UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

HYPOTHESIS GENERATION: TEMPORAL DYNAMICS

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

In partial fulfillment of the requirements for the

Degree of

DOCTOR OF PHILOSOPHY

By

NICHOLAS DAVID LANGE
Norman, Oklahoma
2011

HYPOTHESIS GENERATION: TEMPORAL DYNAMICS

A DISSERTATION APPROVED FOR THE
DEPARTMENT OF PSYCHOLOGY

BY

Dr. Rick Thomas, Chair

Dr. Michael Dougherty

Dr. Scott Gronlund

Dr. Marlys Lipe

Dr. Robert Terry

Dedication

For my mother and father who made my possibilities possible.

Acknowledgements

First and foremost I thank Rick Thomas for his invaluable mentorship and support throughout my development in graduate school. I am intellectually indebted to Rick Thomas and Michael Dougherty for their work and insights upon which this dissertation builds. Without their previous efforts the present work would not have been possible. I thank Scott Gronlund for his many contributions to my development as a researcher over the last six years. Additionally, I thank Robert Terry and Marlys Lipe for their thoughtful considerations concerning my work and my development.

Table of Contents

List of Tables

List of Figures

Abstract

In order to bring structure to many of the judgment and decision making problems people encounter, decision makers are often required to generate, from memory, hypotheses explaining their observations. This dissertation focuses on this predecisional process of hypothesis generation which underlies and supports much judgment and decision making behavior. Although we are beginning to understand a great deal about the mechanisms governing the generation and utilization of hypotheses (Thomas, R.P. et al. 2008) more work is needed to fully appreciate how these retrieval, judgment, and choice processes operate in real-world task environments.

The present research addresses temporal dynamics underlying hypothesis generation processes. As temporal dynamics are an inevitable precondition for information acquisition and utilization, a full understanding of hypothesis generation processes will remain speculative without systematic examination of the influences of such dynamics. Four experiments examined various pertinent issues in an effort to provide fundamental insights upon which more complete theory can be developed. The influence of information order, information activation in working memory, information agreement, information use (grouped vs. isolated), and working memory allocation were examined. Furthermore, two novel methodologies are forwarded providing unique approaches for assessing the active contents of working memory through time. By exploiting biases in which visual attention is drawn towards items matching the contents of working memory

these measures are able to index the current contents of working memory at any given point in time.

Several critical findings emerged from this set of experiments. First, results indicate that people tend to weight later information more heavily than earlier information in some, but not all, circumstances. Second, the role of information activation in working memory was implicated as governing its contribution to the hypothesis generation processes wherein more active items contribute more. Third, it was found that the acquisition of information that is inconsistent with hypotheses under consideration causes people to discard these inconsistent hypotheses. This work provides important insights into how internal working memory dynamics interact with external dynamics in shaping the hypothesis generation process that can be used to support the development of a comprehensive computational theory.

Chapter 1: Introduction

Dynamic decision making (DDM) has been a subfield of investigation within the wider decision making literature since the 1960s (Edwards, 1962; Toda, 1962). Toda's initial investigation utilized a game dubbed "fungus eater" in which participants controlled a robot mining for uranium on a foreign planet. Their task was to decide how best to distribute their time between resources in order to maintain an adequate power supply while optimizing mining activity. Since this novel investigation, the subfield of dynamic decision making has received a fair amount of attention as several dynamic tasks have become the topic of research such as health management (Kerstholt, 1994; Kleinmuntz & Thomas, 1987), fire-fighting (Brehmer & Allard, 1991; Gonzalez, Thomas, & Vanyukov, 2004), and navigation (Anzai, 1984; Jagacinski & Miller, 1978) amongst others.

According to Busemeyer (1999) dynamic decision-making tasks are characterized by three properties. First, a series of actions must be taken over time to achieve an overall goal. Second, these actions are interdependent so that later decisions depend on earlier actions. Third, the environment in which the decision task is situated changes both spontaneously and as a result of earlier actions. By recognizing these task characteristics, investigations in dynamic decision making have made strides in honoring the dynamic complexities one is inevitably faced with in real-world environments. As noted by Gonzalez, Thomas, & Vanyukov

(2004), real-world tasks (and those investigated within DDM) are dynamically complex due to the presence of nonlinear relationships amongst environmental variables, multiple loops, and delays in feedback to the decision maker. Without addressing such complexity a full understanding of decision making behavior will remain elusive.

