REMEMBERING COMPLEX OBJECTS IN VISUAL WORKING MEMORY

A Thesis presented to
the Faculty of the Graduate School
at the University of Missouri

In Partial Fulfillment
of the Requirements for the Degree
Master of Cognition and Neuroscience

by
KYLE HARDMAN
Dr. Nelson Cowan, Thesis Supervisor
MAY 2013

The undersigned, appointed by the Dean of the Graduate School, have examined the thesis entitled:

REMEMBERING COMPLEX OBJECTS
IN VISUAL WORKING MEMORY

presented by Kyle Hardman,
a candidate for the degree of Master of Cognition and Neuroscience and hereby certify that, in their opinion, it is worthy of acceptance.

Dr. Nelson Cowan

Dr. Shawn Christ

Dr. Judith Goodman

Dr. Jeffrey Rouder

ACKNOWLEDGMENTS

I would like to acknowledge my advisor for help with the many revisions of this manuscript and ideas for experiments to pursue this issue. Additional thanks goes to my thesis committee for the ideas they gave me to improve the content of this thesis.

A special thanks for the financial support I received from the Life Sciences Fellowship program.

TABLE OF CONTENTS

CHAPTER

LIST OF TABLES

Table Page

LIST OF FIGURES

ABSTRACT

Visual working memory stores stimuli from our environment as representations that can be accessed by high-level control processes. This study addresses a longstanding debate in the literature about whether storage limits in visual working memory include a limit to the complexity of discrete items. We examined the issue with a number of change-detection experiments that used complex stimuli which possessed multiple features per stimulus item. Some past research that used the same methodology as our experiments found that detection of changes in stimuli was unaffected by how many features of the items participants were required to remember (Luck & Vogel, 1997). However, in none of our eight experiments were we able to replicate that result and instead found that participants were less able to detect changes when they were required to remember more features of the items. We were unable to support the notion that items with multiple relevant features can be processed by visual working memory without loss.

Chapter 1

Introduction

Working memory (WM) is a capacity-limited store for information that is actively in use or which must be maintained over a short interval (Baddeley, 2003; Cowan, 2001). One concern of WM researchers has been to specify how the constituent features of objects are integrated in visual WM into coherent internal representations of the external objects (Fougnie, Asplund, & Marois, 2010; Luck & Vogel, 1997; Treisman, 1988; Wheeler & Treisman, 2002). This study is primarily focused on the issue of whether there is a cost to processing (encoding, storing, and/or retrieving) stimuli for which there is a high feature load. Feature load will be manipulated by varying the number of task-relevant features of stimulus objects between different test conditions.

Some have argued that the features of an object are effortlessly bound to the representation of that object without any cost for additional features (Luck & Vogel, 1997; Zhang & Luck, 2008). An important piece of their evidence is the finding that increasing feature load from one task-relevant feature per object to four task-relevant features had no effect on performance (Luck & Vogel, 1997). The model that came out of this finding is often described as the slot model of WM, in which humans have a limited number of slots which can be filled with discrete stimuli (i.e. individual objects) until there are no more slots available (Zhang & Luck, 2008). The results of Luck and Vogel have often been used as evidence that coherent objects with strongly integrated features are the basic unit of storage in WM, with the implication that the number and/or complexity of the features which make up an object can be ignored when interpreting results; although possessing multiple features, a multi-featured object still only takes up one slot in WM.

Another perspective on feature binding is that coherent objects are created by an attentionally-demanding process that involves assembling objects from features that are stored independently from one another (Treisman, 1988; Wheeler & Treisman, 2002). In this theoretical framework there exist separate feature maps for each feature dimension (e.g. the color map only stores information about object colors) which are capacity-limited within maps, but not between maps, which accounts for findings showing that participants can remember a greater total number of features when those features are drawn from different feature dimensions versus when the features are all from the same dimension (Wheeler & Treisman, 2002). Once features are stored in these feature maps, retrieval of a coherent object comes about by focusing attention on a location, which causes the feature stored at the attended location in each map to be combined with features from the attended location from other feature maps (Treisman, 1988).

Although the results of Luck and Vogel (1997) are very striking, they are not without controversy. Both Wheeler and Treisman (2002) and Delvenne and Bruyer

(2004) failed to replicate the results of one of the feature-conjunction experiments of Luck and Vogel (1997) in which bicolored squares were used. Although Luck and Vogel found no deficit to performance when participants were held responsible for both of the colors of each square, Wheeler and Treisman and Delvenne and Bruyer both found just such a deficit. As far as we are aware, there have been no successful replications of the feature-conjunction experiments of Luck and Vogel (1997). Given the recent focus on problems with replicability of results in psychology and elsewhere (Pashler & Wagenmakers, 2012), another attempt to replicate the results of Luck and Vogel seems warranted.

Additionally, there have been a number of studies showing that there are storage costs associated with features in visual WM (Alvarez & Cavanagh, 2004; Cowan, Blume, & Saults, 2012; Fougnie et al., 2010), suggesting that if the results of Luck and Vogel replicate, they may not generalize to other experimental conditions. If the results of Luck and Vogel can be replicated, we could examine what features of their methods allowed them to obtain such a result while others could not with different methods.

Importantly, no experiments have attempted to replicate a critical result of Luck and Vogel (1997), in which objects possessing four features drawn from different feature dimensions were used (see Figure 2.1 for an example of the stimuli). Although Wheeler and Treisman (2002) used two features per item, the features were from the same feature dimension (color), which prevents their result from being generalized to cases in which the features of the items are drawn from different feature dimensions. Another reason for the focus on this four-feature experiment is that the use of more than two independent features per object is uncommon in the literature. In order to help resolve the issue of feature integration in visual WM, we have attempted to replicate this important result of Luck and Vogel (republished with additional methodological detail as Experiment 14 in Vogel, Woodman, & Luck, 2001). In this task, participants had to determine if a single feature of one object had changed between a sample array and a test array. In single-feature trial blocks, participants knew which one of the four features could change in that block, allowing them to selectively attend to that feature at encoding. However, in the critical multi-feature (or conjunction) block, participants did not know which feature might change and had to attend to all four features.

The question is whether visual WM is limited solely by the number of objects that can be held, or if it is also limited by the complexity of these objects (operationally defined as the number of features of the object that must be known in order to perform perfectly on the task). If strongly integrated objects are the basic unit of storage in WM, it would be expected that, as long as the number of objects in the array is the same, participant accuracy would not vary with the number of features they are required to remember, which was the result observed by Luck and Vogel (1997). In that study, no differences in accuracy were found between the single- and multi-feature conditions or between any of the single-feature conditions. If object complexity (i.e. feature load) matters, it would be expected that accuracy in the multi-feature condition would fall below the average difficulty of the single-feature conditions.

We tested these predictions by attempting to directly replicate the results of Luck and Vogel (1997), which we did in Experiment 1. This experiment used the same change-detection task, stimuli, timings, and secondary verbal load task as the original experiment. Then in Experiments 2 through 6 we performed several confirmatory experiments using the same stimuli in order to rule out a variety of nuisance variables that could have explained our results. Then in Experiment 7 we attempted to extend our findings to somewhat different stimuli by attempting to replicate the results of another experiment of Luck and Vogel.

Chapter 2

Replication attempt and follow-up experiments

2.1 Experiment 1

This experiment was our best attempt at a direct replication of the four-feature experiment performed by Luck and Vogel (1997). The method is as similar as can reasonably be expected, although some minor differences are mentioned.

2.1.1 Method

The experiments reported in this article involve a change-detection procedure with a number of methodological features in common, for which reason some statements about general methodological details are made in this section. Most of the experiments hew closely to the method of this first experiment and details specific to each experiment are described in that experiment's method section.

Participants

Participants were recruited from introductory psychology courses at the University of Missouri – Columbia campus and received partial course credit for participation.

Unless stated otherwise, for all experiments participants were removed from the sample if their accuracy fell below 55% accuracy on at least one trial block (see the procedure for the definition of a trial block). This criterion was designed to remove participants who were performing near chance in at least some conditions. The focus on individual blocks was decided on because while overall accuracy on the tasks tended to be relatively high, there was a distinct pattern of accuracy in many participants' data that seemed to indicate that those participants were not making an attempt to perform the task to the best of their abilities in a consistent manner (i.e. very good performance on some trial blocks while performance on other trial blocks is at chance level). This pattern generally involved at least one trial block on which accuracy was very near chance, which informed our use of the 55% cutoff.

For this experiment, two participants were removed for meeting this low accuracy criterion. An additional participant was removed for having a very high error rate on the secondary verbal load task (37% of their responses were errors compared with a 6% overall average error rate). This left 19 participants (12 female; mean age 19.3) who were used in the analysis.

Materials

The experiments were performed using E-Prime 2 experimental software (Psychology Software Tools, Pittsburgh, PA) on PCs using CRT monitors running at a resolution of 1024 x 768 pixels. For this experiment, the monitor used a refresh rate of 75

Hz. Given that the monitor's refresh period was 13.3 ms, it was not possible to use presentation times in increments of 100 ms as used in the original experiment. The most important timing difference was that the sample array was presented for only 93.3 ms. However, because each participant was presented with every combination of conditions, there is little potential for this presentation time difference to affect the differences between conditions, although overall accuracy may be slightly shifted. In all other experiments, the refresh rate of the monitors varied between 60 Hz and 75 Hz. Again, because each participant completed all conditions on a single computer, there is no potential for the conditions to be differentially affected by the variations in refresh rate. In the procedure section, nominal presentation durations – as would have been achieved by 60 Hz monitors – are given.

In each trial, participants were presented with a sample array of 2, 4, or 6 visual objects about which they would be tested later. The objects possessed four features: orientation (vertical or horizontal), color (red or green), length (short or long), and the presence or absence of a black "gap" in the middle of the rectangle. The objects were rectangles with a length of 2.0° (long objects) or 1.0° (short objects) and a width of 0.15° of visual angle. The gaps were the same width as each object and 0.25° long. Objects were separated by at least 2.0° of visual angle center-to-center to reduce the chance of objects touching. The objects were presented in an area of the screen taking up 9.8° (horizontal) by 7.3° (vertical).

The colors of the objects will be reported as an ordered triple of the red, green, and blue components of the colors, which were 8-bits per component and so varied from 0 to 255 for each component, where a higher number indicates a greater amount of that component. In all experiments the background on which the objects were presented was a shade of grey and the gaps in the objects were always the darkest black that the monitors we used were able to display (RGB: 0,0,0). For this experiment, the background was a light grey (214,214,214) and the objects were either red (255,21,37) or green (66,181,70). These values were identical to those used in Figure 1 of the digital version of Luck and Vogel (1997). However, as reported in Vogel et al. (2001), originally the background was a dark grey with luminosity 8.2 cd/m^2. When measured on a representative monitor used for experiments in our lab using a TSL2561 (Texas Advanced Optoelectronic Solutions, Plano, TX), the luminosity of the background used in this experiment was 123 cd/m^2. In Experiments 4 and 5, we used a darker background and found no effect on the pattern of accuracy between background luminosities.

A sample array and a test array were used on each trial (see Figure 2.1). The test array was identical to the sample array on half of the trials. On the other half of the trials, a single feature of a single object was changed to a different value. For some trial blocks, only one of the four features was allowed to change (single-feature blocks). In the critical multi-feature block, any of the features were allowed to change, but it was still the case that only one feature of one object was allowed to change on any given trial. Object location was held constant between sample array and test array.

Procedure

Participants were tested in a sound-attenuated booth under observation of an experimenter who read the instructions for the task to the participant. Once participants had completed the first set of practice trials, the experimenter left the booth and monitored the rest of the session by way of a video camera and microphone in the booth. The instructions informed the participants about the rules governing the presentation of stimuli in order to assist them to perform optimally.

To begin each trial, participants fixated on a two-digit number presented centrally for 500 ms before the screen was blanked for 1000 ms. Then the sample array was presented for 100 ms before the screen was again blanked for 900 ms. After this retention interval, the test array was presented until participants made a same/different response by pressing ”S” or ”D” on a standard US keyboard. After giving their response, participants were cued to say the number they had seen at the beginning of the trial, with responses coded correct or incorrect by the experimenter. This secondary verbal load task was intended to prevent verbal recoding of visual stimuli. The effect of this verbal load task is further examined in Experiment 2. The procedure for a single trial is shown in Figure 2.1.

