When "capacity" changes with set size: Ensemble representations support the detection of across-category changes in visual working memory

Mark W. Schurgin & Timothy F. Brady
Department of Psychology, University of California, San Diego

Address for correspondence:
Mark Schurgin
maschurgin@ucsd.edu
9500 Gilman Drive #0109
McGill Hall
Department of Psychology
University of California, San Diego

Abstract

Is there a fixed limit on how many objects we can hold actively in mind? Generally, researchers have found participants perform less well at remembering a small number of objects if those objects are more complex, suggesting a limited resource rather than a fixed number of objects best explains working memory performance. However, some evidence has suggested that stimulus similarity better accounts for these effects, and that after accounting for such similarity, the data support a slot-based fixed item limit for working memory. Much of the evidence used to support the latter claim relies on working memory displays containing different categories of items. It has been found that for large, across-category changes, performance does not differ for different kinds of complex stimuli. However, many of these studies fail to adequately control for the potential use of ensemble information in discriminating such large, across-category changes. Here, we sought to identify how much ensemble representations may explain performance across these tasks. In Experiment 1, we observed that as set size increased from 4 to 12 items, capacity estimates for across-category changes increased linearly as well, providing evidence against the claim of a fixed capacity. In Experiment 2, we controlled for stimulus complexity and similarity but varied the utility of ensemble representations for the change detection task. We observed significantly greater capacity when ensemble information could be used. Altogether, these results are contrary to a slot-like, fixed-object constraint on working memory capacity, and consistent with object complexity and ensemble representations strongly affecting working memory performance.

Keywords: visual working memory, ensemble representations, stimulus complexity, stimulus similarity

Introduction

Working memory enables the storage and manipulation of information over brief durations of time. It is widely considered a core cognitive system, used to support a wide variety of behaviors in day-to-day life (Baddeley & Hitch, 1974; Cowan, 2008; Schurgin, 2018). Visual working memory capacity is severely limited, with participants' performance suffering significantly if asked to remember more than 3-4 items (Luck & Vogel, 2013; Ma, Husain, Bays, 2014). Individual estimates of working memory performance have been shown to strongly correlate with scholastic achievement (Alloway & Alloway, 2010; Daneman & Carpenter, 1980) and fluid intelligence (Fukuda, Vogel, Mayr, & Awh, 2010; Oberauer, Schulze, Wilhelm, & Sü, 2005). Accordingly, a large body of research on working memory has focused on understanding the nature of these underlying capacity limitations.

A wide variety of work focuses on visual working memory for simple features or arbitrary bindings of simple features. In this literature, the debate has largely focused on the extent to which capacity limits are best characterized by a fixed constraint on how many objects can be remembered (Luck & Vogel, 1997; Woodman, Vogel & Luck, 2001; Luck & Vogel, 2013) or some form of continuous resource that becomes more thinly distributed as the number of items to be remembered increases, resulting in decreased performance when more than 3-4 items must be remembered (Wilken & Ma, 2004; Bays & Husain, 2008; Ma, Husain & Bays, 2014; Schurgin, Wixted, Brady, 2018).

This debate has led to important insights about the nature of working memory, and the model of using simple stimuli -- which are thought to prevent contamination from other memory systems (Lin & Luck, 2012) in order to allow the direct quantification of memory capacity -- has

uncovered a huge and diverse amount of features pertaining to limits on visual working memory. However, in the real world, we do not encounter arbitrary sets of simple features. Thus, it is also important to understand visual working memory for more complex objects.

Working memory for complex objects

The literature on working memory for complex objects has largely been focused in two distinct directions. One line of work, on meaningful real-world objects, has concentrated on the role of existing knowledge in supporting working memory and the potential contributions from long-term memory systems in the short-term storage of items (e.g., Brady, Störmer, Alvarez, 2016; Curby & Gauthier, 2007; Kaiser, Stein, Peelen, 2015; O'Donnell, Clement & Brockmole, 2018; Schurgin, Cunningham, Egeth, & Brady 2018).

Another line of work has continued to avoid semantically meaningful, real-world objects, but has nevertheless focused on more complex objects than the traditional visual working memory literature. For example, initial research looking at memory for more complex, but still non-meaningful objects, observed that when the items in visual working memory tasks are more complex than combinations of simple features (e.g. arbitrary polygons, 3D cubes, etc), fewer items can be remembered (Alvarez & Cavanagh, 2004). These results provided some of the first strong evidence against an object-based limit, as they showed that the complexity of objects, and therefore amount of information that needs to be held in mind to remember them, affects working memory capacity (see also Taylor et al. 2017; Luria et al. 2010).

The role of change difficulty

One potential limitation of using complex objects to study visual working memory is that as stimulus complexity increases, so does the potential similarity of foils offered at test. As a result, the decreasing working memory performance for more complex stimuli might not be the result of complexity per se, but rather increased sample-test similarity. To investigate this possibility, Awh and colleagues (2007) investigated observers' ability to detect within-category changes (e.g. a cube to another cube), where sample-test similarity was high, and across-category changes (e.g. a cube to a Chinese character), where sample-test similarity was low. They observed that performance was greater for across- relative to within-category changes, suggesting sample-test similarity was affecting working memory performance. Moreover, across-category change capacity estimates were around four objects, leading to the claim that once similarity is taken into account there may be a fixed object limit on working memory (see also Barton, Ester, & Awh, 2009; Fukuda, Vogel, et al., 2010; Scolari, Vogel, & Awh, 2008). Importantly, since with only one item in mind participants are nearly perfect at the within-category discriminations for these objects (Alvarez & Cavanagh, 2004), this claim required the introduction of a resource-like addition to the model rather than true slots – e.g., while supposedly a fixed number of items are held in mind, it must be proposed that the items are stored with varying amounts of precision (see also Zhang & Luck, 2008, which similarly rejected a pure slot model in favor of a discrete resource model). This work of Awh and colleagues (2007) has been extremely influential in sustaining the idea of a slot-like representation, even for complex objects.

The role of spatial ensemble/global texture information

In order to investigate the underlying nature of visual working memory limitations, the majority of studies share one fundamental manipulation – varying set size. By pressuring visual working

memory beyond its capacity, researchers seek to examine what may be the source of this limitation. An implicit assumption guiding these studies is that only memory information about individual items contributes to working memory performance. In fact, the vast majority of working memory studies report "*K*" values, a capacity estimate derived from the assumption that a fixed number of items is represented (according to a high-threshold theory) and that this capacity can be calculated separately across all set sizes (Cowan, 2001; Pashler, 1988). Once the number of items shown is greater than the capacity of the participant, the idea is that this "*K*" value should remain approximately fixed, and this has often been found. Usually participants are shown to remember approximately 3-4 items by this logic (e.g., Cowan, 2001).