In this dissertation I address temporal dynamics from a slightly different approach than has been taken in previous DDM research. My interest in this work is the predecisional process of hypothesis generation by which decision makers develop beliefs that bring structure to judgment and decision making problems. This process relies on the retrieval of hypotheses from long-term memory that are then brought into working memory for further consideration and use. It is my contention that in order to understand the temporal dynamics of hypothesis generation, and the higher-level decision making tasks reliant upon this process, we must gain a greater appreciation of the fine-grained working memory dynamics unfolding over time in support of these tasks. In this way the present work departs from, yet complements, much of the previous work of DDM which has focused largely on dynamics external to the decision maker.

The research perspective adopted in the present work lies at the confluence of internal and external dynamics as it is assumed that an understanding of how internal cognitive dynamics bidirectionally interact with external dynamics will provide a fuller picture of decision making in dynamic task environments. It is with this goal in mind that we look towards a recent model of working memory dynamics within list memory tasks as a guidepost suggesting how working memory may operate in service of data acquisition, hypothesis generation, maintenance, testing, and judgment processes as they evolve over time. These ideas are further elaborated below.

Chapter 2: Hypothesis Generation & Temporal Dynamics

The Problem

Hypothesis generation is a predecisional process by which we formulate explanations and beliefs regarding the occurrences we observe. As such, hypothesis generation represents one of our most fundamental cognitive faculties and its ubiquity in our lives cannot be understated. Given this ubiquity, it is no surprise that hypothesis generation forms a core component of several professions. Auditors, for instance, must generate hypotheses regarding abnormal financial patterns, mechanics must generate hypotheses concerning car problems, and intelligence analysts must interpret the information they receive. Perhaps the clearest example, however, is that of medical diagnosis. A physician observes a pattern of symptoms presented by a patient (i.e., data) and uses this information to generate likely diagnoses (i.e., hypotheses) that explain the patient's disease state. Given these examples, the importance of developing a full understanding of hypothesis generation processes is clear as impoverished hypothesis generation may result in severe consequences.

When engaged in hypothesis generation tasks cognitive limitations place constraints on data acquisition. Important to the present work is the fact that data acquisition is generally occurs serially over some span of time. This, in turn, dictates that individual pieces of data are acquired in some temporal relation to one another. These constraints, individual data acquisition over time and the relative ordering of data, are likely to have significant consequences for hypothesis generation processes. Given these basic constraints it is intuitive that temporal dynamics must form an integral part of any comprehensive account of hypothesis generation processes.

At present there exists limited data on the temporal dynamics of hypothesis generation tasks. Accordingly there exists no comprehensive theory regarding how temporal dynamics influence hypothesis generation processes. Thus, the constraints operating over these processes are not yet well understood and until they become so a full understanding of hypothesis generation processes will remain speculative.

The Goals of this Work

The primary goal of this work is the identification of basic constraints to inform a computational model of hypothesis generation honoring temporal dynamics. By empirically targeting key theoretical issues, the breadth and character of the assumptions required in developing such a formal account can be refined.

The theoretical issues addressed in this work cover five main inquiries. 1) It is likely that decision makers do not use all available data to an equal degree in hypothesis generation processes. The first investigation aimed to delineate the effect of the serial position of data on hypothesis generation and evaluation. The overall goal of this query was to examine how the weightings applied to individual data are influenced by serial position. 2) The second area of inquiry examined if the level of working memory (WM) activation associated with various pieces of data predicts their relative contributions to hypothesis generation. 3) It is not yet clear how people make use of data that is inconsistent with a current set of hypotheses. The specific question investigated here was whether or not inconsistent data prompt the purging of hypotheses from working memory in favor of a new round of generation or if beliefs in current hypotheses are simply revised in light of the new information. 4) The effect of maintaining multiple pieces of data in working memory prior to generation was contrasted with generation following individual pieces of data. This manipulation afforded comparisons between maintenance processes preceding generation with those of belief adjustment and the effect of re-cueing based on data isolated in time. 5) Lastly, there exists no data to date on the dynamic allocation working memory resources during hypothesis generation tasks. Tradeoffs between data and hypotheses in working memory following hypothesis generation were explored.

A secondary goal was the development and validation of novel empirical methodologies designed to assess the active contents of working memory. The main advantages of these techniques are that they are less invasive and potentially more sensitive than existing techniques. The standard technique used to measure hypothesis generation in previous paradigms has been to simply ask the participant what they believe to be the likely explanation(s) of what they have observed (e.g., Fisher et al., 1983). As a consequence of this type of prompting, the contents of working memory are substantially perturbed (e.g., interference of information being maintained, new information called into working memory to develop an answer to the prompt) [1]. Thus, simply asking the participants about the active contents of their memory systems is likely an insufficient method to investigate fluctuations of working memory contents (and their activations) over time.