The presentation of the test array in this experiment was slightly different than the presentation used by Luck and Vogel (1997). In their experiments, the test array was removed after 2000 ms, but the participant was still required to make a response. In all of our experiments, the test array was presented until a response was made. This is very unlikely to have had any effect because in this experiment, only 6% of response times were longer than 2000 ms. This percentage is similar across our experiments.

Participants performed four single-feature trial blocks and one multi-feature trial block, the order of which was counterbalanced across participants using a Latin square. Each trial block began with a screen of instructions indicating which feature or features of the objects should be attended in the coming trial block. Upon

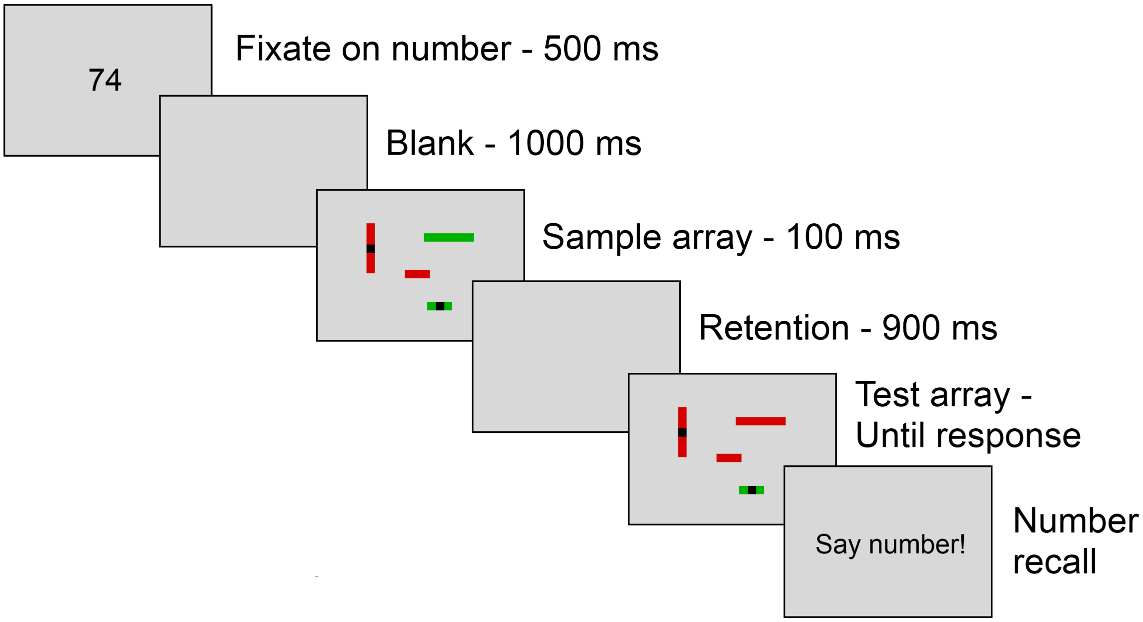

Figure 2.1: An example of a single trial in Experiment 1.

reading the instructions and indicating their intent to continue, participants were given six practice trials after which they were presented with an indication that they had finished the practice trials and were starting the main block. Within a trial block, the number of objects in the arrays varied unpredictably from trial to trial but there were always the same number of trials at each array size. In this experiment, there were 96 trials per trial block. For all experiments, each participant's experimental session lasted no more than one hour.

2.1.2 Results

In keeping with the data analysis procedure of Luck and Vogel (1997), we removed trials on which the spoken number was incorrect, which resulted in the removal of 6%

of trials. Data were analyzed with a 5 (attended feature) X 3 (array size) univariate within-participants ANOVA. In this experiment, there was a main effect of array size on accuracy, $F(2,\ 36) = 84.22$, $MSE = 0.0091$, $p < .001$, $\eta_p^2 = .82$. There was also a main effect of attended feature , $F(4,\ 72) = 21.94$, $MSE = 0.0169$, $p < .001$, $\eta_p^2 = .55$. Finally, an interaction between array size and attended feature was found, $F(8,\ 144) = 2.831$, $MSE = 0.0065$, $p < .01$, $\eta_p^2 = .14$.

The data for this experiment are summarized in Figure 2.2. Because most of the experiments in this study are very similar in design, Figure 2.2 shows the data from several experiments. The data are presented in a variety of ways. For each experiment, accuracy in each attended feature condition is shown at each array size. Additionally, hits and correct rejections are plotted for each attended feature condition at each array size. Within the multi-feature condition, hits for trials on which the change was in each feature dimension are also plotted by array size. Finally, a compound measure of the difference in detection of changes when accounting for response bias is used to compare each single-feature condition and the corresponding feature within the multi-feature condition is plotted. This compound measure is a difference in hits minus false alarms for each feature between the single- and multi-feature conditions. Hits minus false alarms were calculated for each single-feature condition and separately for each feature within the multi-feature condition. Finally, hits minus false alarms for the features within the multi-feature condition were subtracted from the hits minus false alarms for the single-feature conditions.

The interaction between array size and attended feature can be largely attributed to a ceiling effect for some single features that is present at array size 2 but not at other array sizes. This is supported by the strip chart of individual participants' mean

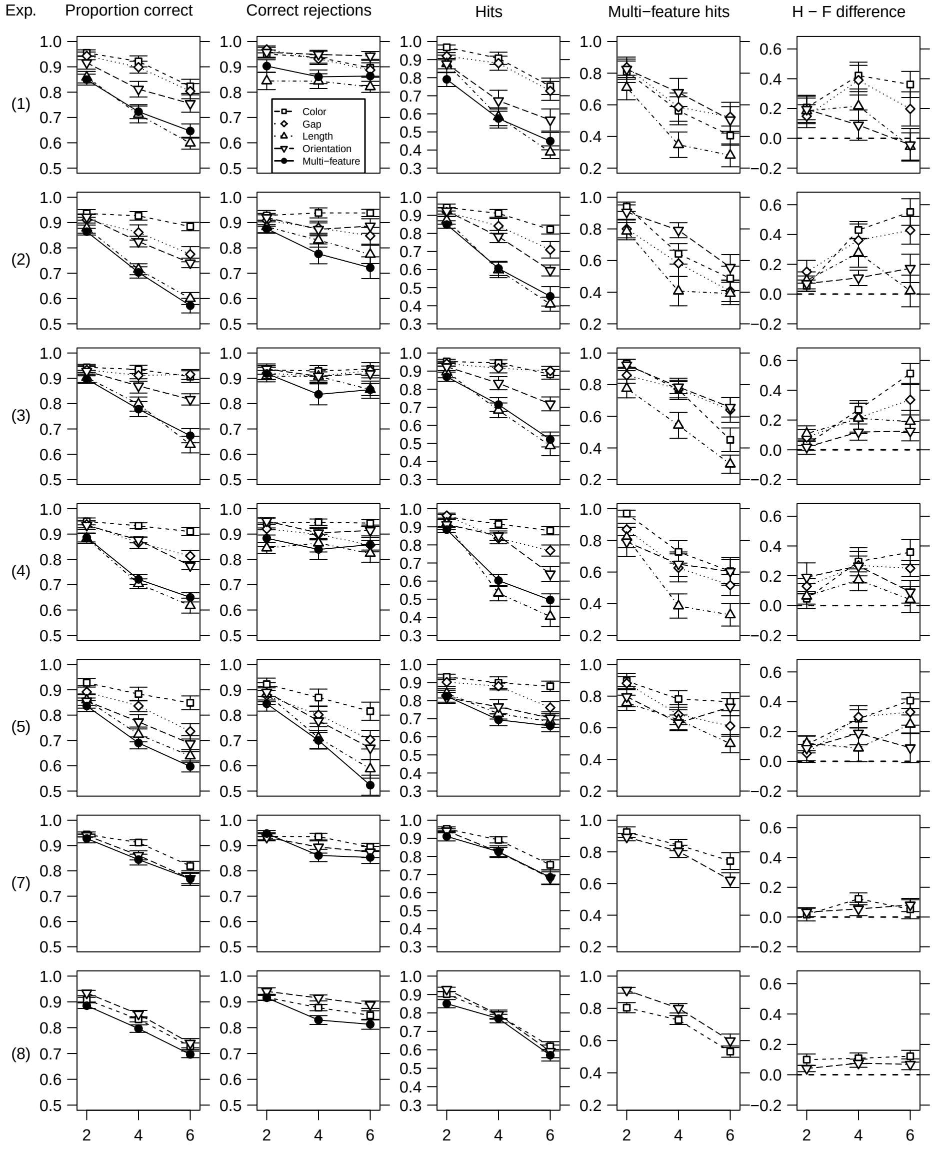

Figure 2.2: Plots of data from Experiments 1- 5, 7, and 8 (in numbered rows). The X-axis shows array size. The values shown on the Y-axis are described by the headers shown at the top of each column of plots, explained here. Proportion correct: Correct response rate. Correct rejections: Rate of correct responses to trials on which there was no change. Hits: Rate of correct responses to trials on which there was a changed feature. Multi-feature hits: Hit rate for trials on which the given feature changed within the multi-feature condition. H - F difference: Hits minus false alarms for each of the single-feature conditions minus hits minus false alarms for the corresponding feature within the multi-feature condition. Note that the scale of the Y-axis varies. Error bars are SEM.

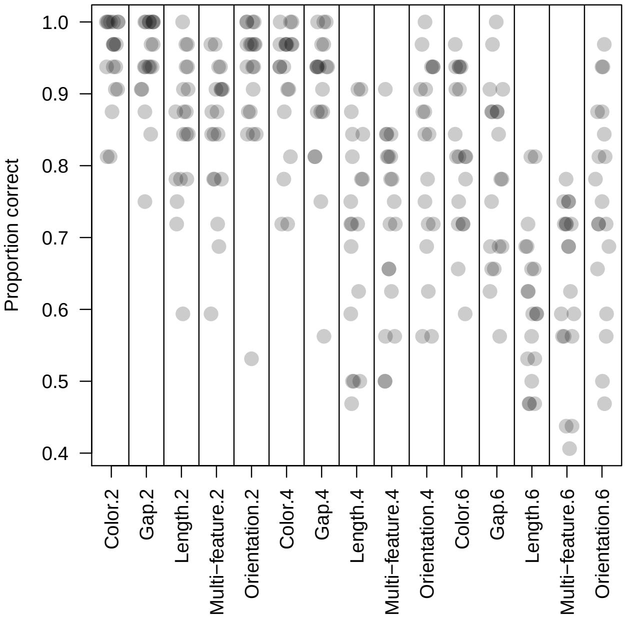

Figure 2.3: Stripchart of individual participant accuracy for each attended feature and array size for Experiment 1. The data are horizontally jittered and overlapping points are represented with darker colors to show areas of high observation density.

accuracy by array size and attended feature shown in Figure 2.3. For this reason, the interaction is not assumed to be the result of an interesting mental process and post hoc analyses will proceed as if there were no interaction.

A Newman-Keuls post hoc analysis of attended feature collapsed across array size showed that the multi-feature condition differed from orientation, color, and gap, but not length. The color and gap conditions were also not different. All other conditions were different from one another. All comparisons were made with a $p < .05$ criterion.

In order to assess the contribution of object load to performance on this task, we would compare two conditions which differed in object load but were equated in feature load (Wheeler & Treisman, 2002). If object load does not contribute to accuracy on the task, we might conclude that feature load determines accuracy. In this experiment, we do not have a condition that allows us to do this direct comparison. However, a comparison that gives useful information can be made between the multi-feature condition at array size two (MF2) and the average of the single feature conditions are array size six (SF6). The MF2 case has two objects with four features each, resulting in eight features total and the SF6 case has six objects with one feature each. In order to carry out this analysis in the most straightforward way, we would like to have observed data in the single-feature condition at array size eight, but we did not. However, we can still do a one-directional test of accuracy based on the assumption that accuracy will not improve as array size increases. If accuracy in MF2 and SF6 are equivalent or accuracy in SF6 is better than MF2, we cannot conclude anything without extrapolating in order to estimate what we might have observed in SF8. We are unwilling to do this sort of extrapolation to specific values. If, on the other hand, accuracy in the SF6 condition is worse than in the MF2 condition we can fairly safely conclude that accuracy in the SF8 condition (unobserved) would have been worse than in the MF2 condition. We are willing to believe that accuracy will not increase as array size increases from six to eight in the single-feature conditions because this kind of increase is not known in the literature. Accordingly, we will restrict ourselves to only interpreting cases in which better accuracy was obtained in MF2 than in SF6.

