, for a main effect of presentation duration. The model with only the two main effects was superior to the full model including the interaction, with a Bayes factor of 6.0. This result provides modest evidence against the interaction. A BANOVA comparing the first two levels of presentation time showed strong evidence for higher accuracy with 1.0 s than with 0.1 s, $BF_{10} = 4{,}284$. In contrast, for the comparison of 1.0 and 1.9 s, we obtained $BF_{10} = 0.16$, which implies that the data support the null model more than the model including the presentation time effect. The Bayes factor in favor of the null model is $BF_{01} = 1/BF_{10} = 6.25$.

Figure 5 shows the effect of number of features for each feature dimension separately and for the no-change condition. Separate BANOVAs for trials with changes on each feature dimension revealed evidence for a number-of-features effect for all dimensions except color, which again yielded performance close to ceiling.

Discussion

Experiment 2 replicated the sizable decrement in change detection when 4 features, instead of 1, had to be remembered for each object. The effect of number of relevant features was of equal size for all three presentation durations, ruling out the possibility that the effect was due to verbal encoding and the possibility that it was due to encoding difficulties. The finding that performance did not increase from 1.0 to 1.9 s shows that beyond 1 s, change detection was not limited by encoding, in agreement with Vogel et al. (2006). This finding also speaks against the possibility that people relied on verbal codes even at the longer presentation durations, because any attempt to verbally encode up to 12 features would be expected to benefit from presentation durations exceeding 1 s. We conclude that participants did not try to encode the memory arrays verbally, probably because the features were very difficult to label, even with generous time. This finding also renders unlikely the possibility that performance relied on encoding of objects into long-term memory, because long-term memory improves gradually with longer encoding durations (Ganor-Stern, Seamon, & Carrasco, 1998). In Experiment 2, the memory arrays for single-feature and multiple-feature trials were physically identical, so that participants received no perceptual assistance in selecting the relevant feature dimension in the single-feature trials. To conclude, we ruled out the first three plausible explanations discussed at the end of Experiment 1 for why we found an effect of number of features whereas Vogel et al. (2001) did not.

Experiment 3 tested the fourth potential explanation for this discrepancy. When there are only two possible feature values per feature dimension, the number of possible objects is fairly small, so that people might learn unified representations of them in long-term memory, which might enable them to encode the objects as unified entities or chunks. In contrast, when there are eight different levels on each

feature dimension, the number of objects that can be created from combining them is too large for people to learn unified representations within an experiment. If this is the case, the ability of VWM to remember as many multifeature objects as single-feature objects could be explained in the same way as the ability of verbal WM to remember (nearly) as many known words as single letters: The multiletter strings that form known words count as a single unit toward WM capacity because they have a unified representation in long-term memory.

Experiment 3

In Experiment 3, we contrasted feature dimensions with eight levels, as in the previous experiment, with feature dimensions with only two levels, as in Vogel et al.'s (2001) Experiment 14. Participants worked on a change detection task in two sessions. In one session, each of four feature dimensions (size, shape, color, and orientation of thick stripes) varied across eight levels. In the other session, each feature dimension had only two levels.

Method

Participants

Twenty-four students at the University of Zurich (age range, 18–35 years) took part in two 1-h sessions for course credit or 30 CHF. Data from one session of 1 participant were lost due to experimenter error; this participant was replaced.

Materials and procedure

The features for the session with eight levels per dimension were the same as in Experiment 2. For the session with binary features, we selected two levels that were four steps apart on the eight-level dimensions, with the exception of orientation, for which we used the canonical orientations (0° and 90°) that were not included in the eight levels of that dimension (see Table 1). On change trials, the changed feature was always changed by four steps; thus, the size of change was the same in the eight-level and the binary feature conditions. In the session with binary features, two of the three objects in the memory display always had the same feature value (e.g., red), while the third object had the alternative value (e.g., green). In this way, none of the features was homogeneous across the whole display.