By adapting paradigms from the visual attention literature and exploiting automatic processes of visual search it may be possible to assess the contents of working memory while obviating significant amounts of perturbation. The methodologies forwarded may allow inference of working memory activation by examining attentional performance over very brief displays (≈ 200 – 500 milliseconds). Given such short durations, these methods can be thought of as taking "snapshots" of the contents of working memory[2].

[1] It is possible to infer likely hypotheses generated though judgments (e.g., probability judgment). However, the elicitation of such judgments will perturb the active contents of working memory in the same manner as described.

[2] The logic and conceptual details of these methodologies are treated in detail in chapter 6.

HyGene (Dougherty, Thomas, & Lange, 2010; Thomas, Dougherty, Harbison, & Sprenger, 2008), short for **hy**pothesis **gene**ration, is a computational architecture addressing hypothesis generation, evaluation, and testing. This framework has provided a useful and comprehensive account through which to understand the cognitive mechanisms underlying these processes. This process model is presented in Figure 1.

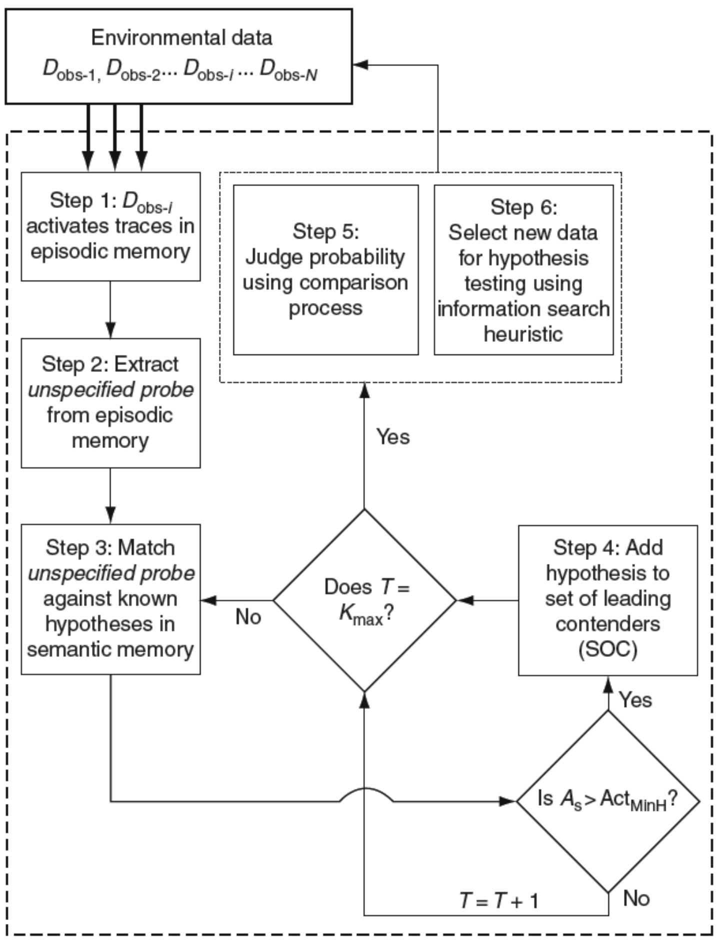


Figure 1: Flow diagram of the HyGene model of hypothesis generation, judgment and testing

HyGene rests upon three core principles. First, it is assumed that hypothesis generation represents a generalized case of cued recall. Data observed in the environment (Dobs), which one would like to explain, act as cues prompting the retrieval of hypotheses from long-term memory (LTM). For instance, when a physician examines a patient, he/she uses the symptoms expressed by the patient as cues to related experiences in LTM. These cues activate a subset of related memories in LTM from which hypotheses are retrieved. These retrieval processes are indicated in steps one, two, & three of Figure 1.