In order to make it possible to find evidence for the null hypothesis that there is no difference between MF2 and SF6, we used a Bayesian *t*-test to compare those conditions (Rouder, Speckman, Sun, Morey, & Iverson, 2009) in addition to a standard *t*-test. The hypotheses were the standard point null that $\mu_1 = \mu_2$ and the alternative was $\mu_1 \neq \mu_2$ (exactly the same as a standard t-test). For this experiment, accuracy in the multi-feature condition at array size two ($M = 0.85$, $SD = 0.09$) and the average accuracy of the single-feature conditions at array size six ($M = 0.74$, $SD = 0.08$) were found to differ, $t(36) = 3.66$, p ¡ .001 (two-tailed). The Bayes factor for this comparison was 5.34, which favors the alternative over the null.

2.1.3 Discussion

The results of this experiment were strikingly dissimilar from those of Luck and Vogel (1997), who found no difference between any of the attended feature conditions. Using the same stimuli and methods as the original experiment, we found differences between many of the conditions. Most importantly, the post hoc tests showed that accuracy in the multi-feature condition was lower than three of the four single-feature conditions. Additionally, differences between the single-feature conditions were found in this experiment but not by Luck and Vogel. This result contradicts the results of Luck and Vogel and rejects the hypothesis that objects in visual WM are stored with all features intact. It is not immediately obvious why we were unable to replicate the results of Luck and Vogel. We made every attempt to bring our methods in line with those reported by Luck and Vogel, even extracting additional methodological detail from Vogel et al. (2001). Over the next several experiments, we attempt to replicate our own result using a variety of minor (and major) changes to the method in order to rule out the possibility that we obtained an unusual sample in this experiment or that there was an error in our methods that caused us to fail to replicate the results of Luck and Vogel.

Our examination of object load while controlling feature load showed that it does not seem to be possible to support the idea that feature load by itself is able to fully account for accuracy. This is in contrast to the finding of Wheeler and Treisman (2002) that object load does not affect accuracy if feature load is equated, suggesting that object load is irrelevant to performance. One caveat of this analysis is that in the single-feature conditions object load and feature load are entirely confounded. As such, we cannot tell if accuracy in the single-feature conditions drops off because participants run out of object slots or because participants run out of feature-specific storage space. It is possible that in the multi-feature condition at array size two, participants are able to fill all four feature-specific stores with a small amount of information relative to the capacity of the stores. However, in the single-feature conditions they may run out of storage for that particular feature. This possibility would allow our results to be interpreted without reference to object load. However, this does not seem to be a complete explanation for our inability to find the same substantive result as Wheeler and Treisman. In that study they used array size six as a single-feature condition which they compared to a multi-feature condition with fewer objects, showing that it should have been possible for us to find their result with the array sizes we used. Another reason for the difference in findings is that the experiment in which they found their result differed significantly from ours in that their objects possessed two different colors. By drawing features from the same feature dimension, they may be examining a different effect than we are. What we will conclude is that it may not be generally true that feature load can wholly account for WM behavior.

What we have found so far is that we can support neither objects nor features as the sole determining factor of accuracy in visual WM tasks. However, we were unwilling to conclude with this as our only experiment and we continued forward attempting to replicate our result in a variety of ways.

2.2 Experiment 2

The purpose of this experiment was to determine if the verbal load task has a significant effect on accuracy in this particular task. Research by Morey and Cowan (2004) showed that secondary verbal loads consisting of two digits do not have an effect on accuracy in visual WM tasks similar to those used in this study. However, Morey and Cowan did not investigate how verbal load affected accuracy in a task that required binding together features of visual objects. Binding information in visual WM may be affected differently by secondary verbal loads than item information, perhaps interacting with the type of memory required. If this is the case, the choice to use secondary verbal loads and the nature of those loads must be carefully considered for these confirmatory experiments. If not, the use of such a task may be discontinued, benefiting both participant and researcher.

2.2.1 Method

In this experiment, data from all 16 participants (10 female, mean age 18.6 years) were used.

This experiment differed from Experiment 1 by the removal of the verbal load task. Instead of fixating on a number, participants in this experiment fixated on a small cross in the center of the screen. The blank interval following fixation in Experiment 1 was important as it allowed time for participants to begin passively rehearsing the number. Because this experiment had no such secondary task, this blank interval was removed in order to increase trial density. As the results will show, this manipulation had no effect on the pattern of results. Because no secondary task was used, there was no need for an experimenter to monitor participants during their session, so the monitoring was discontinued for this and all following experiments. Because participants did not need to be monitored for this and all further experiments, the sessions no longer took place in a sound attenuated booth, but in a private testing room. Participants performed 120 trials per attended feature condition.

2.2.2 Results

The analysis for the experiment was carried out in the same way as for Experiment 1. There was a main effect of array size on accuracy, $F(2,\ 30) = 144.2$, $MSE = 0.0049$, $p < .001$, $\eta_p^2 = .91$. There was also a main effect of attended feature, $F(4,\ 60) = 49.81$, $MSE = 0.0069$, $p < .001$, $\eta_p^2 = .77$. Finally, an interaction between array size and attended feature was found, $F(8,\ 120) = 9.945$, $MSE = 0.0042$, $p < .001$, $\eta_p^2 = .40$. The data for this experiment are summarized in Figure 2.2.

To further analyze the differences between Experiments 1 and 2, we have the option of performing a 2 (experiment) X 5 (attended feature) X 3 (array size) mixed ANOVA, where attended feature and array size are within-participant variables and experiment is a between-participant variable. If there were a difference in accuracy for attended feature conditions that changed between the experiments, that difference would manifest itself in this analysis as a two-way interaction between experiment and attended feature. If such an interaction were found, it would be necessary to specify the nature of the interaction by performing tests of simple effects. One approach would be to test simple effects of attended feature within both levels of experiment. The result of this test would be essentially equivalent to the post hoc tests already conducted in each experiment. The main difference is that the mixed ANOVA approach would have significantly reduced power to detect differences, first because the power to detect interactions is limited and second because a proper simple effects analysis controls Type 1 error at a level below that used by the post hocs we have used. Because the mixed ANOVA approach would be less likely to reject the null hypothesis that the experiments are not different than an informal comparison of the pattern of post hoc differences, we have chosen to simply compare the pattern of post hocs from each experiment rather than carry out the mixed ANOVA in order to have increased power to detect potential differences. This works against our desire to show that when changing various details of the experiments we still find the same basic pattern of results.

Accordingly, a Newman-Keuls post hoc analysis showed that the length and multi-feature conditions were not different, that the gap and orientation conditions were not different, but that all other conditions were different from one another. This result differs from Experiment 1 only in the shifting of relationships between the individual feature conditions. Specifically, color and gap were different in Experiment 1 whereas orientation and gap are different in this experiment. More importantly, the relationship between the multi-feature condition and the single-feature conditions did not change.

For this experiment, accuracy in the multi-feature condition at array size two ($M = 0.86$, $SD = 0.05$) and the average accuracy of the single-feature conditions at array size six ($M = 0.75$, $SD = 0.06$) were found to differ, $t(30) = 5.66$, p ¡ .001 (two-tailed).

The Bayes factor for this comparison was 6.52.

2.2.3 Discussion

As our results show, the pattern of results in this experiment is qualitatively similar to the pattern seen in Experiment 1 with the only difference being a change in the pattern of accuracy in some of the single-feature conditions. However, these changes are not central to the issue at hand, that issue being the question of whether verbal load differentially affects binding and item information. It is clear that the multi-feature and length conditions are still equivalently difficult and are both the most difficult conditions, which is no departure from Experiment 1. This result indicates that there is no reason to continue using the verbal load task in its current form in this type of experiment, so we have chosen to discontinue the use of such a task for further experiments. Although we had the option of increasing the verbal load and examining the effects of such a manipulation, we chose instead to neglect the contributions of verbal memory for this set of experiments with the possibility of continuing this line of research in the future. Due to the rapid presentation of stimuli and short maintenance period, it is questionable if verbal recoding is generally an effective strategy at all. It is even more questionable whether any advantage in accuracy achieved through verbal recoding would be worth the cost of the additional effort required in order to enact such a strategy.

2.3 Experiment 3

The purpose of this experiment was to determine what effect sample array presentation time has on the pattern of results we have observed. Although sample array presentation time was previously ruled out by Luck and Vogel (1997) as a significant contributor to accuracy, because of the striking differences between our results and theirs we were interested to see what effect it might have on the patterns of results we were obtaining.

The relationship between accuracy in the multi-feature condition and the most difficult single-feature condition (length) might be explained by the results of Vogel, Woodman, and Luck (2006), who found that there was a minimum amount of time needed to consolidate a WM representation. If the amount of time it takes to encode an object is limited by the most-difficult-to-encode feature, it could be that when participants are attempting to encode all the features of each object in the multi-feature condition, their accuracy is limited by the amount of time it takes to encode the lengths of the objects, length being the most difficult single feature in our experiments. If participants are given a much longer encoding time, then encoding should no longer be a bottleneck and accuracy in the multi-feature condition would not be limited by the most difficult single feature if encoding time is in fact a limiting factor of accuracy.

2.3.1 Method

This experiment differed from Experiment 2 by increasing the sample array presentation time to 500 ms. The blank interval between sample and test was maintained at 900 ms.

Data from 15 participants (7 female, mean age 20.6 years) were used in this experiment. Three additional participants' data were removed for failing to meet the single-block accuracy cutoff.

2.3.2 Results

There was a main effect of array size on accuracy, $F(2,\ 28) = 72.93$, $MSE = 0.0044$, $p < .001$, $\eta_p^2 = .84$. There was also a main effect of attended feature, $F(4,\ 56) = 28.66$, $MSE = 0.0082$, $p < .001$, $\eta_p^2 = .67$. Finally, an interaction between array size and attended feature was found, $F(8,\ 112) = 9.99$, $MSE = 0.0047$, $p < .001$, $\eta_p^2 = .42$. The data for this experiment are summarized in Figure 2.2.

A Newman-Keuls post hoc test showed that the multi-feature and length conditions were not different and that the color and gap conditions were not different, with all other pairwise comparisons showing differences between conditions. This pattern of differences is identical to that found in Experiment 1.

For this experiment, accuracy in the multi-feature condition at array size two ($M = 0.9$, $SD = 0.07$) and the average accuracy of the single-feature conditions at array size six ($M = 0.82$, $SD = 0.06$) were found to differ, $t(28) = 3.39$, p ¡ .01 (two-tailed). The Bayes factor for this comparison was 4.88.

2.3.3 Discussion

Increasing the encoding time fivefold did not meaningfully affect the pattern of accuracy on this task. This confirms that the sample array presentation time of 100 ms used in these experiments is sufficient for encoding, pointing to maintenance, retrieval, or decision processes as limiting factors for accuracy on this task.

2.4 Experiment 4

This experiment was performed in order to determine if the background color on which the objects are presented affects accuracy. As mentioned in the method for Experiment 1, the background colors initially used by us differed from the values reported by Vogel et al. (2001) because we based our color values on a figure in Luck and Vogel (1997). Presumably, the figure was modified for better visibility in a print format and did not reflect the actual color values that were used. The most important difference was that the background color we used for Experiments 1 through 3 was far brighter than was reported in Vogel et al. (2001). We performed this experiment to determine what effect changing the brightness of the background would have. The color values for the red and green objects were also changed somewhat in order to maintain high contrast between the objects and the background.

2.4.1 Method

Data from 13 participants (12 female; mean age 18.4 years) who took part in this experiment were used in the analysis. Three additional participants failed to meet the accuracy criterion and their data were removed from the analysis.

The method of this experiment was identical to 2 except for which stimulus color values were used. The RGB values of the colors used in this experiment were as follows: background (50, 50, 50), red (255, 0, 0), and green (0, 255, 0).