The change detection task was the same as in Experiment 2, with the following modifications. In each session, there were three blocks, one single-feature block, one 2-feature block, and one 4-feature block. Their order was counterbalanced across participants. Each block consisted of 88 trials (8 practice, 80 test). Single-feature blocks were subdivided into four consecutive subblocks, one for each relevant feature. Each subblock began with 2 practice trials, followed by 20 test trials. Two-feature blocks were subdivided into two subblocks as follows. There are six possible pairwise combinations of four feature dimensions. Each participant started with one of them in the first subblock of 44 trials (4 practice, 40 test) and then switched to the complementary pair in the second subblock. For instance, if the first combination was color–shape, the second combination was orientation–size. The choice of the two pairwise combinations of feature dimensions was counterbalanced across subjects. Four-feature blocks consisted of 8 + 80 trials on which all four features had to be remembered. Each feature dimension was determined to be the relevant one on a random subset of 20 test trials. On half of them, a change was introduced on that feature dimension in one randomly chosen object.

Results

Mean accuracies in each condition are shown in Fig. 6. As in the preceding experiments, accuracy declined with a larger number of features, but this effect was limited to the comparison of one and two features per object; there was no difference between two-feature and four-feature objects. Change detection was better for binary features than for eight-level features. The two effects did not interact.

The overall BANOVA (Table 2) returned Bayes factors strongly supporting the main effects of number of features

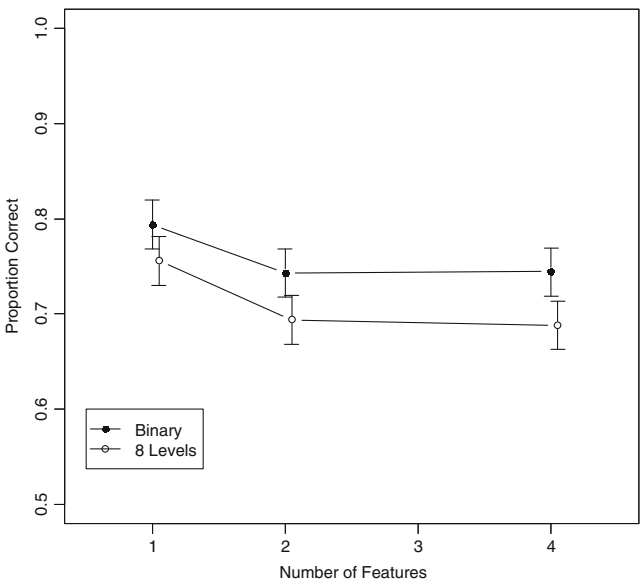

Fig. 6 Mean accuracy in Experiment 3 as a function of number of features and number of feature levels. Error bars are 95 % Bayesian highest-density intervals (HDI95), computed using the method and the R script by Kruschke (2011)

and of number of feature levels (binary vs. eight levels). There was also evidence against the interaction, with BF_{1f} = 10.0 in favor of the model with two main effects only, as compared with the full model including the interaction. The contrast between single-feature and two-feature objects was strongly supported, BF_{10} = 892.4. The contrast between two-feature and four-feature objects received no support, BF_{10} = 0.16, which translates into BF_{01} = 6.3 in favor of the null model.

Figure 7 shows the effect of number of features for individual feature dimensions and for no-change trials. Bayes factors for the number-of-feature effect on trials with changes on individual feature dimensions can be found in Table 3. In contrast to the overall analyses, these Bayes factors provide evidence against a number-of-features effect for three of the four feature dimensions. It appears that the overall effect of number of features was driven by the effect on no-change trials, for which BF_{10} = 1,898.8. One possible explanation for this result is that participants adapted their criterion for reporting a change to the difficulty of detecting a change, keeping hit rate roughly constant over different numbers of features at the expense of correct rejections.