In addition to the potential for rejecting the idea of considering 'how many' items are remembered in such a high-threshold manner (e.g.,Schurgin, Wixted & Brady, 2018), another issues with this approach is that previous research has demonstrated observers can rely on other memory sources, such as ensemble information, to inform working memory performance (Brady & Alvarez, 2011). For example, recalled locations of individual objects in a cluster are pulled toward the centroid location of the set (Lew & Vul, 2015), and participants are much more accurate at detecting changes that change not only items but also ensemble structure of a display (e.g., Brady & Tenenbaum, 2013; Orhan & Jacobs, 2013). In addition, ensemble information can inform memory for complex stimuli, such as the perceived emotion of a neutral face (Corbin & Crawford, 2018). For example, Jiang et al. (2016) found high similarity along a feature dimension (i.e. face identity) actually facilitated memory performance compared to a low similarity condition, likely by reducing noise in the memory representation via ensemble information (in contrast to predictions that more distinct items should improve memory by reducing competition, e.g., Cohen et al. 2014).

Importantly, such ensemble effects also appear to differ as a function of set size (Brady & Alvarez, 2015a). At large set sizes, where many items are present, item information is noisier, and thus models of how participants use ensemble information predict a greater reliance on such information (e.g., Brady & Alvarez, 2011). As a result, any account of working memory limitations utilizing set size manipulations needs to account for the potential role of ensemble information.

Indeed, in a replication of Awh et al. (2007) an analysis of individual displays found greater capacity estimates when items were clustered in ways that resulted in more useful global ensemble information, compared to when items were more individually dispersed throughout the display (Brady & Alvarez, 2015b). This was not a form of all-or-none perceptual grouping ("all the cubes are over there") because participants remained aware of the individual identities to a certain extent as well. Instead, the evidence suggested that people represent both individual items and information about the general spatial distribution of items in the display, perhaps via spatial ensemble (Alvarez & Oliva, 2009) or peripheral texture information (as proposed by Balas et al. 2009; Rosenholtz, 2016). This suggests when given an across-category change, observers may be utilizing ensemble information to inform their judgments, thereby inflating their working memory performance and violating another assumption of high-threshold capacity estimates.

There exist many possible versions of how ensemble information may inform across-category change performance, even accounts with explicit strategies (such as only encoding items of a particular category). Here, we focus on spatial ensemble accounts, informed by research suggesting people represent summary information about the spatial structure in simple displays (Balas et al. 2009; Brady & Alvarez, 2015). Figure 1 provides a schematic of this logic. Participants first encode a display with different stimuli categories (e.g., cubes and Chinese

characters). At test, if a change occurred, participants would be able to rely on two sources of information. If the change occurred for an item for which observers have a working memory representation, they could use this information to correctly indicate a change has occurred. However, even if the observer failed to encode information for that specific item, they could still use ensemble representations to inform their performance. For example, they may know in a particular area of the display (outlined in red) there were no cube-like features present, and subsequently use this information to correctly indicate a change has occurred. Critically, such ensemble representations (here schematized using the texture tiling algorithm of Balas et al. 2009) can be utilized primarily if a change is across-category, as this can create large texture changes between displays, but would not be as beneficial for within-category changes, given the general feature information at the changed location would be consistent between displays.

According to this ensemble representation account, as the number of categorically similar items in a display increases, this increases the potential utility of ensemble representations (e.g. if only four items are in a display, ensemble representations are less useful than when eight items are in a display). In addition, the spatial arrangement of items will affect the utility of this strategy, as displays with categorically similar items clustered in the same locations will tend to have greater capacity estimates than displays where the items are more inter-dispersed (Brady & Alvarez, 2015b). In contrast to this ensemble account, the slot-model prediction is that observers only have item representations available (or not), and can only use these to inform working memory performance, even at high set sizes.

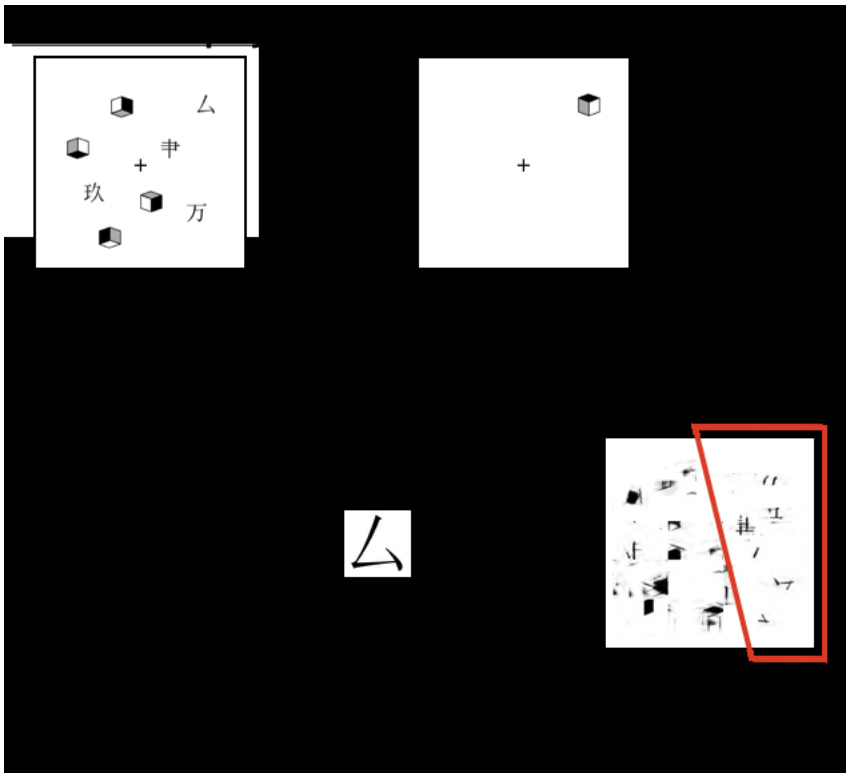

Figure 1. *Schematic of a spatial ensemble account of working memory performance. Here, participants encode a display of cubes and characters, and at test are presented a display with an across-category change (i.e. character to cube). An observer could rely on an item representation (i.e. memory of that particular character) to correctly indicate a change has occurred. However, in the absence of a strong item memory, or in addition to this memory, the observer could utilize an ensemble representation, such as they may know in a particular area of the display (outlined in red) there were no cube-like features present, and thus use this information to correct indicate a change has occurred. Critically, as the number of categorically similar items in a display increases, so too does the ability to utilize an ensemble-based strategy to inform performance. In contrast, according to a slot-model account, observers can only utilize item representations at test to inform performance.*

The current studies

We sought to expand upon these findings and directly address the role of ensemble contributions to working memory capacity for large, across-category changes, and particularly the validity of calculating a "capacity" (*K*) metric for such displays. Specifically, in Experiment 1, we replicated Awh et al. (2007) but analyzed set size separately, with a range of set sizes (4, 8, 12). By comparing performance for across-category changes as set size increased, we could

evaluate whether observers increasingly relied on ensemble information to inform their judgments. We also quantified working memory performance using *d'*, a measure of memory signal strength that does not rely on high-threshold assumptions, to understand working memory performance and the potential utilization of ensemble information at high set sizes.

In Experiment 2, we next assessed the role of ensemble information directly while controlling for the complexity and similarity of stimuli. In particular, participants completed a visual working memory task with Chinese characters and random polygons that were either outlines (less distinct from the Chinese characters in terms of spatial ensemble information) or were filled-in (more distinct from the Chinese characters in terms of ensemble information). There was no significant difference for within-category changes whether a polygon was outlined or filled-in, confirming we had controlled for the complexity and similarity of the stimuli. Thus, we could directly assess whether when stimuli contained greater ensemble information, but complexity and similarity was controlled for, if there was a difference in performance for across-category changes. Overall, we found strong evidence in favor of the idea that participants do not use a slot-like representation and that they take advantage of ensemble information to detect changes.