2.4.2 Results

There was a main effect of array size on accuracy, $F(2,\ 24) = 109.6$, $MSE = 0.0041$, $p < .001$, $\eta_p^2 = .90$. There was also a main effect of attended feature, $F(4,\ 48) = 55.64$, $MSE = 0.0049$, $p < .001$, $\eta_p^2 = .82$. Finally, an interaction between array size and attended feature was found, $F(8,\ 96) = 9.525$, $MSE = 0.0032$, $p < .001$, $\eta_p^2 = .44$. The data for this experiment are summarized in Figure 2.2.

A Newman-Keuls post hoc analysis showed that the length and multi-feature conditions were not different, that the gap and orientation conditions were not different, but that all other conditions were different from one another. This pattern is the same found in Experiment 2 and is not meaningfully different than that found in Experiment 1.

For this experiment, accuracy in the multi-feature condition at array size two ($M = 0.89$, $SD = 0.06$) and the average accuracy of the single-feature conditions at array size six ($M = 0.78$, $SD = 0.04$) were found to differ, $t(24) = 5.24$, p ¡ .001 (two-tailed). The Bayes factor for this comparison was 4.03.

2.4.3 Discussion

The results of this experiment are very similar to the previous experiments, indicating that our results were not dependent on the specific lightness of the background that were used. However, there still exists the possibility that another color combination would result in a different outcome.

2.5 Experiment 5

In this experiment, we modified the procedure by having participants make a change-detection judgment about a single object, rather than the whole array. Finding the same basic pattern of results while significantly modifying the way in which participants' memory was tested would strengthen our previous results. One advantage of this experiment is that the absolute number of decisions that participants are required to make is reduced when participants are responding to a single item probe versus a full array probe (Luck & Vogel, 1997). However, in this experiment participants are still required to make four times as many decisions in the multi-feature condition versus the single-feature conditions, so this experiment only goes part of the way to controlling for decisions at test between conditions.

2.5.1 Method

Of the 26 participants who participated in this experiment, eight were removed for falling below the 55% single-block accuracy criterion, leaving 18 (10 female; mean age 18.9 years) to be used in the analysis.

This experiment is identical to Experiment 4 except for the way in which participants were tested. The method used to present a single object to participants at test was to replace irrelevant objects with a location placeholder. This was done by replacing all but one of the objects in the test array with an unfilled white circle in the location of the original object that was presented in the sample array. The presentation of location information about the irrelevant objects would allow participants to identify the target object in the context of the array. This was important because features were allowed to repeat within a given array, leaving location as the only unique identifier of each object. On any given trial, there was a 50% probability that one feature of the probed object would change.

2.5.2 Results

There was a main effect of array size on accuracy, $F(2,\ 34) = 117.8$, $MSE = 0.0057$, $p < .001$, $\eta_p^2 = .87$. There was also a main effect of attended feature, $F(4,\ 68) = 25.03$, $MSE = 0.0106$, $p < .001$, $\eta_p^2 = .60$. Finally, an interaction between array size and attended feature was found, $F(8,\ 136) = 4.607$, $MSE = 0.0044$, $p < .001$, $\eta_p^2 = .21$. The data for this experiment are summarized in Figure 2.2.

A Newman-Keuls post hoc analysis showed that the multi-feature and length conditions were not different and that the orientation and length conditions were not different, with all other conditions different from one another. This pattern is again very similar to that found in Experiment 1, with the only difference being a shift in the relationships between the single-feature conditions.

For this experiment, accuracy in the multi-feature condition at array size two ($M = 0.84$, $SD = 0.09$) and the average accuracy of the single-feature conditions at array size six ($M = 0.73$, $SD = 0.09$) were found to differ, $t(34) = 3.65$, p ¡ .001 (two-tailed). The Bayes factor for this comparison was 8.17

2.5.3 Discussion

Although in this experiment there was a constraint on the number of decisions that had to be made, the results are very similar to all of the preceding experiments.

However, as stated before, the participants were still required to make more decisions in the multi-feature condition, so this experiment does not wholly control for decision errors. This issue will be explored in more detail in Experiment 6.

Chapter 3

Decision error and estimating the total number of objects in mind

3.1 Experiment 6

This experiment was intended to extend the results of the previous experiments by using a different method of controlling which features of the objects participants were required to attend to. The method is similar to that used by Cowan et al. (2012) in which participants are cued at test to a specific feature dimension in which a change may have occurred. In this experiment we will sometimes present cues at test and sometimes before presentation of the sample array, which will allow for a direct comparison of accuracy between conditions in which participants are able to use information about the target feature at encoding versus conditions in which participants are only made aware of the target feature at test. This comparison is a direct test of whether or not objects are the basic unit of storage in visual WM. If objects are the basic unit of storage, there would be no advantage for the condition in which participants are cued before seeing the sample array because the same number of objects would be stored regardless of when the cue is given. However, objects are not the basic unit of storage, it is expected that selective attention to the target feature at encoding would allow more information about that feature to be stored, resulting in improved accuracy.

We have chosen to use a single-item probe as it has the advantage of limiting the number of decisions that a participant must make when giving a response. By cuing both a feature and an item, only a single decisions will have to be made. A benefit of this design is that it allows us to examine the possibility that the limiting factor in the multi-feature conditions was that participants had to perform a feature-by-feature search of the test array. If participants did not know what feature might have changed, they may have needed to perform this serial search of the test array for a difference from the representation they held, which would result in reduced accuracy if the search was slow enough for the representation to lose fidelity before search terminated. If this were the case, the difference between the single- and multi-feature conditions could have been due to memory search, not storage.

3.1.1 Method

Participants

Thirty participants (19 female; mean age 18.4 years) who participated in this experiment were used in the analysis. Two additional participants were removed for falling below the accuracy criterion.

Materials

The stimulus objects in Experiment 6 were the same as those used in Experiments 1 through 3. Like in Experiment 5, only a single object was presented at test and all other objects were replaced with white, unfilled circles.

The placement of the stimulus objects in arrays were changed in order to reduce collisions between objects, with a new distance of 2.25° between the centers of objects. Collisions were possible using the previous settings because although the objects were at most 2.0° long and center-to-center distance was held at 2.0°, the long rectangles were greater than 2.0° corner-to-corner. This lead to some occurrences of overlapping, touching, or nearly touching objects.

Procedure

Once given instructions, participants performed a short practice block of 18 trials with the experimenter observing. The practice trials included two trials with each combination of cued feature and cue presentation point (discussed further below) with array size selected randomly. Participants then completed five trials blocks with rest periods in between. Each trial block had identical instructions and within each block there were trials of each cue type at each array size.

The sequence of a single trial in this experiment differed from past experiments by the addition of textual cues that indicated which feature was allowed to change on that trial. The feature cues were a single word (e.g., ”Color”). If the participant was not cued to a specific feature at a given point, they were shown a neutral cue, which was a series of dashes (”– – – – – – – – ––”). Cues (including neutral cues) were presented at fixation and test on all trials and were presented 4.65° below the

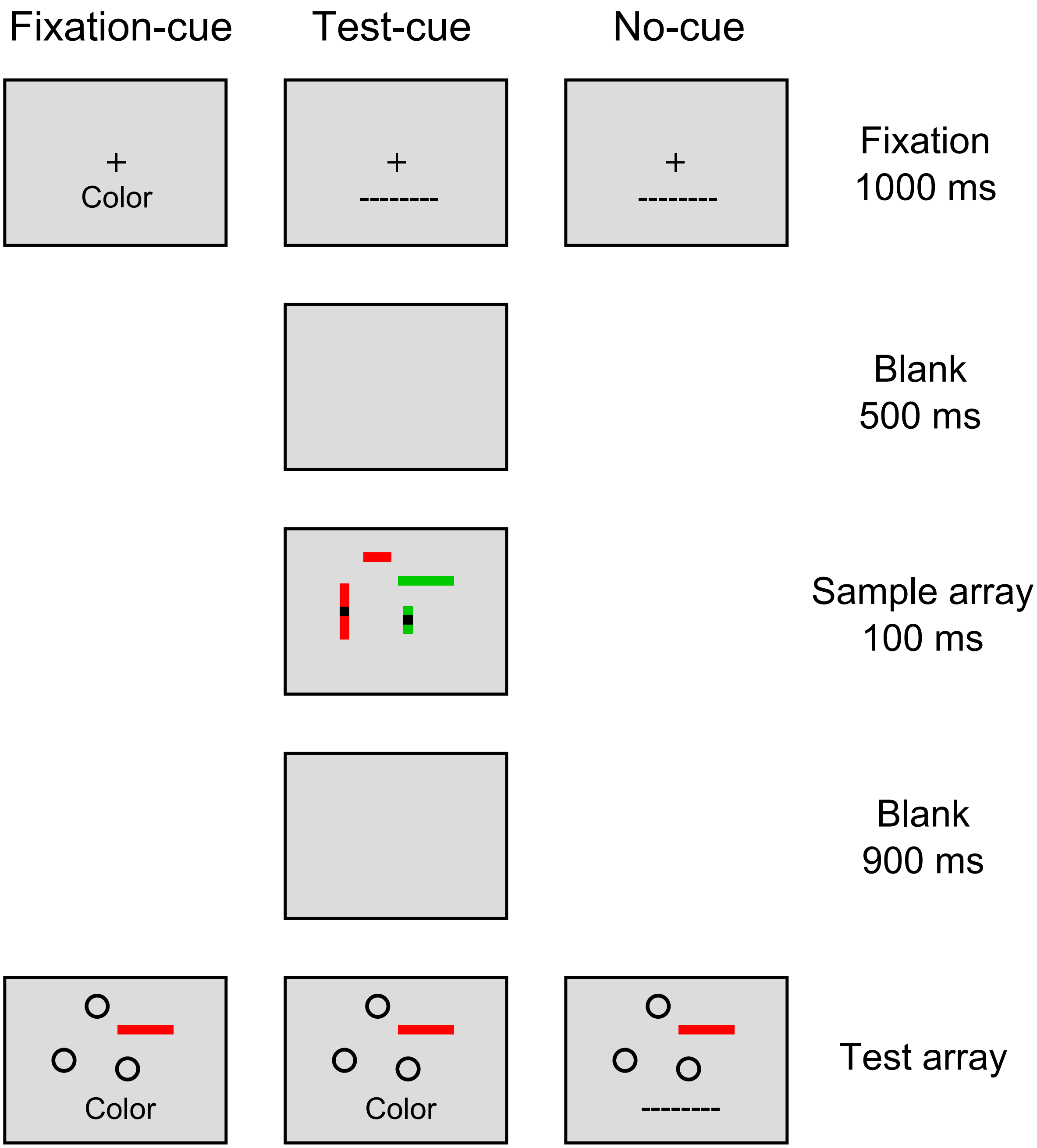

Figure 3.1: Example of the task used in Experiment 6. The three cuing conditions are demonstrated by showing the combinations of cues shown at fixation and at test for each condition. The post-fixation blank, sample array, and retention interval were the same for each condition.

center of the screen. Participants were to interpret the cue words as a fragment of the phrase ”The [cue word] of this object is ______” and to fill in the blank by responding ”the same” or ”different” using the same keys as our previous experiments.

On some trials, which we will call fixation-cue trials, participants were presented with a feature-specific cue at fixation, which allowed participants have full knowledge of the feature that would later be tested while they were encoding the test array. These trials will be similar to trials in the single-feature trial blocks of our previous experiments in that participants will have full knowledge of the target feature throughout the trial. In order to be clear, for fixation cue trials the same cue word was always presented again at test. On other trails, participants were given a neutral cue at fixation and another neutral cue at test, which we will call the no-cue condition. This condition is similar to the trials in the multi-feature blocks in prior experiments due to the fact that any feature may change and the participant has no information about which feature may have changed. We will call these no-cue trials. Finally, in the test-cue condition, participants were given a neutral cue at fixation but a feature-specific cue at test. This condition, when compared to the fixation-cue condition, will allow us to determine if knowledge about the target feature at encoding causes an increase in accuracy, a finding which would not be predicted if objects are the basic unit of storage. In each of the five trial blocks, there were four trials of each trial type at each array size, so there were $4 * 9 * 3 = 108$ trials per block and a total of 540 trials per participant.

