Learning to Form Visual Chunks:
On the Structure of Visuo-Spatial Working Memory

James S. Magnuson (magnuson@bcs.rochester.edu)
David G Bensinger (dgb@cvs.rochester.edu)
Mary Hayhoe (mary@cvs.rochester.edu)
Dana Ballard* (dana@cs.rochester.edu)
Department of Brain and Cognitive Sciences, Center for Visual Science
University of Rochester, Meliora Hall, Rochester, NY 14627 USA
**Also with the Department of Computer Science, University of Rochester*

Abstract

We are interested in a functional account of how capacity constrains memory use in natural, ongoing behaviors, and in how visual memory demands can be reduced through the use of what we have called perceptual *pointers*, or *deictic codes*. Here, we ask whether, with experience, participants can restructure task representations such that single fixations can point to more and more complex chunks of information. We tracked eye movements as participants copied simple model patterns which were presented with different frequencies. At first, participants made multiple fixations to individual pattern components. As patterns were presented repeatedly, model inspections were reduced substantially. This suggests that participants formed more compact representations of the patterns with experience, allowing single fixations to point to larger chunks of information. We also propose that deictic codes provide a short-term store analogous to the visuo-spatial scratchpad or articulatory loop. When the task was structured such that a separate visual search was required for each model component, much less learning was observed than when fixations to known locations were required, suggesting deictic codes were disrupted by active visual search.

Introduction

Models of visuo-spatial working memory have typically been concerned with the limits of human working memory. Results from studies pushing working memory to its limits have led to the proposal of modality-specific "slave" systems which provide short-term stores. Usually, it is assumed that there are at least two such stores: the articulatory loop, which supports verbal working memory, and the visuo-spatial scratchpad (Baddeley & Hitch, 1974) or "inner scribe" (Logie, 1995), which supports visual working memory. Here, we are interested in complementing such work with studies of how capacity limitations constrain performance in natural, ongoing tasks carried out without added time or memory pressures.

Eye Movements in Natural, Ongoing Tasks

The prototypical task we use is block-copying (see Figure 1). Participants are presented with a visual display (on a computer monitor or on a real board) which is divided into three areas. The *model* area contains a pattern of blocks. The participant's task is to use blocks in the *resource* area to construct a copy of the model pattern in the *workspace*. We continuously measure eye and hand position as the participant performs the task.

Note that the task differs from typical laboratory tasks in several ways. First, it is closer to natural, everyday tasks than, e.g., tests of iconic memory or recognition tasks. Second, as a natural task, it extends over a time scale of several seconds. Third, the eye and hand position measures allow us to examine performance without interrupting the ongoing task; that is, the time scale and dependent measures allow us to examine instantaneous performance at any point, but we also have a continuous measure of performance throughout an entire, uninterrupted natural task. Studies using variants of the block-copying task have revealed that information about gaze and hand locations can be used as pointers to reduce the amount of information that must be internally represented (e.g., Ballard, Hayhoe, & Pelz, 1995). These pointers index locations of task-relevant information, and are called *deictic codes* (Ballard, Hayhoe, Pook, & Rao, 1997).

Deictic Codes

In several variants of the block-copying task, the same key result has been replicated. Rather than committing even a small portion of a model pattern to memory, participants work with one component at a time, and typically fixate each model component twice. First, participants fixate a model component and then scan the resource area for the appropriate component and fixate it. The hand moves to pick up the component. Then, a second fixation is made to the same model component as on the previous model

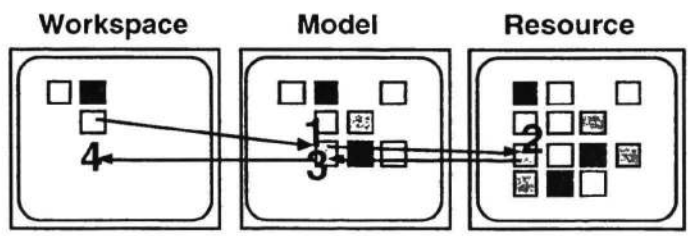

Figure 1: The block copying task. The task is to use blocks displayed in the *resource* (right monitor) to build a copy of the *model* (center) in the *workspace* (left). The arrows and numbers indicate a typical fixation pattern during block copying. The participant fixates the current model block twice. At fixation 2, the participant picks up the dark gray block. At fixation 4, the participant drops the block.