Experiment 1: Replication of Awh Et al. (2007) with Higher Set Sizes

In Experiment 1, we replicated Awh et al.'s (2007) finding that observers are better at detecting across-category changes than within-category changes when remembering complex objects. However, in order to assess the potential role of ensemble information in this task, we added a set size 12 condition and analyzed data separately by set size. It has been argued that similar across-category change capacity estimates of ~4 objects across stimulus type suggest a fixed (i.e.

'slot-like') constraint on visual working memory capacity. However, previous research collapsed capacity over set sizes (i.e. reporting a single capacity estimate averaged across set size 4 and 8). If visual working memory utilizes ensemble information to improve capacity, we expect to see that as set size increases, estimated capacity should not be fixed, since the underlying representations are not actually item-based in the way assumed by the capacity formula. Instead, increasing set size should increase the ensemble information available in the display, and therefore there should be an increase in estimated working memory capacity for large, across-category changes as set size increases.

Methods

Participants. A group of 22 University of California, San Diego undergraduates participated in Experiment 1. The results from one participant was excluded due to noncompliance with instructions. All participants reported normal or corrected-to-normal visual acuity, and no previous experience reading or writing Chinese/Japanese characters. Participation was voluntary, and in exchange for extra credit in related courses. The experimental protocol was approved by the University of California, San Diego IRB.

Apparatus. Experiment 1 took place in a dimly lit sound-attenuated room. Stimuli were presented on a Macintosh iMac computer with a refresh rate of 60 Hz.

Stimuli and Procedure. Stimuli were generated using MATLAB and the Psychophysics toolbox (Brainard, 1997; Pelli, 1997). Participants completed 576 trials of a visual working memory task. Displays contained either 4, 8, or 12 items (equally balanced across trials) consisting of

intermixed cubes and Chinese characters (see Figure 2). Displays always contained an equal number of cubes and characters, and stimuli were randomly selected from the 8 possible stimuli for each category during every trial. Stimuli positions were randomly selected positions within a square region (3.95° per side), constrained to a 5 X 5 grid of possible locations excluding fixation (i.e. 24 total positions). Images were 2.38°, and their positions were randomly jittered horizontally and vertically within region by up to 0.6°. There were no other restrictions on image position.

Participants would press the space bar to initiate a trial. They would then see a fixation cross for 1000 ms, followed by a study display containing intermixed cubes and characters for 500 ms. Afterward a blank delay screen was shown for 900 ms. At subsequent test, a single item reappeared in one of the previous stimulus locations (with location serving as the cue for what object was being tested) and participants had to report whether that test item was the same or different from the item that had been in that same location in the previous display. The test item was shown on screen until participants made a response. For half of the trials the test item was exactly the same as in the previous display (i.e. same), and for the other half of trials the test item changed (i.e. different). On different trials, half the changes were within-category changes (e.g., a cube to a different cube) and half were across-category changes (e.g., a cube to a Chinese character; see Figure 3). Trials were randomly interweaved across set size and type of test.

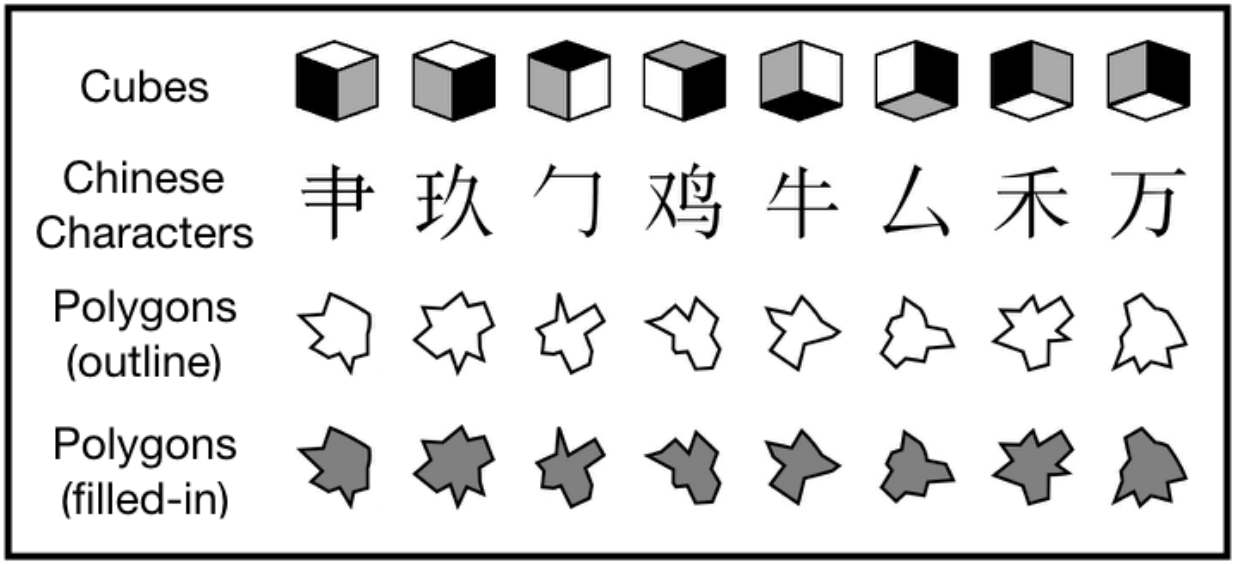

Figure 2. *Possible stimuli of Experiment 1 and 2 (adapted from Awh et al., 2007). In Experiment 1, participants saw a display containing either Cubes or Chinese Characters. In Experiment 2, participants saw a display containing Chinese Characters and random Polygons that were either outlined or filled-in.*