3.1.2 Results

First, all cuing conditions were compared with a 3 (cuing condition) X 3 (array size) univariate ANOVA. The fixation-cue and test-cue conditions were collapsed across cued feature for this analysis. There was no interaction between cue presentation point and array size, $F(4,\ 116) = 1.937$, $MSE = 0.0049$, $p = .11$, $\eta_p^2 = .06$. Predictably, there was a main effect of array size, $F(2,\ 58) = 119.92$, $MSE = 0.0059$, $p < .001$, $\eta_p^2 = .81$. There was also a main effect of cuing condition, $F(2,\ 58) = 41.31$, $MSE = 0.0062$, $p < .001$, $\eta_p^2 = .59$. A Newman-Keuls post-hoc analysis of the effect of cuing condition showed that all three conditions were different from one another. Accuracy was the best in the fixation-cue condition and worst in the no-cue condition. A plot of the data used in this analysis can be found in Figure 3.2, panel A.

An additional analysis of accuracy on the individual features was performed in order to determine if the pattern of accuracy found for the individual features was similar to our previous experiments. In this analysis, the no-cue condition was removed, allowing the data to be analyzed with a 2 (cuing condition) X 4 (cued feature) X 3 (array size) ANOVA. There was no three-way interaction, $F(6,\ 174) = 1.283$, $MSE = 0.0116$, $p = .27$, $\eta_p^2 = .04$. There was a two-way interaction between feature and array size, $F(6,\ 174) = 3.216$, $MSE = 0.0083$, $p < .01$, $\eta_p^2 = .10$. There was also a two-way interaction between cuing condition and array size, $F(2,\ 58) = 4.075$, $MSE = 0.0109$, $p < .05$, $\eta_p^2 = .12$. The was also a two-way interaction between cuing condition and feature, $F(3,\ 87) = 4.663$, $MSE = 0.0084$, $p < .01$, $\eta_p^2 = .14$. There was a main effect of array size, $F(2,\ 58) = 141.40$, $MSE = 0.0143$, $p < .001$, $\eta_p^2 = .83$. There was a main effect of feature, $F(3,\ 87) = 24.14$, $MSE = 0.0191$, $p < .001$, $\eta_p^2 =$

.45. Finally, there was a main effect of cuing condition, $F(1,\ 29) = 26.85$, $MSE = 0.0184$, $p < .001$, $\eta_p^2 = .48$. Plots of the data can be seen in Figure 3.2, panels D and E.

The interactions between array size and both cuing condition and feature appear to both be driven by the fact that there is a relatively small decrease in accuracy for the color and gap features as array size increases in the fixation-cue condition. Thus, when we collapse across features, there would be an increasingly large delta between the fixation-cue and test-cue conditions at larger array sizes for some features, as can be seen in Figure 3.2, panel F. Similarly, if we were to collapse across cuing condition, there would be a similar pattern of increasing delta between the color and gap features and the two other features. Although this interpretation sounds like a three-way interaction and no such interaction was found by the ANOVA, it could be that such an interaction was simply not found due to reduced power to detect interactions in ANOVA. The number of participants used for the experiment was chosen to allow for detection of main effects, not high-order interactions.

The interaction of cuing condition and feature was examined with a Newman-Keuls post hoc test. We first tested for differences between features within each cuing condition. For the fixation-cue condition, accuracy for the gap feature did not differ from accuracy for the length feature, but accuracy for all other features differed from one another. Accuracy for the color feature was the best and accuracy for the length feature was the worst. This pattern is comparable to that observed in the single-feature blocks of previous experiments. A slightly different pattern was found in the test-cue condition, for which gap did not differ from length or orientation, with all other comparisons showing differences. The only difference in the relative accuracy for

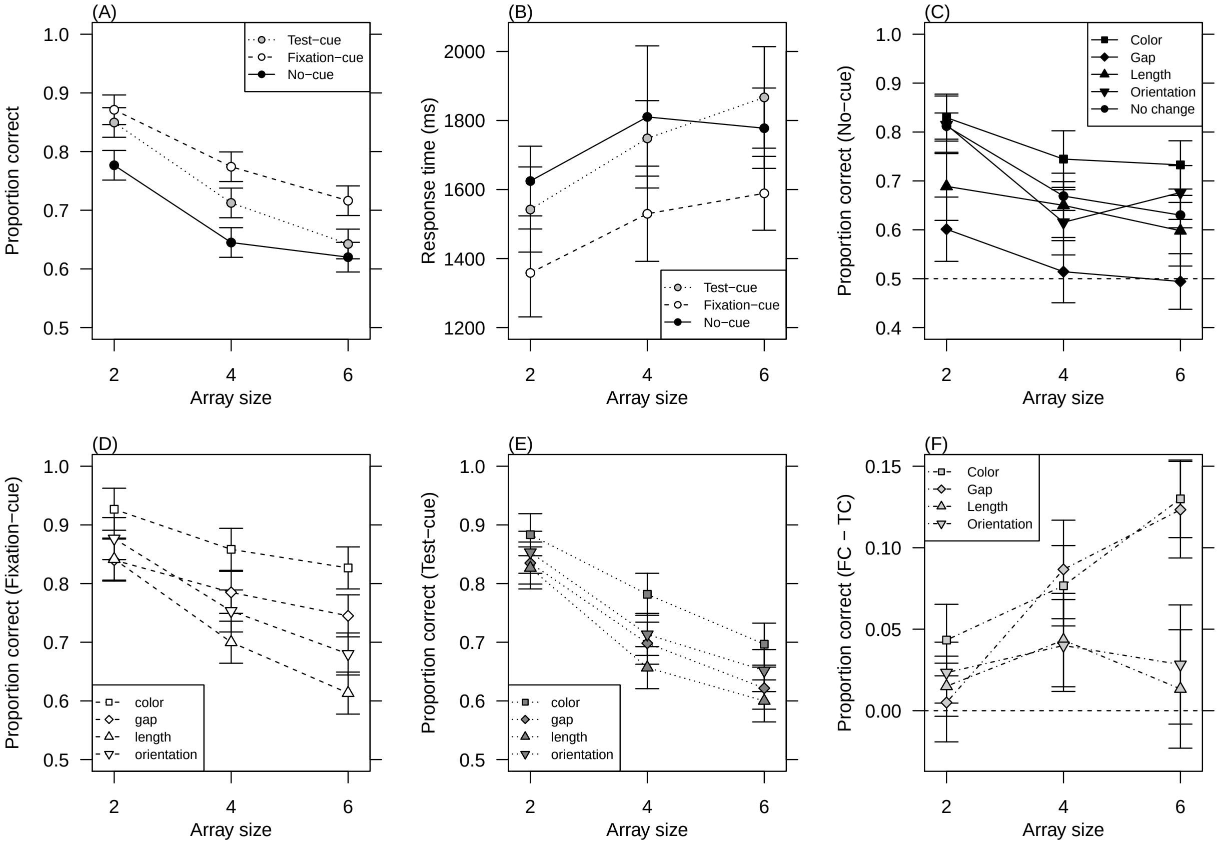

Figure 3.2: Plots of data from Experiment 6. (A) Proportion of correct responses by cuing condition collapsed across features. (B) Response times by cuing condition collapsed across features. (C) Proportion of correct responses by changed feature within the no-cue condition. Proportion of correct responses by cued feature in the fixation-cue (D) and test-cue (E) conditions. (F) Differences in proportion correct by feature between the fixation-cue and test-cue conditions. Error bars for panels A, D, and E are 95% repeated measures confidence intervals (Hollands & Jarmasz, 2010), others are standard error.

the features between the cuing conditions is that gap is not different from orientation in the test-cue condition. The last set of comparisons is between each feature in one cuing condition and the same feature in the other cuing condition, showing a change in accuracy for the feature between conditions. In this comparison, only accuracy for orientation did not change between the fixation-cue and test-cue conditions while accuracy for all other features was diminished in the test-cue condition relative to fixation-cue. The two-way interactions involving array size were not of interest to the research question and are not reported here.

Given that the fixation-cue and test-cue conditions should be as similar as possible in order to allow for clear interpretation of the results, we examined response time data for this experiment in order to determine if there were response time differences between the cuing conditions. The cuing conditions were analyzed with a 3 (cuing condition) X 3 (array size) ANOVA. There was a main effect of cuing condition, $F(2, 58) = 13.1$, p ¡ .001. A Newman-Keuls post hoc showed that the fixation-cue condition had a shorter response time than either the test-cue or no-cue conditions. Response times fixation-cue condition averaged across array sizes were 227 ms faster than response times in the test-cue condition. Response time data is plotted in Figure 3.2, panel B.

Testing for an effect of object load was conducted in a slightly different way than in the previous experiments. In this experiment, the fixation-cue condition should offer us data about accuracy when participants only attend to a single feature of the objects and the test-cue condition offers data about accuracy when participants have to attend to all of the features of the objects. These two cuing conditions were averaged across cued feature for this analysis. Average accuracy in the test-cue condition at array size two ($M = 0.85$, $SD = 0.10$) and the average accuracy in the fixation-cue condition at array size six ($M = 0.72$, $SD = 0.11$) were found to differ, $t(58) = 4.94$, $p = 0$ (two-tailed). The Bayes factor for this comparison was 14.1.

3.1.3 Discussion

The design used in this experiment allowed for a direct test of the hypothesis that feature load has no effect on accuracy. By comparing accuracy on fixation-cue and test-cue trials, it can be determined if participants can improve their accuracy by using information about which feature they will be tested on. If participants have unlimited feature storage capacity, it would not matter whether they are cued before encoding or at test. However, as the results clearly show, it does matter when participants are cued. In particular, accuracy was worse when, at the time of encoding, participants did not know which feature of the objects they were to be tested on. This result is predicted very clearly if it is assumed that feature storage is capacity-limited. If participants do not know which feature they will be tested on at the time of encoding, they have to be able to store information about every feature in order to perform optimally. Given that accuracy in such a case is reduced relative to a case in which participants only needed information about one feature, it is reasonable to think that feature storage is capacity-limited.

One possible explanation for the results that still allows for unlimited-capacity feature storage is that in the fixation-cue condition participants know which feature they will be tested on before they are shown which object they are being tested on. However, in the test-cue condition, participants have to determine both which object and which feature they must make a decision about at test. This added time in the test-cue condition could allow the WM representation to degrade slightly, reducing accuracy for reasons unrelated to storage. Given that participants took more time to respond in the test-cue condition relative to the fixation-cue condition, this explanation is plausible. The reason why participants tool longer in the test-cue condition could simply be that at test, they were presented with the cue word and the test object simultaneously, which would require them to look at both the cue word and the test object before making their decision. In the fixation-cue condition, the cue word presented at test was always the same as the cue presented at fixation, participants did not need to look at the cue word presented at test in order to perform the task. An experimental manipulation that might rule out this possibility would be to move the presentation time of the cue word presented at test in the current experiment forward in the trial in order to have the word be presented slightly before the test object is presented. This would allow participants in the test-cue condition to know which feature they should be making their decision about as soon as the test object was presented, just as in the fixation-cue condition.

The finding that accuracy in the no-cue condition was worse than that in the test-cue condition indicates the importance of performing this experiment. The only difference between those two conditions was that the number of features which participants were tested on differed. This means that there is a factor which affects accuracy when participants are required to make decisions about multiple features of an object at once. One possibility is that participants scan each feature in serial order when making a decision, which would force some features to wait longer before they are scanned. This would allow the quality of the stored representation to degrade, causing errors on later features. Alternatively, there could be a component of decision error that results from the difficulty of integrating information about multiple features when making a decision. Whatever the factor that caused the difference between the no-cue and test-cue conditions is, it was likely also present in the multi-feature condition of our previous experiments and may have led to an underestimate the ability of WM to store multi-feature objects in those experiments. This experiment helps to control for this problem and allows for a more clearly-interpretable result than our previous experiments.

3.1.4 Estimation of total number of objects in WM in test-cue condition

In order to extend the results of Experiment 6, we performed a further analysis of that experiment's data. The purpose of the analysis was to determine the number of objects held in WM for which at least one feature was known when participants were required to attend to all of the features of the objects in the test-cue condition. By using data from the test-cue condition of Experiment 6 it is possible to get an estimate of the number of objects in WM for which the cued feature was known by using Cowan's k (Cowan, 2001). This capacity estimate is not necessarily the total number of objects in WM, just those objects for which the cued feature was known. It is easily possible that a participant could have additional objects in WM for which the cued feature was not known, but a test on knowledge of only a single feature would be unable to determine the presence of knowledge of those other objects.