fixation. Finally, participants fixate the appropriate location in the workspace and move the component from the resource area to place it in the workspace. If we divide the data into fixation-action sequences each time an object is dropped in the workspace, this *model-pickup-model-drop* sequence is the most often observed (~45%, with the next most frequent pattern being *pickup-model-drop*, which accounts for ~25% of the sequences; *model-pickup-drop* and *pickup-drop* each account for ~10% of the sequences, with most of the remaining, infrequent patterns involving multiple model fixations between drops; thus, the majority of fixation sequences involve at least one model fixation per component, with an average of nearly two model fixations per component).

Given such a simple task, why don't participants encode and work on even two or three components between model fixations, which would be well within the range of capacity? Ballard et al. (1997) have proposed that memories for motor signals and eye or hand locations provide a more efficient mechanism than could be afforded by a purely visual, unitary, imagistic representation. In the block copying paradigm, participants seem to encode simple properties one at a time, rather than encoding complex representations of entire components. For example, a fixation to a model component could be used to encode the block's color, and its location within the pattern. This might require encoding not just the block's color, but also the colors of its neighbors (which would indicate its relative location). Alternatively, the block's color and the signal indicating the fixation coordinates could be encoded. With the color information, a fixation can be made to the resource area to locate a block for the copy. The fixation coordinates could serve as a pointer to the block's location in the model. Now, a saccade can be made back to the fixation coordinates, and the information necessary for placing the picked-up block in the workspace can be encoded.

Note that in the copying task, the second fixation is typically made back to exactly the same place in the model. Why can't the information that allows the participant to fixate the same location be used to place the picked-up block in the correct place in the workspace? Because that information is about an eye position -- the pointer -- not about the relative location of the block in the pattern. The fixation coordinates act as a pointer in the sense of the computer programming term: a small information unit which represents a larger information unit simply by encoding its location. Thus, very little information need be encoded internally at a given moment. Perceptual pointers allow us to reference the external world and use it as memory, in a just-in-time fashion. This hypothesis was inspired in part by an approach in computer vision which greatly reduced the complexity of representations needed to interact with the world. On the *active* or *animate* vision view (Bascy, 1985; Brooks, 1986; Ballard, 1991), much less complex representations of the world are needed when sensors are deployed (e.g., camera saccades are made) in order to sample the world frequently, in accord with task demands.

Hayhoe et al. (1998) reported compelling evidence for the pointer hypothesis in human visuo-motor tasks. As participants performed the block-copying task at a computer display, the color of an unworked model block was sometimes changed during saccades to the model area (when the participant would be functionally blind for the approximately 50 ms it takes to make a saccadic eye movement). The color changes occurred either after a drop in the workspace (*before pickup*), or after a pickup in the resource area (*after pickup*). Participants were unaware of the majority of color changes, according to their verbal reports. However, fixation durations revealed that performance was affected. Fixation durations were slightly, but not reliably, longer (+43 ms) when a color change occurred *before pickup* compared to a control when no color change occurred. When the color change occurred *after pickup*, fixation durations were reliably longer (+103 ms) than when no change occurred.

How do these results support the pointer hypothesis? Recall that the most frequent fixation pattern was *model-pickup-model-drop*. When the change occurs *after pickup* – just after the participant has picked up a component from the resource area and is about to fixate the corresponding model block again -- there is a relatively large effect on performance. When the color change occurs *before pickup* – just after a participant has finished adding a component to the workspace -- there is a relatively small effect. At this stage, according to the pointer hypothesis, color information is no longer relevant; what had been encoded for the preceding pickup and drop can be discarded, and this is reflected in the small increase in fixation duration.