In a final analysis we investigated whether the order in which participants worked through the two sessions, one with binary and one with eight-level features, interacted with the number of features and the number of feature levels. The motivation for this analysis was that perhaps participants who started with the eight-level features were motivated to encode the stimuli with higher precision to be able to discriminate between all eight levels, and they might have carried over this mental set for a high precision into the session with binary features.[2] As Fougnie et al. (2010) have shown, increasing the number of features per object impairs precision. Perhaps the reverse is also true: If people try to encode and remember each feature at a high precision, they can remember fewer features per object. Figure 8 plots accuracy as a function of number of features, number of levels, and session order. In line with the above speculation, the effect of number of features appears to be reduced in the subsample of participants that began with binary features in the first session. To evaluate the apparent pattern, we ran a BANOVA with number of features, number of feature levels, and session order as independent variables. The interaction between number of features and session order can be assessed by comparing the full model (including all main effects and interactions) with a model omitting only that two-way interaction. The Bayes factor in favor of the omission was 3.1. Thus, the evidence is ambiguous, but leaning more against that interaction than in favor. Nevertheless, we also ran a BANOVA with number of features and number of levels restricted to the subsample that started with binary features. In this subsample, there was no evidence in favor of (or against) an effect of number of features, BF_{10} = 1.08, or of number of feature levels, BF_{10} = 1.04.

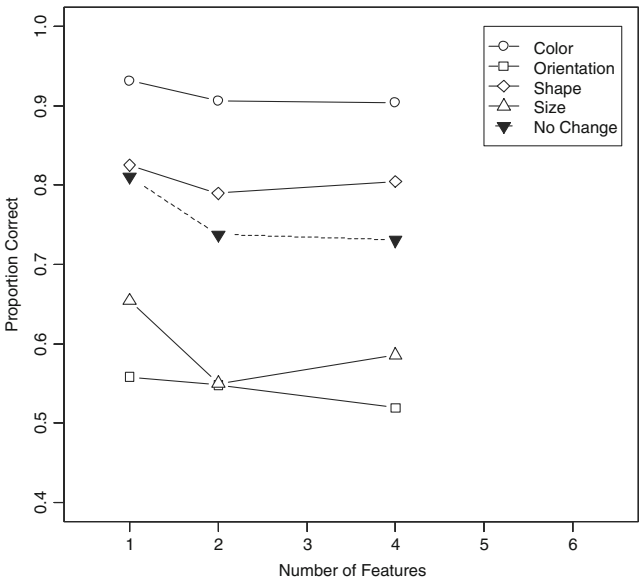

Fig. 7 Mean accuracy in Experiment 3 for change trials separately by changed feature and for no-change trials

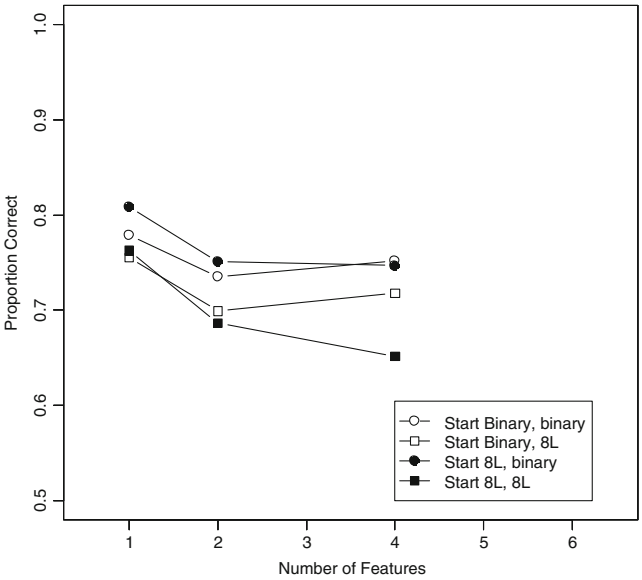

Fig. 8 Mean accuracy in Experiment 3 separately for groups starting with binary feature dimensions in session 1 (Start Binary) and starting with eight-level features (Start 8 L)

Discussion

Our third experiment replicated the detrimental effect of number of features on change detection accuracy.

[2] We are grateful to Ed Vogel for suggesting this possibility.

Reducing the number of feature levels on each dimension from eight to two improved change detection, but this variable did not interact with the number of features. The absence of this interaction rules out the fourth potential explanation for the discrepancy between our results and those of Vogel et al. (2001): It is not the case that the effect of number of features disappears with binary features. Seeing the same objects multiple times in the experiment does not help people to encode them as unified objects, in a way that the number of features has no impact on memory—at least, not within a single-session experiment.