Although we can get estimates for the number of objects in WM for which each individual feature is known, these estimates cannot be combined into an overall estimate of the number of objects in WM without knowledge about the probability of knowing one feature of an object given that other features of the object are known. By making the assumption that knowledge of one feature is independent from knowledge of other features, it becomes possible to combine together the separate estimates of objects for which each individual feature was known into an estimate of the total number of objects held in mind in the test-cue condition. Conveniently, some recent evidence suggests that the features of an object are wholly independent from one another (Vul & Rich, 2010). This allows the analysis to be not only possible without having to guess at the relationship, but also mathematically straightforward.

This type of analysis has been done before by Cowan et al. (2012) with a task that used objects with a color and a shape. By taking advantage of the independence assumption, they were able to show that participants were able to remember the same number of objects when they knew which feature of the object would be cued at test as when they knew that either feature could be tested. The purpose of our analysis is to attempt to extend this finding of Cowan et al. to a situation in which participants are required to know either one or four features of the objects. It appears that our participants are able to preferentially allocate resources to a specific feature when they know that they will be tested on that feature (fixation-cue performance is better than test-cue performance). However, the flexibility of this resource allocation is not clear from our existing analysis. To be specific, based on our current results we cannot tell if participants encode more objects in the fixation-cue condition than in the test-cue condition or if they encode the same number of objects in either case, but have a higher probability of encoding the cued feature of the remembered objects in the fixation-cue condition. If the number of objects held in mind in the test-cue and fixation-cue conditions is the same, then an object limit in WM would be supported.

Given the independence assumption, a general purpose equation can be used to find the number of objects for which any given combination of features was known, given that we know the number of objects for which each of the features was known. This equation is

$$K_{combination} = N \prod_i \left(\frac{k_i}{N}\right) \prod_j \left(1 - \frac{k_j}{N}\right), i \in \{\text{Known features}\}, j \in \{\text{Unknown features}\}$$

where $K_{combination}$ is the number of objects in the array for which the given combination of features was known, N is the number of objects in the sample array, and K_i and K_j are the number of objects for which the ith or jth feature was known. The ratio K_i/N represents the proportion of the objects in the sample array for which the ith feature was known (and the same is true for K_j). The expression $1 - K_j/N$ represents the proportion of the sample array for which the jth feature was not known. By taking the product of the proportions of the sample array for which some features were known and other features were not known, we are given the proportion of the sample array for which the given combination of known and unknown features held. Multiplying this proportion by N simply gives the number of objects for which the combination of features was known. For example, if there are 4 objects in the sample array and the following features are known for the respective number of objects: color, 3; orientation, 2; gap, 2; length, 1. We calculate the number of objects for which color and orientation are known (gap and length are not known) as follows:

$$K_{color,orientation} = 4(3/4)(2/4)(1 - 2/4)(1 - 1/4) = 9/16.$$

By using this equation to calculate every possible combination of known and unknown features and then summing across all cases where at least one feature is known, we can estimate the total number of objects in mind for which at least one feature is known. We will use data from each participant in the test-cue condition and denote the number of objects for which at least one feature was known $K_{testAny}$ ("any" meaning that any feature or combination of features of the $K_{testAny}$ objects was known). This estimate will be compared to the maximum number of objects for which the cued feature was known in the fixation-cue condition (denoted $K_{fixationMax}$). $K_{fixationMax}$ can be thought of as an object (or slot) limit that does not depend on which feature was attended. Although it may not be possible for different features to equally fill the available slots, the easiest feature for each participant in the fixation-cue condition is our best estimate of the number of object slots they have available. Thus, by comparing $K_{fixationMax}$ and $K_{testAny}$ we are testing to see if there is an invariance in the number of objects that participants store. If the number of stored objects is identical, the number could be interpreted as a limit to the number of objects which could be stored in WM.

At array size four, the average of the estimated number of objects for which at least one feature was known in the test-cue condition ($K_{testAny}$) was 3.39 (SD 0.68) and the average of the maximum number of objects for which the cued feature was known in the fixation-cue condition ($K_{fixationMax}$) was 3.10 (SD 0.74). A standard paired t-test was unable to distinguish these conditions, $t(58) = 1.60$, $p = .11$. The Bayes factor for the alternative versus the null was 1.75, slightly favoring the alternative to the null. The correlation between $K_{testAny}$ and $K_{fixationMax}$ at array size four was .55 (p ¡ .001). At array size six, the average of $K_{testAny}$ was 4.45 (SD 1.15) and

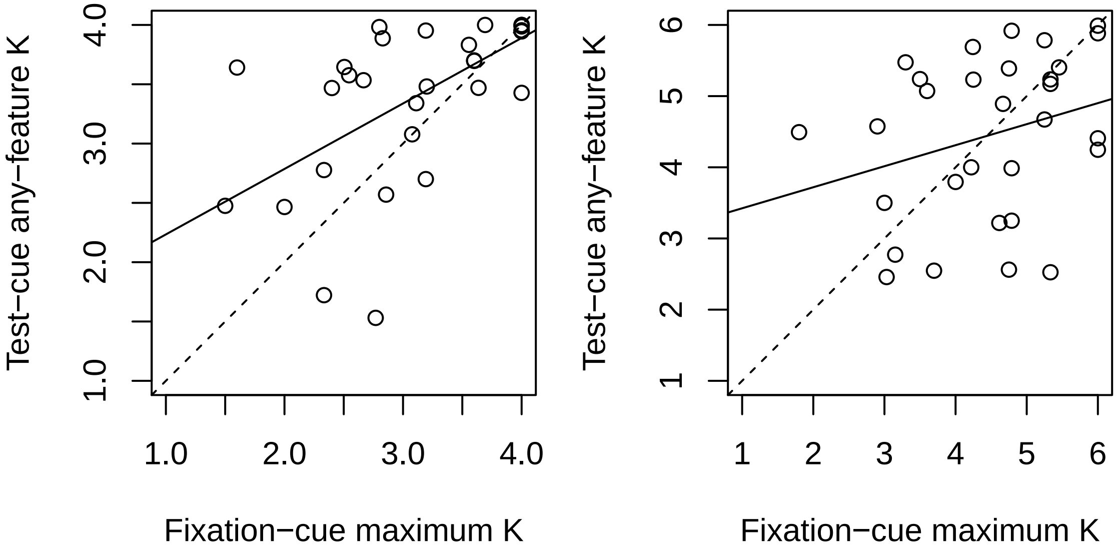

Figure 3.3: Scatterplots of each participant's maximum K-value obtained for any single feature in the fixation-cue condition ($K_{fixationMax}$; X-axis) by the estimated number of objects for which at least one feature was known ($K_{testAny}$; Y-axis). Array size four is shown in the left panel and array size six is shown in the right panel. The solid line is the correlation between $K_{fixationMax}$ and $K_{testAny}$. The dashed line is a line with slope 1 and intercept 0.

the average of $K_{fixationMax}$ was 4.46 (*SD* 1.08). A standard paired *t*-test was unable to distinguish these conditions, $t(58)$ = -0.046, p = .96. A Bayesian *t*-test gave a Bayes factor of 0.14, suggesting that the null is approximately seven times as likely as the alternative. The correlation between $K_{testAny}$ and $K_{fixationMax}$ was .30 but non-significant, $p = .056$. Scatterplots of the data used in these analyses with correlations marked are shown in Figure 3.3. The noise that is apparent in these plots can be at least partially attributed to the low number of trials from which each data point was generated: There were only 20 trials in each array size X cuing condition X cued feature cell for each participant.

Array size two was not analyzed because most participants were assumed to be at ceiling performance on those trials. Similarly, array size six is assumed to be a better data set than array size four due to the possibility that many participants were still at ceiling performance at array size four. To put numbers to it, at array size six, 20 out of the 30 participants had a $K_{fixationMax}$ greater than four, suggesting that most of the participants would have been at ceiling at array size four. Given that $K_{fixationMax}$ was less than $K_{testAny}$ at array size four and that most of the participants were at ceiling in that condition, it is possible that the difference between the K values should be diminished if participants were not at ceiling, which is exactly what we observed at array size six. Accordingly, although the some of the analyses of data from array size four are reported, array size six is the focus from here on.

Averaging $K_{testAny}$ across all participants and using only data from array size six, the number of objects for which no features were known was 1.55, for which one feature was known was 2.26, for which two features were known was 1.45, for which three features were known was 0.58, and for which all four features were known was 0.15. These summary data were drawn from the data shown in Table 3.1. We can use these numbers of features known per object to calculate the number of features known in the sample array (again at array size six) by taking a weighted sum, giving a result of 7.51 features known. By dividing by the number of objects for which at least one feature was known we get the number of features known per object for which at least one feature was known, which is 1.69.

This analysis suggests that regardless of whether participants are attending to a single feature of visual objects (fixation-cue condition) or some combination of features of those objects (test-cue condition), they still load the same number of objects into mind. When only raw accuracy in the two conditions is compared (not

Table 3.1: Number of Objects Held in Mind at Array Size Six for each Combination of Features in Experiment 6

#[a]	Color[b]	Gap[b]	Length[b]	Orientation[b]	K[c]
0	-	-	-	-	1.55
1	X	-	-	-	0.83
1	-	X	-	-	0.46
1	-	-	X	-	0.38
1	-	-	-	X	0.59
2	X	X	-	-	0.30
2	X	-	X	-	0.25
2	X	-	-	X	0.35
2	-	X	X	-	0.14
2	-	X	-	X	0.27
2	-	-	X	X	0.13
3	X	X	X	-	0.12
3	X	X	-	X	0.22
3	X	-	X	X	0.13
3	-	X	X	X	0.12
4	X	X	X	X	0.15

[a] *The number of known features.*

[b] *For each feature, an X indicates that the feature was known and a dash indicates that the eature was not known*

[c] *The number of objects for which the indicated combination of features was known.*

using the results of this analysis), participants seem to be holding less information specific to the cued feature in mind in the test-cue condition. However, it could be that in the test-cue condition participants are holding more information in mind overall, but less information about the cued feature due to there being no reason to preferentially attend to that feature. Provided that our assumptions are correct, this analysis provides evidence for a model of visual WM that has a limit to the number of objects which can be held in mind. This in no way conflicts with the results of our other experiments, which we believe show that objects are not the only limiting factor in WM performance, not that they are not a limiting factor.

According to this analysis, it is a rare event for all of the features of an object to be known, with on average only 0.15 objects out of 6 having all four features of their features known. This conflicts with the original slot model proposed by Luck and Vogel (1997) which posits that all of the features of a stored object are known. On the other hand, it is still compatible with a model that has an object limit, as long as the objects are allowed to be stored without all features intact. Such a model could be the slots plus resources model (Zhang & Luck, 2008).

It is possible that some of the assumptions of our analysis are wrong. Although there is some evidence that the features of visual objects are independent (Vul & Rich, 2010), this is only a recent finding that has not withstood the test of time. Cowan et al. (2012) were not able to reject the hypothesis of independent encoding of the features of objects, which indicates that the finding of Vul and Rich (2010) might be generally applicable to WM research. However, Fougnie and Alvarez (2011) were unable to support the idea that features of objects are completely independent. They found that for color and orientation, memory for the features was largely independent, but not wholly. Additionally, Fougnie and Alvarez found that knowledge of the width and height of rectangles (ostensibly two separate features) were closer to being inseparable than independent. For the features used in this experiment, is seems likely that length and orientation might not be very separable, but the other combinations of features should be fairly separable. If Fougnie and Alvarez are correct, assuming independence of our features might lead us to a result that approximates the true state of the world, but in order to properly carry out this analysis we would need precise information about the separability of our features.

A different potentially problematic assumption is that $K_{fixationMax}$ may not be an appropriate estimate of the number of slots that participants have available to them. It could be that participants are distracted by the irrelevant features of the objects in the fixation-cue condition and cannot effectively fill their object slots with the relevant feature. In that case, a task in which objects possess only a single feature could be used to estimate the number of slots that participants have available. This would create another problem by creating clear differences between the stimuli used to estimate object slots and the stimuli used to estimate the number of objects for which each feature was known.