Bensinger (1997) explored various alternatives to this explanation. He found that the same basic results hold when: (a) participants can pick up as many components as they like (in which case they still make two fixations per component, but with sequences like *model-pickup*, *model-pickup*, *model-drop*, *model-drop*), (b) images of complex natural objects are used rather than simple blocks, or (c) the model area is only visible when the hand is in the resource area (in which case the number of components worked on drops when participants can pick up as many components as they want, so as to minimize the number of workspace locations to be recalled when the model is not visible).

Deictic Codes and Chunking

The concept of chunking (or recoding information into small units such that the number of informational units that must be held in short-term memory is reduced -- e.g., recoding the twelve digits "200117761492" as the three chunks "2001, 1776, 1492") has been well-known since Miller's seminal studies (Miller, 1956). However, it is not necessarily clear how to quantify the notion when we study natural behaviors involving multidimensional stimuli. Deictic codes provide a potentially informative framework for studying this issue. As described above, in the copying tasks we use, participants tend to employ highly stereotyped, serialized eye movements when first presented with a model to copy. Using multiple fixations allows participants to encode different aspects of a stimulus in a just-in-time fashion. If we give participants the opportunity to become familiar with the model, it is possible that the eye movement patterns will change. Specifically, if participants are able to recode the features of the model pattern such that a single

fixation can stand for more features, we should observe reductions in model fixations. Indeed, Magnuson, Sagerer, Hayhoe & Ballard (1997) found such a pattern when they presented participants with the same highly complex model object (e.g., of a scooter) constructed from wooden modeling toys on several consecutive copying trials.

In the current experiment, we extended the work on deictic codes in several ways. First, we used several different simple model patterns ("connected" blocks) within participants. Second, we varied the frequency with which the patterns were presented. This allowed us to study the amount of experience necessary to yield reductions in the number of model fixations.

Third, we varied the internal consistency of some of the model patterns. Two of the models were identical in two of their three components. This could have several possible outcomes. The patterns might be treated independently, which would be reflected in similar amounts of reduction as found for frequency-matched, 100% consistent models; the two shared components might be learned much more quickly than the unshared component, resulting in a larger reduction than for frequency-matched items; or, the variability might disrupt learning such that the amount of reduction observed would actually be less than that for frequency-matched items.

A fourth contribution of the current study was an examination of task consistency. For one group of participants (the *consistent resource* group), the resource area containing the components necessary for the task was presented in the same arrangement on each trial. For the other group (*variable resource*), the resource area was randomly rearranged for each trial. This allowed us to examine the importance of consistency in another part of the task besides the model, in order to ask whether only the model pattern information is being learned.

Experiment

Method

Participants. Eleven students at the University of Rochester were paid for their participation. All had normal or corrected-to-normal vision.

Materials The stimuli were simple three-component patterns consisting of two colored blocks connected by a thin colored rectangle (see Figure 2). There were five different model patterns. One was presented with relatively high frequency (50% of trials), one with medium frequency (20% of trials), and one with low frequency (10%). This frequency manipulation allows us to examine the effects of statistical regularities at the overall pattern level. The last two patterns were both presented with low frequency (10%) but were identical except in the colored block on the right of the pattern (see Figure 2). These patterns allowed us to examine the effects of regularity within patterns. For the high, medium, and low patterns, any one block or connector predicts the entire pattern. For the patterns which share components, the combination of the left block and the connector predict the right block, but with only a 50% transitional probability (we will refer to the latter patterns as *transitional*, or *trans.*, *A* and *B*). In all, 100 trials were

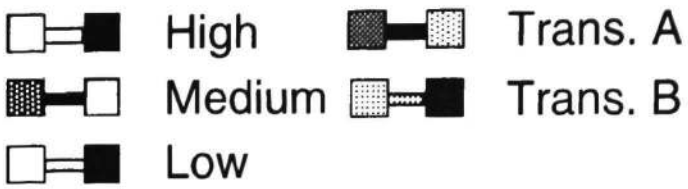

Figure 2: The patterns used in the current experiment. Note that the actual experimental items used solid, easily-distinguishable colors rather than patterned shades of gray.

presented to each participant, with the trials pseudo-randomly ordered such that the overall proportions of patterns occurred every 10 trials.