The present results add evidence against the possibility that verbal encoding contributed to the effect of number of features. Binary features are easier to encode verbally than features with eight levels. If verbal encoding had anything to do with the number-of-features effect, that effect should be larger with binary features than with eight-level features. This was not the case.

Analyzing the data by whether participants started the experiments with binary or eight-level features provided inconclusive results. Although there was no evidence that the order of sessions interacted with the number-of-features effect, when we focused the analysis on those subjects who began with binary features, we found, for the first time, a pattern of results compatible with a null effect of number of features. It is conceivable that the number of feature levels that participants experience at the beginning of the experiment is a critical variable: The only previous experiment we are aware of that investigated VWM with up to four features per object (Experiment 14 in Vogel et al., 2001) used binary features throughout and found no difference in memory for single-feature and four-feature objects.

Assuming that there is, in fact, such an effect, one possible explanation for it is that people form a mental set for the required precision of encoding feature values and that, when the precision is set high, they can encode fewer features per object. Note that the required trade-off is not one of higher precision on one feature dimension for lower precision on another dimension. In Experiment 3, changes were equally large in all conditions, implying that low precision is sufficient for accurate change detection. For this explanation to work, it would have to be the case that trying to encode features with high precision leads to the complete loss of information about some of an object's features. In addition, we would have to make two further assumptions. First, people set the precision not to the value required for successful change detection—which was equally low in all conditions of Experiment 3—but according to how many different feature levels they observe. Second, they do not change that set from one session to another, separated by several days, even when the number of levels changes between sessions.

General discussion

Across three experiments, we observed a substantial decrement in change detection when people had to remember multiple features per object, as compared with single-feature objects. This finding questions the generality of the claim that performance in VWM tasks is independent of the number of features integrated into a single object.

Before we discuss substantive implications from our results, we need to address a possible methodological objection: Could it be that our use of Bayesian statistics instead of null-hypothesis testing contributed to some of our results—in particular, those results that are not in agreement with previous findings? The answer is a clear *no*. First, we analyzed our data with conventional ANOVAs and obtained highly significant main effects and clearly nonsignificant interactions in all three experiments, exactly matching the conclusions from the BANOVAs. Second, it has been shown that Bayesian *t*-tests using the JZS default prior are more conservative than conventional *t*-tests or ANOVAs (Rouder et al., 2009; Wetzels et al., 2011).

Our results are in contrast to those of Vogel et al. (2001), who found no cost for remembering up to four features per object, as compared with a single feature per object. As such, our experiments might be dismissed as a failure of replication. Doing so would be a mistake on four counts. First, failures to replicate are informative and should not be dismissed unless they can be attributed to methodological deficiencies of the replication attempt (Nosek, Spies, & Motyl, 2012). Second, typical failures of replication fail to reproduce an effect reported in the literature. Here, we have the opposite situation: The literature reports a null effect, whereas we consistently obtained an effect of the number of features. This fact rules out one major potential cause of failures to replicate: lack of power. Third, with regard to the comparison of single-feature and two-feature objects, the literature is already mixed, with some reports finding no difference (Delvenne & Bruyer, 2004; Olson & Jiang, 2002; Riggs et al., 2011; Vogel et al., 2001) and others finding worse performance when two (or more) features need to be remembered per object (Cowan et al., 2012; Johnson et al., 2008; C. C. Morey & Bieler, 2012; Stevanovski & Jolicoeur, 2011; Wheeler & Treisman, 2002; Wilson et al., 2012). Therefore, our finding is not completely at odds with previous findings. Fourth, we did not attempt to replicate the experiment of Vogel et al. (2001) exactly. It is plausible that the null effect of Vogel et al. (2001) could be replicated with an exact reproduction of their method. Whether or not this is so is of minor theoretical interest, because the theoretical conclusions from the null effect of number of features do not hold if that null effect is obtained only under a narrow set of specific conditions (cf. Fiedler, 2011).