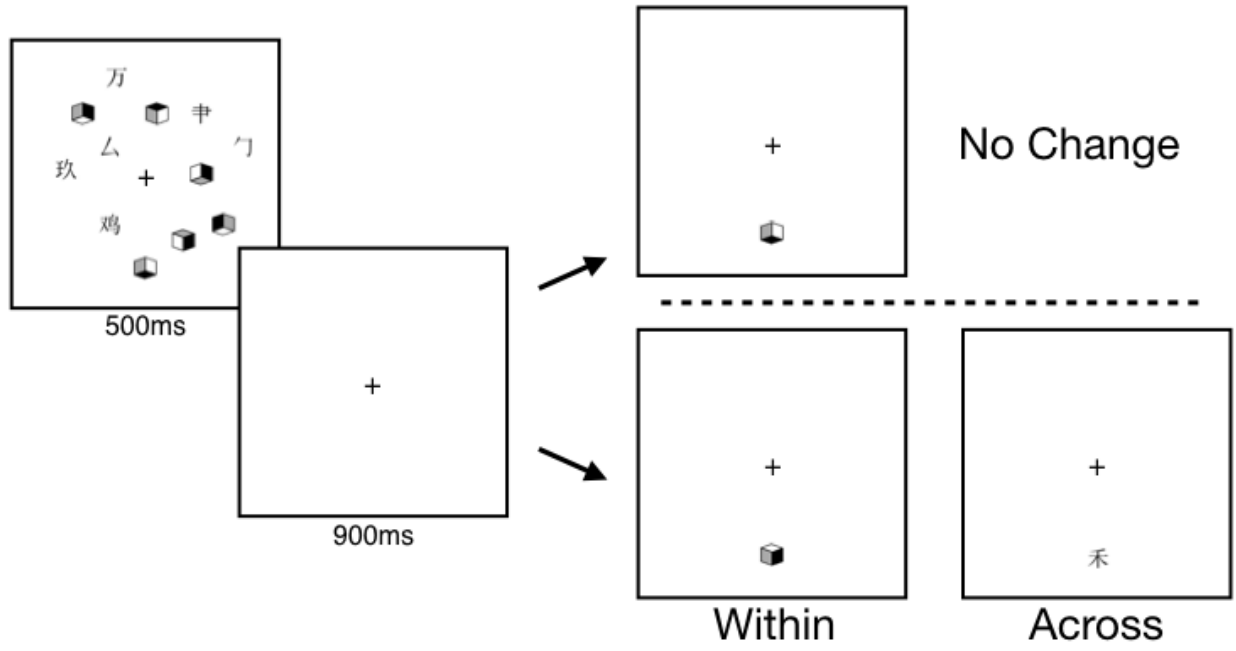

Figure 3. *Participants performed a change detection task containing half cubes and half Chinese characters at set sizes 4, 8, or 12. At test when a change occurred (50% of trials), this change was either within-category (small) or across-category (large). The test item was shown on screen until participants made a response.*

Data Analysis. To estimate visual working memory capacity that allowed for a direct comparison of Awh et al. (2007), we estimated the 'number' of individual items remembered using Cowan's K (Cowan, 2001): $K = (H - FA) * N$, where K is the number of items stored, H is the hit rate, FA is the false alarm rate, and N is the number of items presented. This is the appropriate high-threshold model for the case where only a single item reappears at test (Pashler, 1988; Rouder et al. 2011).

In addition, we report d' values (a measure of memory signal strength), to quantify performance without reliance on a high-threshold, slot-like model: $d' = Z(H) - Z(FA)$ (Macmillan & Creelman, 2004). In the present manuscript, Cowan's K is referred to as working memory 'capacity,' since it is hypothesized to be fixed across set size, whereas d' is referred to as working memory 'performance'. We also report reaction time data, to assist in the interpretation of accuracy-based analyses (excluding reaction times greater than 5 seconds).

Results and Discussion

First, we investigated whether we replicated the general effect between within-category and across-category changes observed in Awh et al. (2007). Collapsing across stimulus type and set size, we observed a capacity of approximately one object in the within-category change condition and four objects in the across-category change condition (K = 1.02 vs 3.88), $t(20) = 10.7$, $p < 0.01$. These results are consistent with a near-direct replication of previous results (albeit with the addition of a set size 12 condition). Quantifying performance, we observed a d' = 0.49 for the within-category and d' = 1.70 for the across-category change condition, respectively.

We found a similar pattern for reaction times, with slower responses for within-category (1.20 seconds) versus across-category (1.05 seconds) changes, $t(20)=5.10$, $p < 0.01$.

Our primary interest was performance in the across-category change conditions. As expected, performance dropped at greater set sizes. An ANOVA of across-category change detection performance estimates found a significant main effect of set size (4, 8 or 12), $F(2,40)=134.17$, $p<0.001$, $\eta^2 = 0.87$. Follow-up analyses revealed d' decreased as a function of set size, from 4 items (d'=2.69) to 8 items (d'=1.58) to 12 items (d'=1.14) (all p's < 0.001; see Figure 4B). For reaction time, an ANOVA also found a significant main effect of set size, $F(2,40) = 19.04$, $p < 0.001$, $\eta^2 = 0.49$, with follow-up analyses revealing reaction time increased as a function of set size, from 4 items (0.95 sec) to 8 items (1.06 sec) to 12 items (1.13 sec) (all p's < 0.01). Altogether, these results provide evidence suggesting performance decreased as a function of set size. Of critical interest was whether the difference in across-category change performance as set size increased was consistent with the high-threshold 'capacity' model as proposed by Awh et al. (2007). In particular, the 'capacity' formula (K) is designed to correct performance for the greater difficulty at higher set sizes and reveal "how many items" were remembered at each set size. If participants had a fixed K value at all high set sizes, this would support the idea that there was a fixed item limit, assuming the other requirements of a high-threshold theory were met as well. However, if this high-threshold view is incorrect, or if observers utilized ensemble information to facilitate across-category change performance, then as the number of stimuli increases this should lead to a greater reliance on non-object-based spatial ensemble information, and thus 'capacity' should increase.

An ANOVA of across-category change detection capacity estimates found a significant main effect of set size (4, 8 or 12), $F(2,40)=18.77$, $p<0.001$, $\eta^2 = 0.48$. Follow-up analyses

revealed capacity estimates increased as a function of set size, from 4 items (K=3.06) to 8 items (K=3.98) to 12 items (K=4.61) (all *p*'s < 0.001; see Figure 4A). In other words, performance as a function of set size dropped less sharply than proposed by slot models. A potential limitation of these results is the possibility of ceiling effects at set size 4, as performance was quite high. However, we found K values well below 4 (i.e., 3.06), and most importantly set size 4 is not the critical comparison in this design, as we are primarily interested in working memory performance when this system was pressured (i.e. set size 8 and 12).

Thus, despite being widely cited, the results of Awh et al. (2007) do not in fact suggest a fixed 'capacity' limit for across-category changes, even according to the model of capacity used in the paper and frequently used by proponents of slot-like views (Cowan, 2001). Consistent with previous reports, this may be because observers utilize ensemble representations to improve performance during across category-changes, particularly in displays with large set sizes (e.g., Brady & Alvarez, 2015b). In addition, capacity estimates suggested working memory performance for across-category changes actually *increased* as a function of set size, rather than remaining approximately fixed as expected by models who believe this formula accounts for how performance changes with set size (e.g., Awh et al. 2007).