In addition to the frequency manipulation, the consistency of the resource area was manipulated. For one group of six participants (the *consistent resource* group), the same arrangement of items appeared in the resource area every trial. For the other five participants (the *variable resource* group), a novel, random arrangement of the same objects appeared in the resource area on each trial (see Figure 3).

Procedure Stimuli were presented using a Macintosh PowerPC 8500 and three 14" Apple monitors. Participants were seated at a comfortable distance from the monitors. The right-most monitor was the resource area, and contained an assortment of blocks and connectors. The center monitor was the model area, and the left-most monitor was the workspace. On each trial, one of the patterns shown in Figure 2 was presented. A participant's task was to select items from the resource area (by clicking on them with the computer mouse), move them to the workspace, and construct a copy of the model pattern. Participants could pick up multiple items simply by clicking on them in succession. In the work area, the items could be dropped by clicking again. Multiple items were dropped in a first-in, first-out fashion.

We tracked eye movements with an Applied Scientific Laboratories E4000 eye tracker. Two cameras mounted on a lightweight helmet provide the input to the tracker. An eye camera provides an infrared image of the eye, sampled at 60 Hz. The center of the pupil and the first Purkinje corneal reflection are tracked to determine the orbit of the eye relative to the head. Accuracy is better than 1 degree of arc, with virtually unrestricted head and body movements. A scene camera is aligned with the participant's line of sight. A calibration procedure allows software running on a PC to superimpose crosshairs showing the point of gaze on a HI-8 video tape record of the scene camera. The scene camera samples at a rate of 30 Hz, and each frame is stamped with a time code. The ASL provided the position of the eye with respect to the head. The head position was monitored using an Ascension Technology 6 df Flock of Byrds. This consists of a transmitter that emits an electromagnetic field and a receiver that allows us to read the position of the head in a full 6 degrees of freedom. The Flock of Byrds allows an area of movement within a volume of about a cubic. The signal from the eye and head tracker were integrated with software provided by ASL to give point-of-gaze with respect to the world. Analyses were based on the integrated point-of-gaze record.