The most far-reaching theoretical conclusion from the null effect of number of features has been that the capacity of VWM can be characterized in terms of a limited number of objects, regardless of the objects' complexity or number of features. If this were true, it would have to hold generally, not only for a specific set of boundary conditions—unless the boundary conditions can be justified by theoretical assumptions that are compatible with the assumption of a discrete, object-based capacity. The finding that multiple features on the same dimension (e.g., two-colored objects) are harder to remember than a single feature (Olson & Jiang, 2002; Wheeler & Treisman, 2002; Xu, 2002) could be reconciled with the object-based capacity hypothesis by arguing that two-color objects are not perceived as a single, unitary object but, rather, are encoded as two separate parts. The finding that remembering two features reduces memory precision relative to a single feature (Fougnie et al., 2010) could be reconciled with the object-based capacity view by assuming that VWM is characterized by two parameters: the number of objects that can be stored (i.e., the capacity) and the precision with which objects are stored (Awh et al., 2007; Fukuda, Vogel, Mayr, & Awh, 2010). Our findings cannot be reconciled with the object-based capacity view with either of these additional assumptions. Change detection became worse as the number of features per object was increased, regardless of whether these features had to be retained with high precision (to detect small changes) or with low precision (to detect large changes). Therefore, it is not the case that increasing the number of features merely impairs the precision with which they are maintained.

The effect of number of features cannot be explained by limitations of encoding, because it persisted with approximately equal size for encoding durations between 0.1 and 1.9 s. The effect cannot be attributed to verbal encoding, because the stimuli of Experiments 1 and 2 were very difficult to encode by verbal labels in a way that allowed discrimination between different feature values. Moreover, verbal encoding would take time, such that we should expect a larger contribution of verbal encoding with longer presentation durations. Nevertheless, the effect of number of features did not increase with presentation duration in Experiment 2. Finally, it should be much easier to use verbal labels when there are only two possible values per feature dimension than when there are eight, and yet, in Experiment 3, the effect of number of features was of equal size for both these conditions. For similar reasons, the effect of number of features cannot be attributed to a presumed contribution of long-term memory to change detection: Encoding into long-term memory should improve with longer presentation duration, so that the relative contribution of long-term memory to performance should be larger with longer presentation duration, but the effect of number of features did not increase with presentation time.

Our results, together with previous reports of an effect of number of features, imply that the original null effect reported by Vogel et al. (2001) does not hold generally. Therefore, the hypothesis that the capacity of VWM is limited to a constant number of objects regardless of their features is no longer tenable. At least under some condition, the capacity of VWM is also limited on a second dimension, the number of features. On a third dimension, the precision of each feature is limited (Awh et al., 2007). This multidimensional characterization of capacity is still compatible with the assumption that, on one dimension, there is a discrete capacity to hold a maximum number of objects in VWM. For instance, Cowan et al. (2012) have proposed that VWM can hold a maximum of k objects (with k around 3), such that for each object, at least one feature is maintained. Additional features are maintained with independent probabilities smaller than one, such that when more than one feature must be retained for each object, memory for an object's feature can fail even if that object is in WM. Because, for each object in WM, one of its features is maintained with certainty, whereas all other features are not, the chance that the tested feature is in memory, given that the tested object is in memory, decreases with the number of features in a decelerating fashion, as observed in our Experiments 1 and 3. In the Appendix, we show that the data of those two experiments are consistent with a formal model of the proposal of Cowan et al.

Cowan et al. (2012) do not explain why people remember one feature per object with certainty but additional features only with a probability lower than one. One potential rationale for this assumption comes from the dimensional-weighting hypothesis raised in the context of visual search studies (Müller, Heller, & Ziegler, 1995; Müller, Reimann, & Krummenacher, 2003). Search for a singleton in a visual field is facilitated if the feature dimension on which the singleton stands out against distractors is known in advance, suggesting that people can prioritize one feature dimension over others. It is plausible that they do so also when encoding visual stimuli into WM. As a consequence, features on the prioritized dimension might be encoded with higher probability than features on nonprioritized dimensions. When more than one feature needs to be encoded per object, only one of them is encoded with priority. One new prediction from this idea is that visual WM should be impaired if the array as a whole contains objects characterized by different feature dimensions, even when only one feature needs to be remembered for each object. For instance, an array could contain a colored blob (potentially changing its color), a gray bar (potentially changing orientation), and a geometric figure (potentially changing shape). Detecting a change in such a display should be harder than detecting a color change among three blobs, detecting an

orientation change among three bars, or detecting a shape change among three figures.