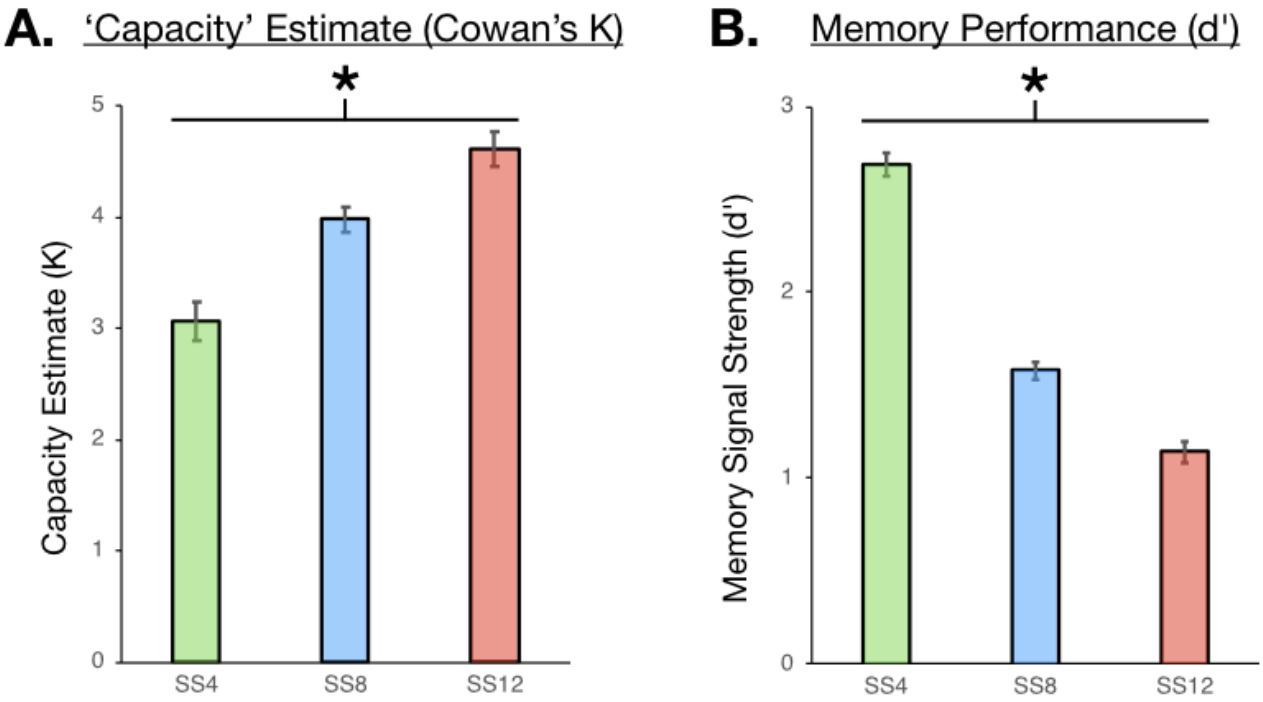

Figure 4. *(**A**) In contrast to Awh et al. (2007), which collapsed results across set size 4 and 8, we observed increasing 'capacity' estimates for across-category changes as set size increased. This may be because observers rely on ensemble strategies to detect across-category changes. (**B**) When analyzing the same across-category change data using d', a measure of memory signal strength, we observed decreasing performance as set size increased. * designates $p < 0.001$. Error bars represent within-subject error (Cousineau, 2005).*

Experiment 2: Controlling for Complexity and Similarity - Greater Capacity with Increased Utility of Texture

In Experiment 1 we observed greater visual working memory capacity estimates in across-category change conditions as set size increased, consistent with our hypothesis that observers utilize ensemble information to improve visual working memory performance, particularly at large set sizes (e.g., Brady & Alvarez, 2015b). These results are inconsistent with the fixed 4-item limit observed by Awh and colleagues (2007) and in line with more recent research suggesting observers remember far less information in across-category changes when the utility of ensemble information is reduced (Brady & Alvarez, 2015b). To pursue this claim even further, we tested the potential utilization of ensemble information while directly controlling for

other possible factors, such as stimulus complexity (Alvarez & Cavangh, 2004) and sample-test similarity (Awh et al., 2007).

Specifically, participants completed a visual working memory task across two sessions involving Chinese characters and random polygons. During one session, the random polygon stimuli were outlined, and during the other session the random polygon stimuli were exactly the same but filled-in with gray. As a result, both sessions were identical in their stimulus complexity and the similarity of their foils, but critically differed in how ensemble information may be utilized. We reasoned when polygons are filled-in, this added more distinctive, discriminative texture information that should increase the ability of observers to utilize ensemble information to inform working memory performance, compared to when the polygons were outlines. That is, peripheral texture information and other non-item based representations should be more informative for telling apart filled-in polygons from characters than outlined-polygons from characters -- even though if only one item was remembered, both would be effectively maximally dissimilar changes from a character.

Methods

Participants. A separate group of 20 University of California, San Diego undergraduates participated in Experiment 2. All participants reported normal or corrected-to-normal visual acuity, and no previous experience reading or writing Chinese/Japanese characters. Participation was voluntary, and in exchange for extra credit in related courses. The experimental protocol was approved by the University of California, San Diego IRB.

Stimuli, Apparatus, Procedure. Experiment 2 was identical to Experiment 1, with the following exceptions: Participants completed 288 trials of a working memory task. All trials were at set size 8. The stimuli in the displays were intermixed Chinese characters and random polygons. Participants completed this task twice across two sessions (taking place at least one day apart but no longer than 7 days apart). During one session the random polygons were outlined, and during the other session the random polygons were filled-in (session order counterbalanced equally across participants; see Figure 5A).

Data Analysis. The analyses used were identical to that of Experiment 1.

Results and Discussion

Collapsing across stimulus type (i.e. sessions) we observed a significant difference in performance (d') for within-category ($d' = 0.44$) and across-category changes ($d' = 1.54$), $t(19) = 9.41$, $p < 0.001$, with a corresponding difference in 'capacity' for within-category ($K = 1.31$) and across-category changes ($K = 4.13$), $t(19) = 9.58$, $p < 0.001$. We observed the same pattern for reaction times as well, with slower responses for within-category (1.17 seconds) versus across-category (1.11 seconds) changes, $t(20) = 2.34$, $p = 0.04$. Thus, we again replicated the general effect observed by Awh and colleagues (2007).

Next, in order to evaluate the potential utilization of ensemble representations affecting working memory capacity estimates, we compared performance for within-category and across-category changes across outlined vs filled-in polygons. We reasoned that both a filled-in and outlined polygon should be extremely dissimilar to a Chinese character and are thus not limited by 'precision', as in the model of Awh et al. (2007), but that they would markedly differ in the

ability of participants to make use of more spatial ensemble-based information to discriminate them.

We observed no difference for within-category changes whether the polygons displayed in the task were outlined ($d' = 0.27$, $K=0.73$) or filled-in ($d' = 0.21$, $K=0.54$), $t(19) = 0.74$, $p = 0.47$ for d', $t(19)=0.82$, $p=0.42$ for K. This pattern is consistent across reaction time as well, with similar reaction times for outlined (1.21 seconds) and filled-in (1.21 seconds) polygons, $t(19) = 1.04$, $p = 0.31$. This lack of any observable effects reinforces that the relative complexity and similarity of polygon stimuli used across sessions was equal.

Critically, we found a significant effect for across-category changes, with greater performance when the polygons contained in the display were filled-in ($d' = 1.75$, $K=4.49$) compared to when they were outlined ($d' = 1.41$, $K=3.77$), $t(19) = 3.38$, $p < 0.01$ for d', $t(19) = 3.71$, $p = 0.001$ for K (Figure 5B). A similar trend to this pattern was observed in reaction time (filled-in RT = 1.07 seconds, outlined RT = 1.16 seconds), $t(19) = 1.07$, p = 0.26. Thus, in a working memory task when the texture information was more discriminative between the Chinese characters and polygons (i.e. filled-in), participants performance (quantified either as d' or K) increased, likely due to the use of less item-based, more global ensemble strategies to increase performance.