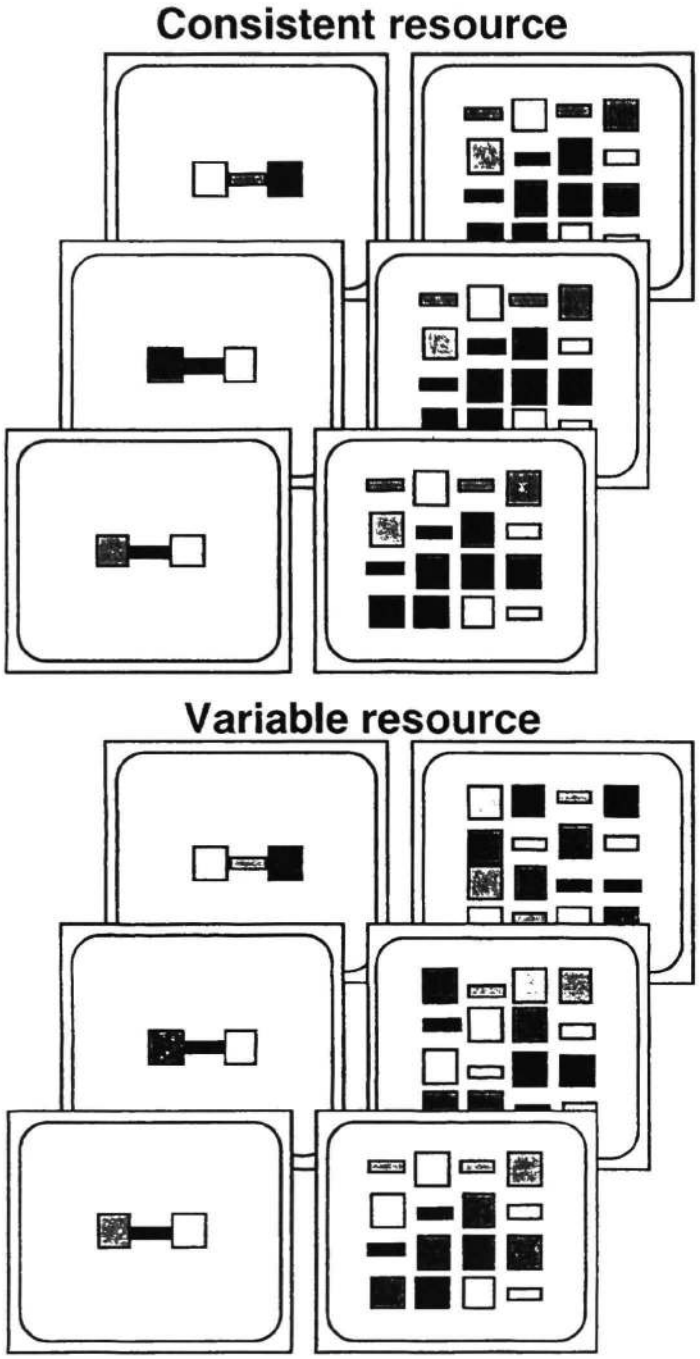

Figure 3: The model and resource areas on three trials in the consistent resource and variable resource conditions. In the consistent resource condition, the resource area contains the same arrangement of objects each trial. In the variable resource condition, the objects appear in a random arrangement each trial.

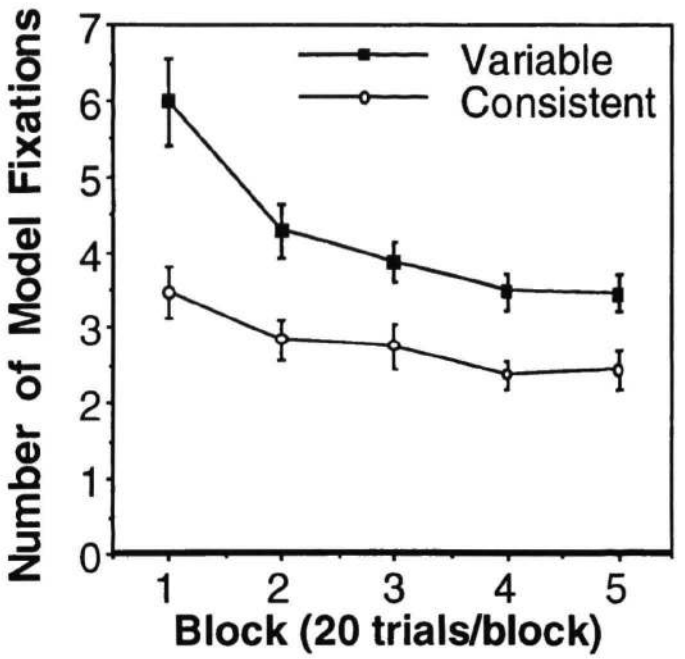

Figure 4: Model fixations by block in the *consistent resource* and the *variable resource* conditions.

Results

The measure we will report here is the number of fixations to the model area. We will assume that a shift from several model fixations to few indicates that more information is being acquired per fixation. This is a strong assumption, but one which appears justified given the series of results discussed in the first section. Those results support the hypothesis that fixations are used to sample just the information needed at the time of the fixation.