Chapter 4

Extension to fewer feature dimensions with more states

4.1 Experiment 7

In this experiment we attempted to replicate the results of another experiment from Luck and Vogel (1997) in which participants were only required to remember at most two features, in this case the color and orientation of long rectangular bars like those used in the four-feature experiments. In addition to reducing the feature dimensions from four to two, in keeping with Luck and Vogel we allowed each of the feature dimensions we used to take on any of four values on any given trial. We have shown that requiring participants to remember all of the features of a certain type of visual object composed of four features results in reduced accuracy when compared to some of the individual features which make up the object. How well our finding extends to stimulus objects with fewer features or features drawn from different feature dimensions is, however, unclear. It is possible that having four features per object overloads our ability to organize each object in such a way that it can be reliably remembered. If participants are only held responsible for two features in the multi-feature condition, the overload may not occur and accuracy might only be related to array size and not feature load. Alternately, it could be that the specific feature dimensions we used for our stimuli are responsible for our results, which this experiment partially addresses by removing two feature dimensions. Finally, allowing each feature dimension to vary between only two values (for example, objects could only be red or green) may have allowed participants to do a very good job of chunking objects in single-feature blocks, an ability that might have been reduced in effectiveness in the multi-feature condition because attention was more likely to be spread between a number of feature dimensions in that condition.

This experiment will also provide evidence related to the question of whether feature load can predict accuracy without reference to object load. Wheeler and Treisman (2002) found that participants could just as easily remember six colors when those colors were in six separate objects as when the colors were displayed with two colors per object in three separate objects. In this experiment we can compare conditions which are matched in terms of feature count but differ in object count.

4.1.1 Method

Data from 21 participants (8 female, mean age 19.9 years) were used. One additional participant was removed for falling below 55% accuracy on two trial blocks. Three of the participants in this experiment were recruited from the community and were paid $15 for their participation.

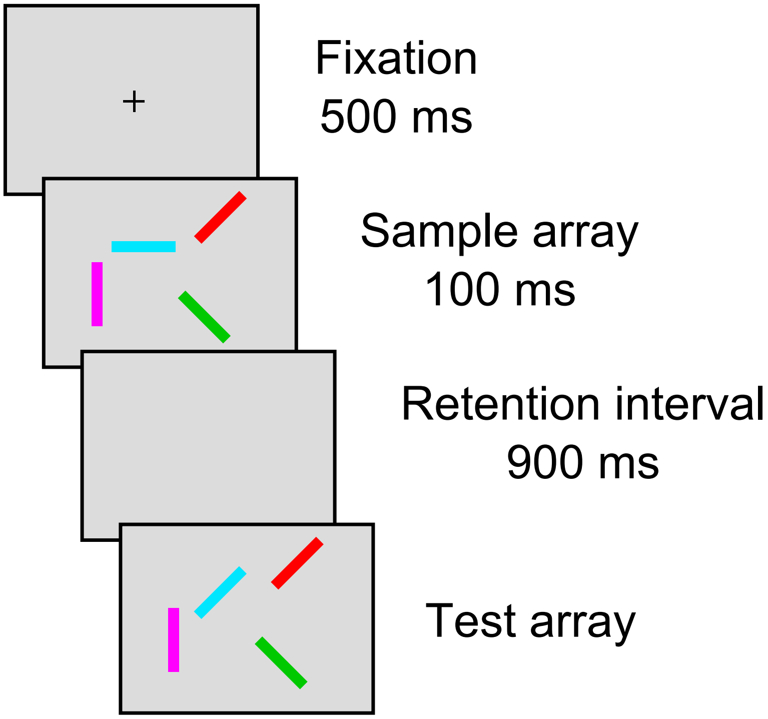

Figure 4.1: Diagram of the task used in Experiment 7. This figure shows how the stimuli for this experiment can take on a greater range of feature values than stimuli in the earlier experiments while also reducing the feature load in terms of the number of feature dimensions per object.

The materials used in this experiment differed from those of other experiments in that each object could possess any of four colors and orientations. The orientations were 0, 45, 90, and 135 degrees. The rectangles were the same dimensions as the long rectangles used in the other experiments and did not have a gap. The colors used in this experiment were red (255,0,0), green (0,246,0), cyan (0, 254, 255), and magenta (255, 0, 255) for the objects and background (60, 60, 60). These were not the same colors used by Luck and Vogel (1997), who used red, green, blue, and black. Our stimuli were still easily discriminable, so we find it unlikely that the colors would have an effect of the results. The increased center-to-center spacing of objects of

2.25° introduced in Experiment 6 was used in this experiment as well. There was no secondary verbal load task in this experiment. See Figure 4.1 for a diagram of the method for this experiment. In this experiment, there were two-single feature trial blocks and one multi-feature trial block, each with 180 trials.

4.1.2 Results

There was a main effect of array size on accuracy, $F(2, 40) = 81.85$, $MSE = 0.0044$, $p < .001$, $\eta_p^2 = .80$. There was also a main effect of attended feature, $F(2, 40) = 3.876$, $MSE = 0.0089$, $p < .05$, $\eta_p^2 = .16$. In this experiment, only a marginally significant interaction was found, $F(4, 80) = 2.284$, $MSE = 0.0022$, $p = .0674$, $\eta_p^2 = .10$. The data for this experiment are summarized in Figure 2.2.

A Newman-Keuls post hoc test revealed that the color condition was different from both the orientation and multi-feature conditions. However, the orientation and multi-feature conditions were not different from one another.

For this experiment, accuracy in the multi-feature condition at array size two (M = 0.93, SD = 0.08) and the average accuracy of the single-feature conditions at array size four (M = 0.89, SD = 0.05) were not found to differ, $t(40) = 2.02$, $p = 0.0504$ (two-tailed). The Bayes factor for this comparison was 0.07, providing substantial evidence for the null hypothesis of no difference.

4.1.3 Discussion

The results of Experiment 7 are similar to what we have found before, namely that there is a cost to remembering multiple features of visual objects. The use of feature sets that included more than two possible feature values helps to reduce the likelihood that our previous results were solely caused by an artifact of using feature sets that possessed only two of the many possible values. As a result of this change, in this experiment participants were less able to chunk objects together by common features. However, they were also able to rely more on memory for which features values were presented than in previous experiments due to the reduced repetition of feature values resulting from a larger sample of possible values. There are a wide variety of possible stimulus sets we could have used in this experiment, so although some questions were answered by this experiment, they have only been answered for a narrow range of possible features.

We again found that accuracy in the multi-feature condition was equivalent to that in the most difficult single-feature condition. This is notable because in this experiment orientation was the most difficult single feature, whereas length was the most difficult single feature in previous experiments. It was previously hypothesized that there was something about the length feature of the objects that was limiting accuracy on the multi-feature condition. However, we see here that accuracy on the multi-feature condition is generally limited by the most difficult single feature in the set of salient features, not by a specific feature dimension.

This is the first experiment in this study that has conditions that are exactly matched on feature load but differ in object load. We were able to directly compare the multi-feature condition at array size two to the single-feature conditions at array size four. Unlike our four-feature experiments, we did not find any difference in accuracy when we carried out this analysis. This result supports the similar finding of Wheeler and Treisman (2002) for the stimuli used in this experiment.

4.2 Experiment 8

In Experiment 7 we once again found that the multi-feature condition was as difficult as the most difficult single-feature condition. In this experiment we attempted to examine this effect in a case where the difficulty of the features used was equated (thanks to Klaus Oberaurer for suggesting this experiment). In Experiment 7, accuracy for color was better than accuracy for orientation, so in this experiment we changed the colors to be less discriminable in order to decrease accuracy for color.

4.2.1 Method

In this experiment, data from 33 participants (25 female; mean age 18.5 years) were used. One additional participant was removed from the sample due to falling below the accuracy cutoff.

The colors of the stimuli used in this experiment were dark pink (235,76,90), pale violet red (210,94,140), medium purple (165,108,214), and light slate blue (120,116,253). The method was otherwise identical to Experiment 7.

4.2.2 Results

There was a main effect of array size on accuracy, $F(2, 64) = 246.77$, $MSE = 0.0036$, $p < .001$, $\eta_p^2 = .89$. There was also a main effect of attended feature, $F(2, 64) = 7.54$, $MSE = 0.0082$, $p < .01$, $\eta_p^2 = .19$. In this experiment, no interaction was found, $F(4, 128) = 0.31$, $MSE = 0.0023$, $p = .87$. The data for this experiment are summarized in Figure 2.2.

A Newman-Keuls post-hoc test showed that accuracy in the multi-feature condition was lower than in either the color or the orientation conditions. It also showed that the color and orientation conditions were not different from one another.

For this experiment, accuracy in the multi-feature condition at array size two ($M = 0.89$, $SD = 0.06$) and the average accuracy of the single-feature conditions at array size four ($M = 0.84$, $SD = 0.05$) were found to differ, $t(64) = 2.96$, p ¡ 0.005 (two-tailed). The Bayes factor for this comparison was 3.46.

4.2.3 Discussion

With this experiment we were able for the first time to separate the multi-feature condition from all single-feature conditions. This shows that our previous finding that the multi-feature condition is as difficult as the most difficult single feature does not hold in all cases.

Unlike Experiment 7 but in keeping with earlier experiments, we found differences between conditions equated in feature load but differing in object load. An explanation for this difference from Experiment 7 is that with the increased difficulty of the colors used, participants were unable to store as many objects in WM in the color single-feature condition. In the multi-feature condition at array size two, participants may have been able to distinguish the colors more easily due to the small number of objects they were required to encode.

Chapter 5

Summary and concluding remarks

5.1 General Discussion

The initial purpose of this series of experiments was to determine if a commonly-cited result supporting objects as the fundamental unit of storage in visual WM (Luck & Vogel, 1997) could be replicated under a variety of conditions – including a direct replication. Our results so far have consistently failed to replicate the results of Luck and Vogel in experiments using two or four features per object. These failures to replicate consist of seven experiments with a total of 165 participants, with a very clear and strongly significant pattern of results that is strongly opposed to the pattern of results observed by Luck and Vogel. Given how well our results have replicated across a number of experimental manipulations and different participant samples, we believe our results are reliable and should be taken as a strong refutation of the results of Luck and Vogel. When our failures to replicate the results of Luck and Vogel are combined with earlier failed replications (Delvenne & Bruyer, 2004; Wheeler & Treisman, 2002), it becomes clear that every feature conjunction experiment of Luck and Vogel has failed to replicate. It should be noted that our results and the results of the cited failures to replicate all show the same pattern of results: an effect of feature load. It should also be noted that Luck and Vogel observed their pattern of results (no effect of feature load) with a sample size of only 10 participants who performed no more trials than our participants, indicating that they may have had too little power to detect an effect present in their data.

Although we have shown that the results of Luck and Vogel do not replicate, simply failing to replicate a result does not necessarily invalidate a theory supported by that result. One must seek to eliminate confounds in order to allow for unambiguous interpretation of results with respect to a theory. In particular, the theoretical issue we are interested in is storage in visual WM, so we should attempt to remove confounds related to the encoding and retrieval of stimuli in order to verify that our effects are due to the behavior of visual WM storage.

Experiment 3 controlled for the possibility that participants' ability to perform the task was limited by an encoding time bottleneck. Experiment 5 controlled for the number of objects about which a decision must be made at test, suggesting that our pattern of results observed in experiments with a full-array probe are not due to accumulated decision error across multiple objects. However, it is possible that accuracy in the multi-feature condition of Experiment 5 was limited by the fact that participants were required to make four decisions, one for each feature of the object. Experiment 6 extended the results of Experiment 5 by further controlling the number of decisions that participants were required to make. This showed that when participants are only required to make a single decision at test, there is an advantage if participants are able to selectively encode information about the feature of interest, just as they were able to do in the single-feature conditions of previous experiments. This confirms that the results of Experiments 1 - 5 were not entirely due to decision error. We also verified that precise stimulus characteristics do not significantly affect our results by using different stimulus and background colors in Experiments 4 and 5 than we used in Experiments 1 to 3. In Experiments 7 and 8, we further verified that we were able to find an effect of feature load in experiments in which only two features were allowed to vary, but were able to vary between a greater number of feature values per feature dimension than were used in the four-feature experiments. When combined, these experiments clearly support the argument that feature load has an effect on WM performance.