As viable hypotheses are retrieved from LTM they are placed in the Set of Leading Contenders (SOC) as demonstrated in step four. The SOC represents HyGene's working-memory construct to which the second principle applies. HyGene's second principle holds that the quantity of hypotheses that can be maintained at one time is constrained by cognitive limitations as well as task characteristics. That is, the more working memory resources that one has available to devote to the generation and maintenance of hypotheses, the more accommodating the SOC will be of additional hypotheses. Working-memory capacity places an upper bound on the amount of hypotheses (and data) that one will be able to maintain at any point in time. In many circumstances, however, attention will be necessarily divided by a secondary task. Under such conditions this upper bound is reduced as the alternative task siphons resource that would otherwise allow the population of the SOC to its unencumbered capacity (Dougherty & Hunter, 2003a; Dougherty & Hunter, 2003b; Sprenger & Dougherty, 2006; Sprenger et. al., 2011).

The third principle states that the hypotheses maintained in the SOC form the basis from which probability judgments are derived and provide the frame from which hypothesis testing is implemented. This principle underscores the function of hypothesis generation as a predecisional process underlying higher level decision making tasks. The tradition of much of the prior research on probability judgment and hypothesis testing has been to provide the participant with the options to be judged or tested. HyGene highlights this as limiting the scope of the conclusions drawn from such procedures as decision makers in real world tasks must generally generate the to-be-evaluated hypotheses themselves. As these higher-level tasks are contingent upon the output of the hypothesis generation process, any conclusions drawn from such experimenter-provided tasks are necessarily limited to such conditions.

These assumptions form the core of HyGene's theoretical framework and as such have been essential to the model's ability to explain the extant phenomena regarding hypothesis generation, probability judgment, and information search. It is likely, however, that the HyGene framework is in need of additional assumptions to more fully capture hypothesis generation processes as deployed in real-world tasks and environments. Perhaps the most vital assumption yet to be incorporated is that temporal dynamics will influence hypothesis generation processes.

HyGene in its current form is *static* with regards to data acquisition and use. The model receives all available data from the world simultaneously and engages in only a single iteration of hypothesis generation. Given the static nature of the model, each piece of data used to cue memory contributes equally to the recall process. There is reason to suspect, however, that all available data do not contribute equally[3]. What is needed is an understanding of working memory dynamics as data acquisition, hypothesis generation, and maintenance processes unfold and evolve over time in hypothesis generation tasks. As we assume that hypothesis generation represents a generalized case of cued recall, we look towards a recent model in the memory literature as a guide for how working memory dynamics might operate in hypothesis generation tasks.

The Context Activation Model (Davelaar et al., 2005) is one of the most comprehensive models of list memory recall dynamics to date. This model was developed to inform an ongoing debate in the memory literature concerning the necessity of postulating the involvement of working memory to account for the extant data from a host of recall paradigms. As various single store models have provided good explanations of much of the extant data (e.g., SIMPLE (Brown, Neath, & Chater, 2007), TCM (Howard &

[3] Additionally HyGene allows a large set of cues to be used in the recall process (e.g. 9). Given well known capacity limitations on working memory it is unlikely that such a large set could contribute to the recall process.

Kahana, 2001)) many researchers have suggested that there is no reason to posit that working memory plays a significant role in memory recall. The Context Activation Model (CAM) provides an argument counter to this trend by demonstrating that the utilization of a working memory store not only allows for a better account of the extant data, but also accurately predicts novel (and surprising) phenomena. As the novel effects predicted by this model are not readily addressable by single store models, CAM provided a strong argument for the involvement of working memory processes in list recall tasks.

The model's ability to predict nuanced phenomena results from the manner in which the activations of individual items (e.g., words from a list) dynamically fluctuate over time in concert with the assumption that working memory is capacity limited. The activations of individual items in the model are multiply determined by several factors continuously interacting over time. These factors are: 1) bottom up sensory input currently received, 2) recurrent self-excitation, 3) global inhibition from competing items, 4) excitation from semantic associates, 5) the item's activation in the previous time step, and 6) stochastic noise.

A representative example of the activation trajectories produced by these forces working in tandem is presented in Figure 2 which displays the activation trajectories of twelve items over the course of a study period of a recall task. The activation levels are plotted on the y-axis as F(X) with time on the x-axis. The threshold for inclusion in working memory is demarcated by the horizontal line at F(X) = 0.2. As can be seen, the individual items enter and exit working memory in accordance with their activation levels achieved over time.