The 100 trials were divided into 5 blocks of 20 trials. An analysis of variance showed a significant main effect of *block* ($F[4,36]=8.71$, $p<.001$), which can be seen in Figure 4 (error bars in all of the figures represent standard error of the mean across participants). Model fixations dropped from 4.61 in the first block to 2.92 in the fifth. There was a significant main effect of pattern frequency ($F[3,27]=32.50$, $p<.001$), which can be seen in Figure 5. Across blocks, the average number of model fixations ranged from 2.52 for the high-frequency model to 4.06 for the low-frequency model. The results for the transitional patterns were inconclusive. Although the number of model fixations to the transitional pattern were intermediate between those for the medium and low patterns for the consistent resource group, planned comparisons showed that the differences were not significant.

As can be seen in Figures 4 and 5, there was also a main effect of resource group ($F[1,9]=7.02$, $p=.026$). On average, participants in the *consistent resource* group made fewer model fixations throughout the experiment (2.78) than participants in the *variable resource* group (4.22). As can be seen in the figures, there were no significant interactions; the overall patterns were very similar for both groups.

The results are broken down by frequency and block in Figure 6. With the exception of the low-frequency pattern, there was substantially more reduction in model fixations by the participants in the consistent resource group for each pattern. Note that participants in the variable resource group made substantially more model fixations beginning from the first block of trials. An examination of the first block of 20 trials shows that the groups start at the same level, but diverge rapidly as a result of the consistency of the resource area. This suggests that the consistency of the

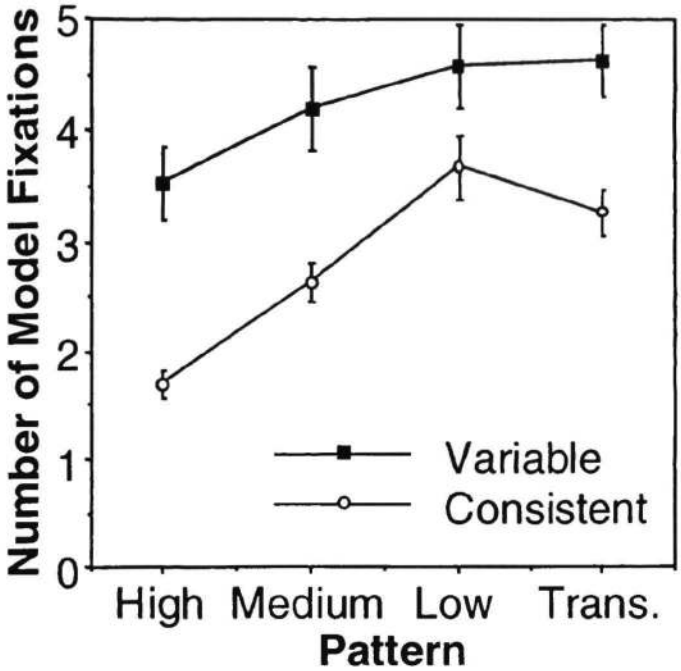

Figure 5: Model fixations by frequency type in the *consistent resource* and the *variable resource* conditions.

resource area has a rapid and powerful effect on participants' ability to reduce the number of model fixations required to copy even the most frequently presented patterns.

Figure 6 shows the number of fixations to each pattern across the five blocks. Participants in both groups began the experiment using an initial procedure in which they fixated each model component two times -- as is usually observed in the copying task with unrepeated patterns. The results discussed in the first section suggest that when participants are making two fixations per component, they are using a highly serialized procedure in which separate fixations encode separate features of model components (e.g., color and relative location).

Participants in the consistent resource area quickly shifted from features to components (i.e., three fixations per model, or one per model component), and even to the entire high-frequency model pattern. As can be seen in Figure 6, participants in the consistent resource group were making just over one model fixation per trial for the high frequency pattern beginning with the second block of trials. For the other patterns, participants were making about three model fixations by the fifth block. This suggests that participants were able to encode chunks of features (when they were making one fixation per component) or even chunks of components (when they were making one fixation per model) with single fixations.