A potential limitation of the present study is that rather than solely being based on ensemble structure, there is a potential that some aspects of the benefit for filled vs. outlined polygons could arise from a difference in across-category similarity when polygons were outlined, compared to when they were filled. Specifically, it may be that outlined polygons are slightly more similar to Chinese characters, resulting in worse performance in across-category changes compared to when those polygons are filled-in (and thus less similar). This would

require an expansion of the idea of item 'precision' (as previously proposed) such that there could be items stored that are so imprecise that they can be discriminated from filled-in polygons but not from outlined polygons, and that there could be (approximately) 5 items represented. For example, consider the high-threshold model used by Cowan's K, where if it an item is represented, a change is always detected, but if the item is not represented, participants must guess. Such a model, where there is no noise in responses but simply unrepresented items, would require that participants are storing, on average, one high-precision item (sufficient to distinguish within-category changes), three medium-precision items (sufficient to distinguish most across-category changes) and one low-precision item (sufficient to distinguish only some across-category category changes). While possible, this is clearly quite distinct from a slot-like model and from the claims of Awh et al. (2007), and is effectively a continuous resource model.

This model could be softened to keep a slot-like limit on performance but move to a more signal-detection-like explanation of responses, where rather than specifying a certain number of items are encoded vs. not encoded, items are conceived as having different memory strengths and these memory strengths do not result in all-or-none correct vs. incorrect responses. Under such a framework our results can be accounted for in a more nuanced way, with memory strength varying between items such that, on average across trials, the within-category change can be successfully detected for around 1 out of 5 memorized items, the across-category change for the outline polygons for 4 out of 5, and the across-category change for the filled polygons for 5 out of 5, with the greater performance for filled polygons arising from either their greater apparent dissimilarity to the Chinese characters or the memory strength derived from ensemble representations.

In general, we believe that previous evidence of ensemble structure being used in displays like the current experiments suggest that the usage of ensemble information is a likely explanation of the results (e.g., Brady & Alvarez, 2015; Brady & Tenenbaum, 2013), particularly in light of the set size difference in Experiment 1. In other words, we propose that participants *are* using low-precision information about several items, but these representations are not strictly item-based but also ensemble-based.

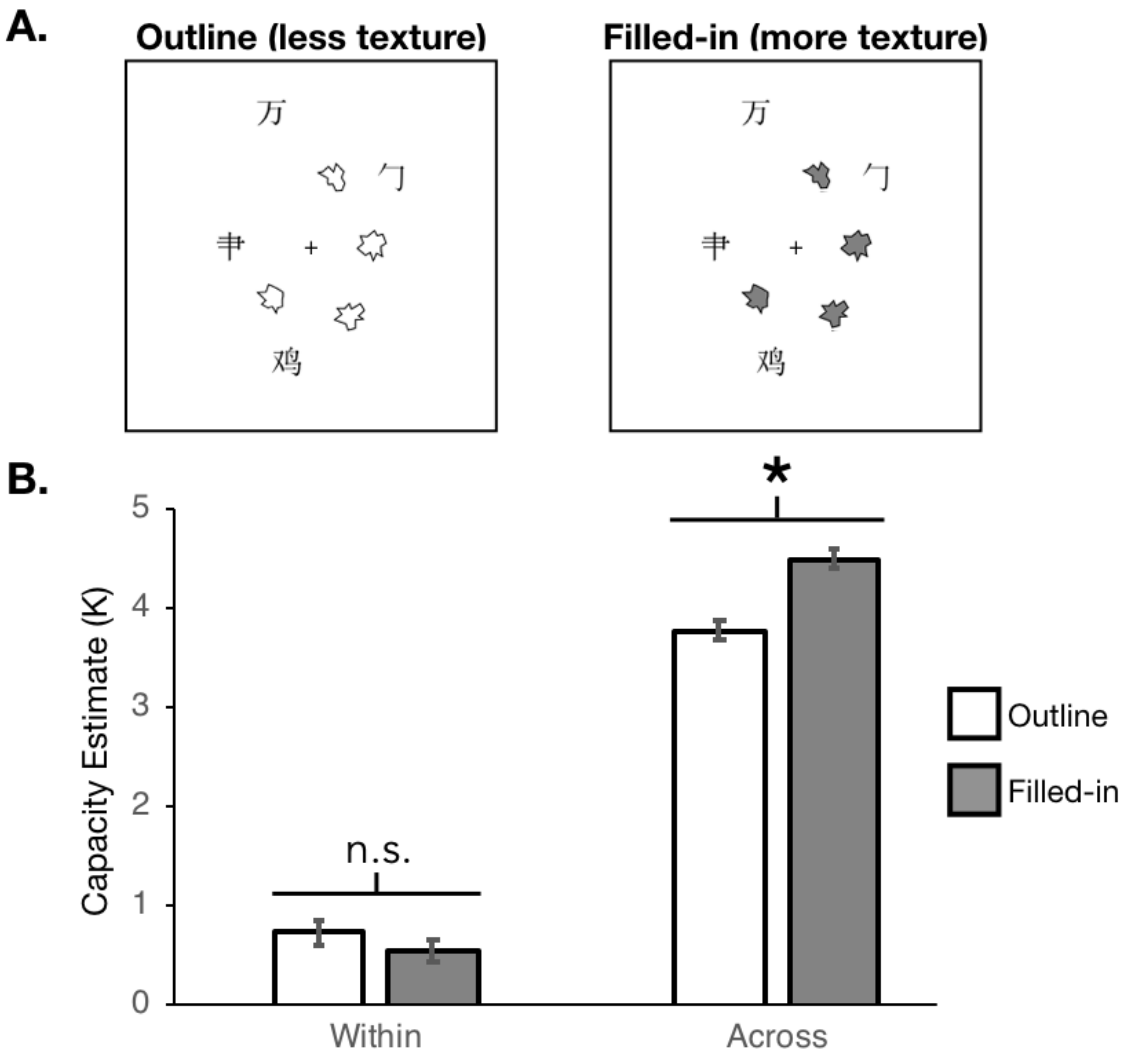

Figure 5. *(A) In Experiment 2, participants completed change-detection task at set size 8. Displays consisted half of Chinese characters and half random polygons. However, depending on the session, the polygons shown were either outlines (less discriminative from the characters in terms of texture) or filled-in (more discriminative from the characters in terms of texture).*

*Thus, similarity and complexity were controlled for across the different types of polygons, with the only difference across sessions the amount of texture information that could be utilized by a global ensemble-based strategy. (**B**) There was no difference for within-category changes for polygons whether they were outlines or filled-in, confirming complexity and similarity were controlled. However, we found significantly greater capacity for filled-in relative to outline polygons, suggesting greater texture information and the use of ensemble information drives effects observed for across-category changes. This same pattern of results was observed using d', which does not rely on high-threshold, slot-like assumptions of working memory performance. *designates p = 0.001. Error bars represent within-subject error.*

Discussion

Overall, we found that as the utility of ensemble information in a display increased, whether through increased set size (Experiment 1) or via the stimuli themselves (Experiment 2), across-category performance increased as well. In addition, we found evidence refuting the claim of previous work that there is a fixed 'capacity' for across-category changes. These results demonstrate that across-category change performance is not fixed across set size, and so it is not an informative index of if performance is based on some slot-like limit on the number of objects. Further, we provide direct evidence against a fixed object limit in visual working memory, as it appears that across-category changes are not solely an effect of low sample-test similarity but reflect the utilization of spatial ensemble information. These results are consistent with a continuous, distributed resource account of working memory, especially those continuous resource accounts that allow for ensemble-based in addition to item-based representation.