The idea that visual WM capacity is limited on multiple dimensions (i.e., number of objects, number of features, feature precision, and perhaps others) would receive support if the capacity on one dimension was found to be independent of the load on the other dimensions, such that people can always hold *k* objects in VWM regardless of how many features need to be remembered and regardless of the required level of precision. The multidimensional capacity view becomes less attractive if there are trade-offs between the dimensions, because such trade-offs suggest that the load effects on different dimensions are merely different manifestations of load on a single underlying dimension limiting capacity. The finding of Fougnie et al. (2010) that precision suffers when more features need to be remembered is an instance of such a trade-off. Our Experiment 3 hints at the possibility that there is also the reverse trade-off: When people try to encode and maintain features with higher precision, they can retain fewer features. Precision also decreases with an increasing number of objects to be remembered (Anderson, Vogel, & Awh, 2011; Bays, Catalao, & Husain, 2009; Wilken & Ma, 2004; Zhang & Luck, 2008), demonstrating a trade-off between number of objects and precision. The latter trade-off appears not to be under the person's control: Zhang and Luck (2011) found no evidence that people could trade lower precision for a larger number of objects in VWM when given incentives to do so (cf. Murray, Nobre, Astle, & Stokes, 2012). We are not aware of any evidence speaking to the possible trade-off between number of objects and number of features; this should be a goal for future research.

To the extent that capacity limits on the three dimensions trade off against each other, a unidimensional notion of capacity becomes attractive as the more parsimonious description. A unidimensional capacity limit could be conceptualized as a constant resource for maintaining visual objects (Bays et al., 2009), with the assumption that a larger amount of that resource needs to be allocated to each object when more of its features need to be remembered and when they need to be remembered with higher precision. One challenge for such a constant-resource model comes from our finding that the effect of number of features levels off rapidly as more features are added to each object, with only a modest decline in accuracy from three to six features (Experiment 1) and no decline between two and four features (Experiment 3). If every additional feature consumes a constant additional resource quantity, then six features should require twice as much of the resource than three features, and four features should require twice as much as two features. Resource theorists could assume that the first feature of every object requires a large amount of the resource and every additional feature requires an increasingly smaller amount, but this would be an arbitrary assumption of the sort that renders resource theories infinitely flexible and, thereby, untestable (Navon, 1984). For this reason, we do not regard a resource model an entirely satisfying explanation of our findings.

To conclude, our results show that the capacity of VWM cannot be exhaustively described as a limited number of objects. For a more complete characterization of VWM capacity, we need to map out the capacity limits in three dimensions—number of objects, number of features, and precision—and their trade-offs.

Author note The research reported in this article was supported by a grant from the Swiss National Science Foundation (Grant number 100014_135002) to the first author. We are grateful to Sonja Peteranderl for collecting the data in Experiment 3.

Appendix

Applying the model of Cowan et al. (2012) to Experiments 1 and 3

Cowan et al. (2012) proposed that a constant number of *k* objects can be held in WM, and for each object, one of its *n* features is stored with certainty, whereas all remaining features $i = 2{:}n$ are stored with a probability x_i. In our experiments, the feature to be tested in the selected object is chosen at random, so the chance that the one feature encoded with certainty is $1/n$. We can therefore express the probability of giving a correct response to a test of feature *i* in a condition with *n* features per object as

$$P(\text{correct } i|n) = \frac{1}{n}Pm + \left(1-\frac{1}{n}\right)Pm{\cdot}x_i, \tag{1}$$

where *Pm* is the probability of remembering the tested object. If *k* is constant across manipulations of *n*, then *Pm* should be constant too, $Pm = \min(1, k/N)$, where *N* is the number of objects to remember.