There were also large decreases between the number of model fixations in the first and second blocks for each pattern by the variable resource group. Note that for the high-frequency pattern, this accounts for the majority of the reduction. Participants moved from around six fixations per model to, at best, about three, for the high-frequency pattern. Participants in the variable resource group were able to adopt a *model fixation -- resource pick-up, model-fixation -- resource pick-up, model fixation -- resource pick-up* procedure. Thus, by the end of the experiment, participants in the variable resource group were able to shift from individuating model component *features* (e.g., one fixation for color and one for relative location for each model component) to individuating model *components* or chunks of features (in the case of the high-frequency pattern).

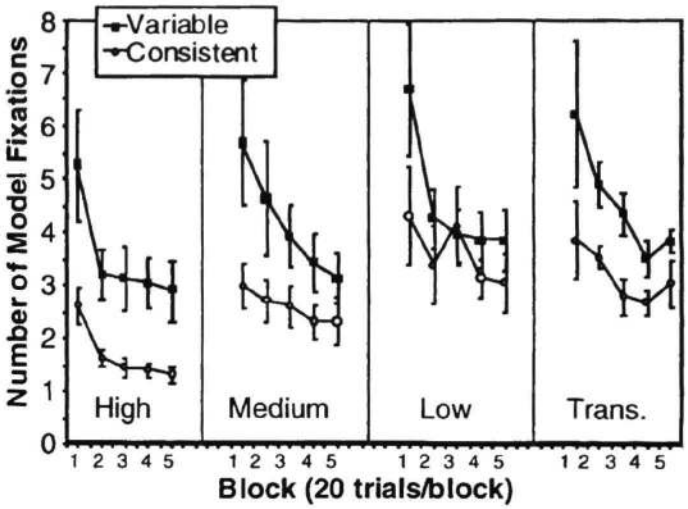

Figure 6: Model fixations per block for each pattern.

The results can be summarized in three points. First, as participants become familiar with a pattern, they are able to reduce the number of fixations required to reproduce it. Further, the frequency with which a pattern was presented modulated the reduction effect, demonstrating that the reduction was not due to task familiarity, but rather to experience with particular patterns.

Second, there was a substantial effect of the consistency of the resource area. More fixations were required in the variable resource condition in general, and the best performance in the variable resource condition (on the high frequency pattern) was about as good as that on the medium frequency pattern in the consistent resource group.

Third, the results for the *transitional* patterns were inconclusive. Contrasts comparing the number of fixations or inspections for *transitional* and *medium*- or *low*-frequency patterns in the fifth block for the consistent resource group did not reveal reliable differences. Thus, while performance on the *transitional* patterns was roughly intermediary between *medium* and *low* patterns, we cannot say what the precise effect of varying pattern-internal consistency was; experiments using more training are called for to examine this issue.

Discussion

These results provide support for the deictic codes hypothesis. Even with simple, three-component models, participants began the experiment using a highly serialized procedure of sampling individual components multiple times, presumably to acquire important features of the components separately. With sufficient practice, participants were able to reduce the number of fixations required to copy frequent models. In the case of the

consistent resource group, the complete high-frequency pattern could be acquired with a single fixation. For other patterns, and for the high-frequency pattern for the variable resource group, participants were still able to greatly reduce the number of fixations needed to perform the copying task. Instead of making two fixations per component, they were able to make about one fixation per component. These reductions indicate that single fixations can come to represent chunks of features. That is, participants were able to restructure their representations of the task such that fewer fixations were required to encode the same information. With repeated exposures, participants were able to build pointers to increasingly complex chunks of information in the visual world. Future work will explore the nature of the reduction more precisely by examining the details of the changes in the fixation patterns.

The results suggest that the memory mechanism at work does involve deictic codes. The result in the case of the consistent resource group is straightforward: given practice and stable task constraints, participants need fewer fixations to perform the task. The key to the connection with the deictic codes hypothesis is the effect of varying the arrangement of objects in the resource area. Why should this have a strong effect on participants' ability to perform the task more efficiently?