Recent evidence for the importance of stimulus complexity

Research on visual working memory has consistently shown that stimulus complexity affects performance significantly in a variety of different tasks, and this is true even in simpler stimuli (e.g., conjunctions of orientations and colors; Cowan et al., 2013, Fougnie et al., 2010, Hardman & Cowan, 2015; Oberauer & Eichenberger, 2013), and even when using continuous report

techniques, where no foil is offered and thus accounts based on the difficulties of change detection are not relevant (e.g., Oberauer & Eichenberger, 2013).

In general, there is a convergence of evidence that performance for complex objects is limited even with only 1-2 objects in mind. For example, Taylor et al. (2017) administered a change detection task to participants, but where participants completed blocks of a single stimulus type that varied in complexity relative to one another (e.g. letters, words, colors or shapes). They analyzed their data with a measurement model that gave an estimate not only for memory capacity, but for the probability of comparison errors as well. Critically, this model did not require within and across-category changes, and thus freed participants from potential issues that may arise with encoding items from different categories. They observed that capacity estimates, and not comparison error estimates, varied as a function of stimulus complexity. Furthermore, their most complex stimuli (shapes) resulted in a capacity estimates of ~1 object. It is important to note that other studies have also attempted to estimate error rates with comparing complex within-category stimuli and have found evidence suggesting complexity does not matter (Barton et al., 2009, Umemoto et al., 2010), but the specific models used to support these claims have been found to rely on incorrect assumptions and implementation (Morey, Morey, Brisson, & Tremblay, 2012).

A general performance limit with performance dropping and resource limits approached when remembering only ~1-2 complex but non-semantically meaningful objects is also corroborated by neural evidence. For example, when participants completed a change detection task containing either colored squares or polygons (similar to the stimuli in Experiment 2), researchers observed lower behavioral performance for more complex objects (i.e. polygons compared to colored squares). In addition, neural activity was collected via EEG, and a well-

known marker of perceptual maintenance in visual working memory, the sustained posterior contralateral negativity (SPCN; equivalent to contralateral delay activity), was larger during the delay period for complex objects than for simple objects, suggesting visual working memory needed to maintain more perceptual information for complex objects (despite lower overall behavioral performance) (Luria et al., 2010). Taken together, these studies provide strong evidence in favor of the idea that stimulus complexity affects visual working memory performance.

Evidence against a 'slot-model' of visual working memory

The present results provide strong evidence against a 'slot-model' of complex objects in visual working memory, which suggests people always represent information as a fixed number of objects (~4) regardless of complexity, and these items vary only in precision (e.g., Awh et al. 2007; Barton, Ester, & Awh, 2009; Fukuda, Vogel, et al., 2010). Much of this research relies on tasks with set sizes either at this theorized limit (i.e. 4 objects) or by averaging performance across set sizes. However, in our replication of Awh et al. (2007) that included set sizes 4-12 (Experiment 1) we found that when analyzing each set size separately across-category, estimates of 'capacity' increased significantly when more items were in the display (e.g., performance as a function of set size dropped less sharply than proposed by slot models). At the highest set size, we obtained across-category capacity average capacity estimates of almost 5 items in working memory, considerably higher than those reported by Awh et al. (2007) in a similar task and, most critically, with performance not reflecting a fixed 'K' across set sizes.

Overall, these results are incompatible with a potential object-based limit of visual working memory, which continues to require post-hoc amendments to explain extant data (see

Schurgin, 2018 for a review). This work adds to the growing literature suggesting that even when using all the assumptions of the discrete-resource versions of the slot model to quantify data, as in the current work, the data do not obey the regularities required of this model, like a fixed capacity estimate across set size (see also Bays 2018).

Experiment 2 provided additional evidence against 'slot'-like representations. Again, we found that performance exceeded what would be expected for a slot-like model, and that even accounting for possible 'precision' differences between items, as in previous discrete resource models (Awh et al. 2007; Zhang & Luck, 2008) does not support a fixed item limit. To account for our data such a model would require that participants are storing one high-precision item (sufficient to distinguish within-category changes), three medium-precision items (sufficient to distinguish most across-category changes) and one low-precision item (sufficient to distinguish only some across-category category changes). This is effectively a continuous resource model.

The importance of accounting for ensemble information

More broadly, these results reinforce the need to dissociate item-level and ensemble-level representations contributions to working memory performance. We are not the first to make this point, as previous research has found that ensemble statistics can inflate working memory performance in a variety of different tasks (Brady & Alvarez, 2011; Orhan & Jacobs, 2013; Brady & Tenenbaum, 2013; Brady & Alvarez, 2015b). However, this emerging and consistent evidence highlights the importance of taking into account the potential use of global ensemble information or explicit chunking strategies (Nassar, Helmers & Frank, 2018) in both visual working memory tasks and the models used to explain those tasks.

Ensemble information is especially relevant to the study of working memory, considering the vast majority of research on working memory is concerned with understanding its apparent limitations. In order to study these limitations, many experiments vary set size in ways meant to pressure or exceed working memory capacity, with conditions that include a large number of items in a display. As the number of items in a display increase, it is important that researchers understand the role ensemble information may be contributing to performance. Many researchers using high set sizes to investigate working memory limitations have not considered how ensemble information might affect performance in their tasks, potentially inflating estimates of working memory performance. This becomes even more problematic when different stimuli are used across conditions, which may have varying ensemble information available. As we have shown, even the same stimuli but with differing usefulness of texture information (i.e. outlined vs filled-in) can significantly change working memory performance in a task (inflating 'capacity' estimates by almost one item!).

It is also worth noting that the most widely used estimates of working memory capacity, Cowan's *K*, relies not only on a strong high-threshold memory assumption, but on an assumption that individual representations are the sole contribution to working memory performance. As a result, *K* estimates may not be the best measure of working memory performance, especially at high set sizes when the contribution of ensemble information likely increases.

One potential solution may be to analyze data using a signal detection framework, where the assumptions guiding memory performance are well established. Indeed, in Experiment 1 we observed, as expected, that as set size increased, *d'* decreased. This appears to more accurately reflect participants performance at high set sizes than the assumptions of Cowan's K, which

suggested that performance increased at high set sizes (above previously documented limits of a 'capacity' around 3-4 objects).

Consistent with the present results, recent research suggests that working memory performance is best captured under a signal detection framework that takes into account the psychological representation of the relevant feature space (Schurgin, Wixted & Brady, 2018). Such a model, which directly assesses the role of stimulus perceptual characteristics, could be used to more directly incorporate ensemble information into estimates of performance.

Implications for visual working memory and fluid intelligence

Previous research has found a strong correlation between visual working memory capacity and fluid intelligence, further indicating visual working memory's importance as a core cognitive system. However, this relationship has been observed only for large, across-category change performance (r=0.66), whereas no relationship between fluid intelligence has been observed for smaller, within-category changes (r=-0.05) (Fukuda et al., 2010). Previously these results were interpreted to support that the relationship between visual working memory and fluid intelligence was tied to capacity (i.e. large change performance), not the resolution of working memory representations (i.e. small change performance).