For single-feature objects, Eq. 1 reduces to

$$P(\text{correct } i|1) = Pm. \tag{2}$$

Solving Eq. 1 for x_i yields

$$x_i = \frac{P(\text{correct } i|n) - \frac{1}{n}Pm}{\left(1-\frac{1}{n}\right)Pm}. \tag{3}$$

With three or more levels of *n*, we can use $n = 1$ to compute *Pm* and the second level of *n* to compute x_i for each feature dimension *i*. These values can be used to predict $P(\text{correct } i|n)$ for all remaining levels of *n*. To the extent that this prediction holds, the data are consistent with

Table 4 Means and 95 % highest density intervals (in brackets) for observed accuracy and accuracy predicted by the Cowan, Blume, and Saults (2012) model for the highest number of features

Experiment	Observed	Predicted
1	69.6 [66.6, 72.6]	70.6 [67.6, 73.6]
3, binary	74.4 [70.4, 78.5]	71.3 [67.1, 75.4]
3, 8-level	68.6 [64.6, 72.5]	66.1 [62.1, 70.0]

the assumption that k is constant across variations of the number of features per object.

We tested the predictions for the highest level of n in Experiments 1 and 3. In Experiment 1, we removed the small-change condition because Cowan's model predicts virtually perfect performance for comparisons of probes with memory representations only when the changes are large enough so that the role of precision is negligible. For Experiment 3, we carried out separate tests for the session with binary and the session with eight-level feature dimensions. Thus, there were three BANOVAs with observed versus predicted values as one independent variable and tested feature as the second independent variable.

In both experiments, the model with a main effect of observed versus predicted was not supported over the null model: BF_{10} = .108 for Experiment 1, BF_{10} = .322 for Experiment 3 (binary), and BF_{10} = .202 for Experiment 3 (eight-level). The reciprocal values of these Bayes factors reflect the support for the null hypothesis that there is no difference between predicted and observed values, which can be interpreted as support for the assumption that k is invariant across n. The joint support from all three tests for the invariance assumption is the product of the three BF_{01} values: 1/.108 * 1/.322 * 1/.202 = 142.

Table 4 shows the observed and predicted accuracies at the highest level of number of features in each experiment. The model predicts the data of Experiment 1 virtually perfectly but slightly underestimates accuracy for Experiment 3. The reason for the underestimation is that the model must predict a drop in accuracy from two to four features, which was not observed in Experiment 3. Nevertheless, the data of Experiment 3 provide more support for the model than against it.

References

Anderson, D. E., Vogel, E. K., & Awh, E. (2011). Precision in visual working memory reaches a stable plateau when individual item limits are exceeded. *Journal of Neuroscience, 31,* 1128–1138.

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.

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.

Berger, J. O. (2005). Bayes Factor. In S. Kotz, N. Balakrishnan, C. Read, B. Vidakovic, & N. L. Johnson (Eds.), *Encyclopedia of Statistical Sciences, Second Edition* (2nd ed., Vol. 1, pp. 378–386). Hoboken, NJ: Wiley.

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

Cowan, N. (2005). *Working memory capacity*. New York: Psychology Press.

Cowan, N., Blume, C. L., & Saults, J. S. (2012). Attention to attributes and objects in working memory. *Journal of Experimental Psychology*: *Learning*, *Memory*, *and Cognition*.

Delvenne, J.-B., & Bruyer, R. (2004). Does visual short-term memory store bound features? *Visual Cognition, 11,* 1–27.

Fiedler, K. (2011). Voodoo correlations are everywhere - not only in neuroscience. *Perspectives on Psychological Science, 6,* 163–171.

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

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

Fukuda, K., Vogel, E. K., Mayr, U., & Awh, E. (2010b). Quantity, not quality: The relationship between fluid intelligence and working memory capacity. *Psychonomic Bulletin & Review, 17,* 673–679.

Ganor-Stern, D., Seamon, J. G., & Carrasco, M. (1998). The role of attention and study time in explicit and implicit memory for unfamiliar visual stimuli. *Memory & Cognition, 26,* 1187–1195.