We suggest that two different mechanisms are at work in the two resource consistency conditions. It may be that what is learned in the consistent resource condition is not a propositional or imagistic representation of the model, but rather a series of fixation locations, which presumably can be represented more compactly than an imagistic representation. That is, given recognition of the high-frequency pattern in the model area, participants can generate a fixation to location X, Y in the resource area, where they pick up the first component. The coordinates for the next item needed in the resource area (A, B) could be associated with position X, Y, and the location of the third component could be associated with location A, B.

However, given a variable resource area, this mechanism cannot work. Instead, participants must either maintain the strategy of serializing the task by making multiple fixations to the model, or construct a propositional or imagistic representation of the model. The current results suggest that participants prefer serialization. That is, in the variable resource condition, participants were forced to individuate model components with separate visual searches for each one; in the consistent resource condition, participants could potentially work on a chunked representation of the model, by fixating a learned series of resource area locations to obtain needed components.

This suggests that fixation locations (or other representations of task-relevant locations; see Ballard et al., 1998) may provide a store akin to the articulatory loop or the visuo-spatial scratchpad. Or it may be that instead of a series of fixation locations being placed in a rehearsal-based store, a simpler mechanism may be at work in which one state (e.g., a fixation location) is associated with the next. Such a mechanism would provide several advantages over a rehearsal mechanism. The information that needs to be encoded (e.g., fixation locations) is very compact, and can stand for variable amounts of information. A non-imagistic store for guiding visuo-spatial vision would allow relatively intensive visual processing without disrupting the store – compared, say, to an imagistic store like the visuo-spatial scratchpad -- as long as that processing did not require many active fixations. In the variable resource condition, such states cannot be learned and the participant must perform active, appearance-based visual search for each component and so relies on serializing the task via deictic codes for model locations.

While many questions remain open, the current results provide first steps towards understanding the role of fixations in learning and visuo-spatial working memory.

Acknowledgments

Supported by NIH grants EY05729, EY01319, RR06853, and an NSF Graduate Research Fellowship to JSM. We thank Inge-Marie Eigsti for comments which substantially improved this paper.

References

Bajcsy, R. (1985). Active perception vs. passive perception. In *Proceedings of the Workshop on Computer Vision*, 55-59.

Ballard, D. H. (1991). Animate vision: An evolutionary step in computational vision. *Journal of the Institute of Electronic, Information, and Communication Engineers*, 74, 343-348.

Ballard, D. H., Hayhoe, M. M., and Pelz, J. (1995). Memory representations in natural tasks. *Journal of Cognitive Neuroscience*, 7, 66-80.

Ballard, D. H., Hayhoe, M. H., Pook, P., and Rao, R. (1997). Deictic codes for the embodiment of cognition. *Behavioural and Brain Sciences*, 20, 723 - 767.

Baddeley, A. D., & Hitch, G. J. (1974). Working memory. In G. Bower (Ed.), *The Psychology of Learning and Motivation*, (V. 8, 47-90). New York: Academic Press.

Bensinger, D. G (1997). *Visual Working Memory in the Context of Ongoing Natural Behaviors*. Unpublished Ph.D. thesis, University of Rochester, Dept. of Brain and Cognitive Sciences.

Brooks, R. (1991). *Intelligence Without Reason*. Massachusetts Institute of Technology Technical Report 1293.

Hayhoe, M. M., Bensinger, D. G, and Ballard, D. H. (1998). Task constraints in visual memory. *Vision Research*, 38, 125-137.

Logie, R. H. (1995). *Visuo-Spatial Working Memory*. Hillsdale: Lawrence Erlbaum Associates.

Magnuson, J. S., Sagerer, G., Hayhoe, M. M., and Ballard, D. H. (1997). The role of fixations in task automatization. Investigative Ophthalmology and Visual Science, 38(4), S963.