However, our results suggest an alternative interpretation to this relationship. Specifically, it may be that observers with greater fluid intelligence are better able to shift their strategy in visual working memory tasks to utilize ensemble information or to switch -- between trials -- their reliance on item vs. ensemble information. Thus, the correlation between visual working memory performance and fluid intelligence may not be directly attributable to capacity per se, but rather the ability to shift strategies in a way that maximizes the utility of ensemble

information. Future research should investigate what factors might be driving this relationship, i.e. whether this correlation is due to visual working memory capacity or strategic utilization of ensemble information.

Conclusion

The present results provide evidence against the idea that fixed item limits are present for more complex objects, and argue that observers utilizing ensemble strategies to improve working memory performance, particularly at high set sizes. As this ensemble-based strategy becomes more beneficial (e.g. high set size or via increased texture in stimuli), this inflates performance and violates one of the assumptions of capacity models of working memory that generally are based on the assumption only item representations contribute to performance. In the real-world, observers likely utilize both individual item and ensemble representations to inform their performance in working memory tasks. As a result, tasks and models of working memory need to better account (or discount) for the role ensemble representations may have in informing performance.

Acknowledgments

We would like to thank Kelvin Lam for his invaluable role in collecting data for this project. For funding, we would like to acknowledge NSF CAREER (BCS-1653457) to TFB.

References

Alloway, T. P., & Alloway, R. G. (2010). Investigating the predictive roles of working memory and IQ in academic attainment. *Journal of Experimental Child Psychology*, 106(1), 20–29.

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

Alvarez, G. A., & Oliva, A. (2009). Spatial ensemble statistics are efficient codes that can be represented with reduced attention. *Proceedings of the National Academy of Sciences*, 106(18), 7345-7350.

Awh, E., Barton, B., & Vogel, E. K. (2007). Visual working memory represents a fixed number of items regardless of complexity. *Psychological Science,* 18(7), 622–628.

Baddeley, A. D., & Hitch, G. (1974). Working memory. *Psychology of Learning and Motivation, 8*, 47–89.

Balas, B., Nakano, L., & Rosenholtz, R. (2009). A summary-statistic representation in peripheral vision explains visual crowding. *Journal of Vision*, *9*(12), 13-13.

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

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

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

Brady, T. F. and Alvarez, G.A. (2011). Hierarchical encoding in visual working memory: ensemble statistics bias memory for individual items. *Psychological Science*, 22(3), 384-392.

Brady, T. F. and Alvarez, G.A. (2015a). Contextual effects in visual working memory reveal hierarchically structured memory representations. *Journal of Vision*, 15(15):6.

Brady, T. F. and Alvarez, G.A. (2015b). No evidence for a fixed object limit in working memory: Ensemble representations inflate estimates of working memory capacity for complex objects. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 41(3), 921-9.

Brady, T. F., & Tenenbaum, J. B. (2013). A probabilistic model of visual working memory: Incorporating higher order regularities into working memory capacity estimates. *Psychological Review*, 120, 85–109.

Brainard, D. H., & Vision, S. (1997). The psychophysics toolbox. *Spatial vision*, *10*, 433-436.

Corbin, J. C., & Crawford, L. E. (2018). Biased by the Group: Memory for an Emotional Expression Biases Towards the Ensemble. *Collabra: Psychology*, *4*(1).

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

Cowan, N. (2008). What are the differences between long-term, shortterm, and working memory? *Progress in Brain Research, 169*, 323– 338

Curby, K. M., & Gauthier, I. (2007). A visual short-term memory advantage for faces. *Psychonomic bulletin & review*, *14*(4), 620-628.

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of verbal learning and verbal behavior*, *19*(4), 450-466.

Fukuda, K., Vogel, E., Mayr, U., & Awh, E. (2010). Quantity, not quality: The relationship between fluid intelligence and working memory capacity. *Psychonomic Bulletin & Review*, 17(5), 673–679.

Kaiser, D., Stein, T., & Peelen, M. V. (2015). Real-world spatial regularities affect visual working memory for objects. *Psychonomic bulletin & review*, *22*(6), 1784-1790.

Lin, P. H., & Luck, S. J. (2012). Proactive interference does not meaningfully distort visual working memory capacity estimates in the canonical change detection task. *Frontiers in psychology*, *3*, 42.

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, 390(6657), 279–281.

Luck, S. J., & Vogel, E. K. (2013). Visual working memory capacity: from psychophysics and neurobiology to individual differences. *Trends in cognitive sciences*, *17*(8), 391-400.

Luria, R., Sessa, P., Gotler, A., Jolicoeur, P., & Dell'Acqua, R. (2010). Visual short-term memory capacity for simple and complex objects. *Journal of Cognitive Neuroscience*, 22, 496 –512. http://dx.doi.org/ 10.1162/jocn.2009.21214

Ma, W. J., Husain, M., & Bays, P. M. (2014). Changing concepts of working memory. *Nature Neuroscience*, 17(3), 347–356.

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

Nassar, M. R., Helmers, J. C., & Frank, M. J. (2018). Chunking as a rational strategy for lossy data compression in visual working memory. *Psychological review*, *125*(4), 486.

Oberauer, K., Schulze, R., Wilhelm, O., & Süß, H. M. (2005). Working memory and intelligence--their correlation and their relation: comment on Ackerman, Beier, and Boyle (2005).

O'Donnell, R. E., Clement, A., & Brockmole, J. R. (2018). Semantic and functional relationships among objects increase the capacity of visual working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(7), 1151.

Orhan, A. E., & Jacobs, R. A. (2013). A probabilistic clustering theory of the organization of visual short-term memory. *Psychological Review*, 120, 297–328.

Pashler, H. (1988). Familiarity and visual change detection. *Perception & Psychophysics*, *44*(4), 369-378

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial vision*, *10*(4), 437-442.

Rosenholtz, R. (2016). Capabilities and limitations of peripheral vision. *Annual Reviews of Vision Science.*

Rouder, J. N., Morey, R. D., Morey, C. C., & Cowan, N. (2011). How to measure working memory capacity in the change detection paradigm. *Psychonomic Bulletin & Review*, *18*(2), 324-330.

Schurgin, M. W. (2018). Visual Memory, The Long and the Short of it: A Review of Visual Working Memory and Long-Term Memory. *Attention, Perception, & Psychophysics*.

Schurgin, M. W., Cunningham, C. A., Egeth, H. E., & Brady, T. F. (2018). Visual Long-term Memory Can Replace Active Maintenance in Visual Working Memory. *bioRxiv*, 381848.

Schurgin, M. W., Wixted, J. T., & Brady, T. F. (2018). Psychophysical Scaling Reveals a Unified Theory of Visual Memory Strength. *bioRxiv*, 325472.

Taylor, R., Thomson, H., Sutton, D., & Donkin, C. (2017). Does working memory have a single capacity limit?. *Journal of Memory and Language*, *93*, 67-81.

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

Van den Berg, R., Shin, H., Chou, W. C., George, R., & Ma, W. J. (2012). Variability in encoding precision accounts for visual short-term memory limitations. *Proceedings of the National Academy of Sciences*, 109(22), 8780-8785.

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*(1), 92.

Wilken, P., & Ma, W. J. (2004). A detection theory account of change detection. *Journal of Vision*, 4(12), 1120–1125.