Johnson, J. S., Hollingworth, A., & Luck, S. J. (2008). The role of attention in the maintenance of feature bindings in visual short-term memory. *Journal of Experimental Psychology. Human Perception and Performance, 34,* 41–55.

Kruschke, J. K. (2011). *Doing Bayesian data analysis: A tutorial with R and BUGS*. New York: Academic Press.

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

Morey, C. C., & Bieler, M. (2012). Visual short-term memory always requires general attention. *Psychonomic Bulletin & Review, 20,* 163–170. doi:10.3758/s13423-012-0313-z

Morey, C. C., & Cowan, N. (2004). When visual and verbal memories compete: Evidence of cross-domain limits in working memory. *Psychonomic Bulletin & Review, 11,* 296–301.

Morey, R. D., & Rouder, J. N. (2012). BayesFactor (Version 0.8.8). Retrieved from http://cran.at.r-project.org/web/packages/BayesFactor/index.html

Müller, H. J., Heller, D., & Ziegler, J. (1995). Visual search for singleton feature targets within and across feature dimensions. *Perception & Psychophysics, 57,* 1–17.

Müller, H. J., Reimann, B., & Krummenacher, J. (2003). Visual search for singleton feature targets across dimensions: Stimulus- and expectancy-driven effeccts in dimensional weighting. *Journal of Experimental Psychology. Human Perception and Performance, 29,* 1021–1034. doi:10.1037/0096-1523.29.5.1021

Murray, A. M., Nobre, A. C., Astle, D. E., & Stokes, M. G. (2012). Lacking control over the trade-off between quality and quantity in visual short-term memory. *PLoS One, 7*(8), e41223. Retrieved from doi:10.1371/journal.pone.0041223

Navon, D. (1984). Resources - a theoretical soupstone? *Psychological Review, 91,* 216–234.

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific Utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Cognitive Science, 7,* 615–631.

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

R-Development-Core-Team. (2012). R: A language and environment for statistical computing (Version 2.15.2). Vienna, Austria: R Foundation for Statistical Computing. Retrieved from URL: http://www.R-project.org

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology, 25,* 11–163.

Riggs, K. J., Simpson, A., & Potts, T. (2011). The development of visual short-term memory for multifeature items during middle childhood. *Journal of Experimental Child Psychology, 180,* 802–809.

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology, 56,* 356–374.

Rouder, J. N., Speckman, P. L., Sun, D., & Morey, R. D. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16,* 225–237.

Stevanovski, B., & Jolicoeur, P. (2011). Consolidation of multifeature items in visual working memory: Central capacity requirements for visual consolidation. *Attention, Perception, & Psychophysics, 73,* 1108–1119.

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.

Vogel, E. K., Woodman, G. F., & Luck, S. J. (2006). The time course of consolidation in visual working memory. *Journal of Experimental Psychology. Human Perception and Performance, 32,* 1436–1451.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review, 14,* 779–804.

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E.-J. (2011). Statistical evidence in experimental psychology: An experimental comparison using 855 t Tests. *Psychological Science, 6,* 291–298.

Wheeler, M. E., & Treisman, A. M. (2002). Binding in short-term visual memory. *Journal of Experimental Psychology. General, 131,* 48–64.

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

Wilson, K. E., Adamo, M., Barense, M. D., & Ferber, S. (2012). To bind or not to bind: Addressing the question of object representation in visual short-term memory. *Journal of Vision, 12,* 1–16.

Woodman, G. F., & Vogel, E. K. (2008). Selective storage and maintenance of an object's features in visual working memory. *Psychonomic Bulletin & Review, 15,* 223–229.

Xu, Y. (2002). Limitations of object-based feature encoding in visual short-term memory. *Journal of Experimental Psychology. Human Perception and Performance, 28,* 458–468.

Zhang, W., & Luck, S. J. (2008). Discrete fixed-resolution representations in visual working memory. *Nature, 453,* 233–236.

Zhang, W., & Luck, S. J. (2011). The number and quality of representations in working memory. *Psychological Science, 22,* 1434–1441.