We do not argue that object load is unimportant, simply that it is not the sole factor contributing to WM performance. In all of our experiments we tested conditions which were equated or nearly equated in feature load but which differed in object load. In seven out of eight experiments, we found an effect of object load. The reason why we did not find the same effect in one of the experiments (Experiment 7) is unclear and may warrant further study. Nevertheless, it seems clear from our data that we cannot ignore the importance of object load in our experiments.

A result that might appear to conflict with ours is that of Awh, Barton, and Vogel (2007). They found that the reduction in accuracy that appears when multiple complex stimuli are remembered should be attributed to difficulty comparing insufficiently-precise representations of sampled stimuli to test objects, not to a reduction in the number of stored representations. Although our analysis of the total number of objects in mind in Experiment 6 suggests that there may be reduction in the number of stored objects, that results of that analysis are constrained by the modeling assumptions that were used. The high-threshold model that we used (Cowan, 2001) assumed that objects or features must be known well enough in order to pass some high threshold that allows participants to make a high-confidence response on a trial. This model does not make a provision for imprecise representations and would underestimate the number of objects in mind if some of those representations were fairly imprecise. Most importantly, we make the claim that storage of the features of a visual item is not without cost, which is not really different from the Awh et al. claim that when a number of complex objects are stored, the quality of the representations suffers. It could be the case that our participants were able to hold the same number of objects in mind in both single- and multi-feature conditions, but have their performance suffer in the multi-feature condition due to an inability to store all of the features of the objects they had in mind, which is suggested by the analysis of total number of objects held in mind in Experiment 6.

One pattern in our results for which we have no explanation is the fact that accuracy in the multi-feature conditions is equivalent to the most difficult single-feature condition in most of our experiments. We found this pattern in Experiments 1 through 5, where the most difficult single feature was length and also in Experiment 7, where the most difficult single feature was orientation. Because the effect occurs regardless of which feature dimension is the most difficult, it is unlikely to be caused by a limitation in visual processing of certain attributes of objects, but rather a limitation of WM. We were able to create a situation in which the multi-feature condition was more difficult than either of the single-feature conditions in Experiments 8 by equating the difficulty of the single-feature conditions. If the most difficult single feature was the limitation to multi-feature accuracy, then accuracy in the multi-feature condition would be the same as accuracy in all of the matched-difficulty single-feature conditions in Experiment 8. This is not what we found in that experiment, showing that accuracy in the multi-feature condition is not strictly limited by the difficulty of the most difficult single feature.

5.1.1 Analysis of high performers

One difference between our data and the data of Luck and Vogel (1997) was that our participants did not seem to perform as well as the participants of Luck and Vogel. Averaging across feature conditions at array size six, their participants achieved approximately 82% accuracy, whereas our participants in Experiment 1 only achieved 73% accuracy. It could be that many of the participants used by Luck and Vogel were at ceiling performance and that the pattern of results we observed only occurs when most participants are not at ceiling performance. To investigate this possibility, we chose to examine high performers from our three most similar experiments: Experiments 1, 2, and 4. These experiments were the same expect for the use of a secondary verbal load task in Experiment 1 and a different set of stimulus colors in Experiment 4. We collapsed across experiments in order to get a large enough sample of high performers.

When we took the top quartile of our participants (N = 12 from a set of 48 participants), average accuracy at array size six was 79%. As can be seen in Figure 5.1 (left panel), the basic pattern of results that we observed in our general population was still present in the data of our high performers. An effect of attended feature

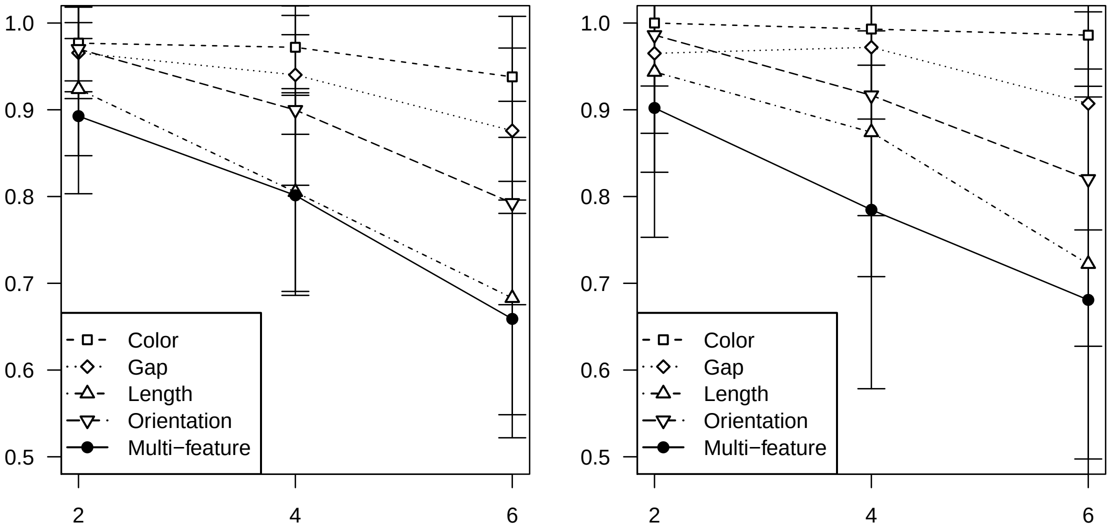

Figure 5.1: Plots of accuracy for the top performing participants in Experiments 1, 2, and 4. Data from the top quartile of participants is presented in the left panel. Data from the top four participants, whose average accuracy is matched to the accuracy of the participants of Luck and Vogel (1997), is presented in the right panel. Error bars are standard error.

on accuracy was found in the high performers, $F(4,\ 44) = 38.37$, $MSE = 0.0055$, $p < .001$, $\eta_p^2 = .78$. Given the effect size, this effect appears to be even stronger in the high performers than in the general population, although this interpretation must be tempered by the fact that these data are from three different experiments with somewhat different methods. A Newman-Keuls post hoc found that only the length and multi-feature conditions did not differ from one another, much like in the general population. In order to get a set of participants who had an average accuracy of 82% at array size six, we selected only the top four participants from Experiments 1, 2, and 4. Although this sample is too small to meaningfully analyze with ANOVA, the data are plotted in Figure 5.1 for reference. Although this sample of participants is very small, the same basic pattern of results that we have observed elsewhere is beginning to appear. It appears that although the participants of Luck and Vogel (1997) performed significantly better than our participants, this difference is not the reason that we were unable to replicate their results.

5.1.2 Potential chunking confound

It seems likely that participants were able to engage in a significant amount of chunking in our experiments due to the type of objects we used. In our four-feature experiments, features could take on only one of two values, meaning that feature value repetitions in multiple adjacent objects could lead participants to encode those features as a chunk that may exist separately from the underlying objects. That is to say, that a region of the array may take on a value for a specific feature and any objects in the region are considered to have the same value. Since there were only two values available for each feature, it was fairly common that there were easily chunked sections of the array. Perhaps the most clear evidence for this chunking is in Experiment 3 where the color and gap features seemed to suffer very little performance decrement as array size increased. In fact, in a number of the four-feature experiments accuracy for color only decreased slightly as array size increased. The major difference between Experiment 3 and the other experiments was that participants were given 500 ms to encode the sample array in Experiment 3 versus 100 ms in the other experiments. It is possible that participants were better able to chunk in Experiment 3 than in the other experiments because they had more time to actively form chunks while viewing the sample array.

If chunking is an active process that requires attention and participants were relying heavily on chunking, our finding that the multi-feature condition was more difficult than most single-feature conditions could be an artifact of the stimuli we used. To explain: In the multi-feature condition, in order to chunk well enough to "keep up" with the single-feature conditions, participants would have had to form chunks which contained information about all four features. They might have had to create separate chunks for each feature dimension, which except in unusual circumstances would be located in different spatial regions. This task is clearly much more difficult than creating chunks for only one feature dimension at once. Alternately, participants might have been able to create a chunk from objects which shared the same value for most or all feature dimensions, allowing them to effectively store information about all feature dimensions in a single chunk. That nearby objects would share most or all feature values is, however, quite unlikely compared to the probability that nearby objects would share a feature value for just a single feature dimension, making chunking objects that shared multiple features a strategy that would only be effective on a small number of trials. It seems plausible that if participants were using an attention-demanding chunking strategy, they would be limited in their ability to form chunks in the multi-feature condition relative to the single-feature conditions due to the added attentional demands of attempting to chunk in multiple feature dimensions at once. Given this, the differences we found between the single- and multi-feature conditions may be due to difficulty creating chunks in the multi-feature condition, and not directly due to difficulties storing multi-feature objects. This issue could be addressed by an experiment in which feature values are not allowed to repeat in a given array, preventing identical features from being placed in chunks.

One piece of evidence that suggests that chunking may not be the only explanation for our results comes from Experiments 7 and 8. In those experiments, each feature dimension was allowed to take on any of four values, which should reduce the ability of participants to chunk objects together relative to the four feature experiments. In spite of reduced chunking potential, we still found an effect of feature load, suggesting that chunking may not be entirely driving the effect. That said, a specific manipulation to control for chunking would still be required to rule out that explanation for our results.

5.1.3 Final conclusions

Although more experiments are required to account for some potential confounds, our results so far are still evidence that storage in visual WM is limited in part by the number of features of stored objects, not just the number of objects. We have consistently observed a pattern of results that does not support the belief that the number of objects which can be stored in visual WM is the sole determinant of accuracy. Neither have we found evidence that feature load is the sole determinant of performance in our data. We believe that claiming either that objects or that features are the single facet of WM that mediates visual WM performance in all cases is unreasonable in light of these data. We suggest that rather than attempting to specify the most important individual factor which mediates visual WM performance, we would be better served by improving our understanding of the contributions of all of the factors which meaningfully impact visual WM performance.

References

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

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

Baddeley, A. (2003). Working memory: looking back and looking forward. *Nature reviews. Neuroscience*, *4*(10), 829–39. doi: 10.1038/nrn1201

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*(1), 87-185.

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

Delvenne, J.-F., & Bruyer, R. (2004). Discrete fixed-resolution representations in visual working memory. *Visual Cognition*, *11*(1), 1–27. doi: 10.1080/13506280344000167

Fougnie, D., & Alvarez, G. A. (2011). Object features fail independently in visual working memory: Evidence for a probabilistic feature-store model. *Journal of Vision*, *11*(12), 1–12. doi: 10.1167/11.12.3

Fougnie, D., Asplund, C. L., & Marois, R. (2010). What are the units of storage in visual working memory? *Journal of Vision*, *10*(12), 1-11.

Hollands, J. G., & Jarmasz, J. (2010). Revisiting confidence intervals for repeated measures designs. *Psychonomic Bulletin and Review*, *17*(1), 135-138.

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

Morey, C. C., & Cowan, N. (2004). When visual and verbal memories compete: evidence of cross-domain limits in working memory. *Psychonomic bulletin & review*, *11*(2), 296–301.

Pashler, H., & Wagenmakers, E.-J. (2012). Editors introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, *7*(6), 528–530. doi: 10.1177/1745691612465253

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin and Review*, *16*(2), 225–237. doi: 10.3758/PBR.16.2.225

Treisman, A. (1988). Features and objects: The fourteenth bartlett memorial lecture. *The Quarterly Journal of Experimental Psychology Section A*, *40*(2), 201–237. doi: 10.1080/02724988843000104

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), 92-114.

Vogel, E. K., Woodman, G. F., & Luck, S. J. (2006). The time course of consolidation in working memory. *Journal of Experimental Psychology: Human Perception and Performance*, *32*(6), 1436-1451.

Vul, E., & Rich, A. N. (2010). Independent sampling of features enables conscious perception of bound objects. *Psychological Science*, *21*(8), 1168–1175. doi: 10.1177/0956797610377341

Wheeler, M. E., & Treisman, A. M. (2002). Binding in short-term visual memory. *Journal of Experimental Psychology: General*, *131*(1), 48-64.

Zhang, W., & Luck, S. J. (2008). Discrete fixed-resolution representations in visual working memory. *Nature*, *453*(8), 233–235. doi: 10.1038/nature06860