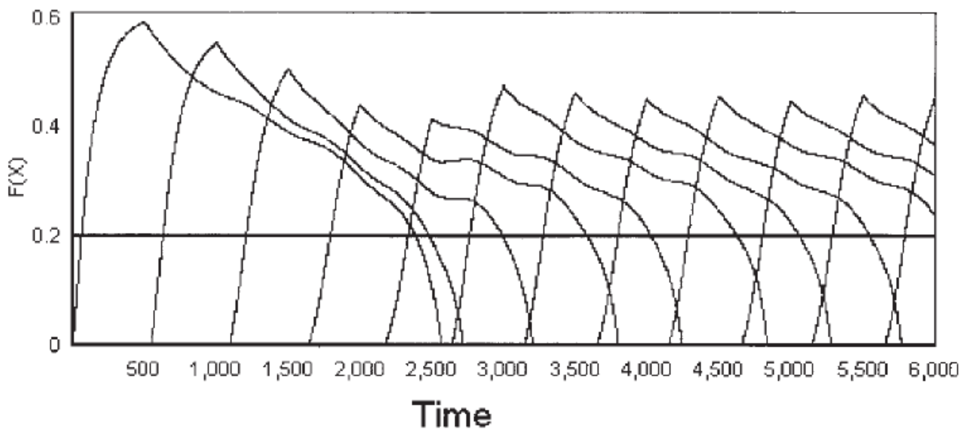


Figure 2: Activation trajectories of 12 items during a study period of a free recall task produced by the Context Activation Model

Although the activation trajectories displayed in Figure 2 are somewhat characteristic of the trajectories produced by the model under commonly employed task characteristics in free and cued recall tasks, these trajectories are highly sensitive to changes in task parameters. For instance, if the display rate of each item is increased from 400 to 200 iterations (i.e., when the display rate is twice as fast), the activation trajectories behave remarkably different, as displayed in Figure 3. In the top panel we see the trajectories exhibiting the gradual replacement of earlier list items with later list items as greater competition emerges throughout the study period. However, as demonstrated in the lower panel, when the presentation rate is high it is not the later items that are active in working memory at the end of the study period, but the initial items. The stunted activation levels of the later items results from the decreased period of strong bottom up activation for each item that would, in most other cases, allow these items to surmount the global inhibition of items currently in working memory and facilitate their activation beyond the WM threshold.

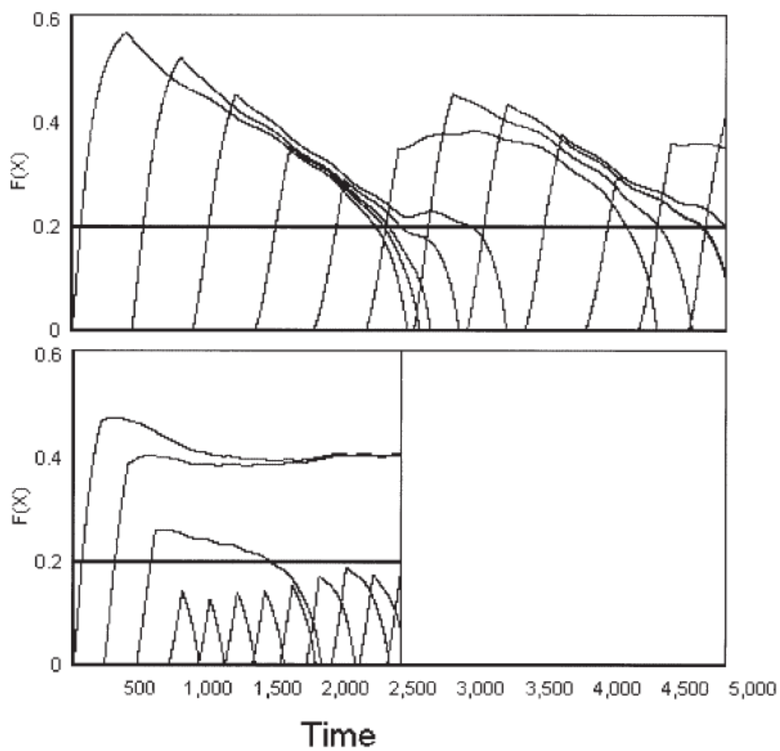


Figure 3: Activation trajectories of 12 items under a relatively slow presentation rate (top) and a fast presentation rate (bottom) produced by the Context Activation Model

It is this dynamic sensitivity that allows the model to account for a host of findings in the memory literature. The verification of novel predictions generated by the model, such as the presentation rate effect described above, bolsters the validity of the model's account of working memory dynamics in recall tasks. Given that HyGene is based on the premise that hypothesis generation is a case of cued recall, it stands to reason that the incorporation of the WM dynamics of CAM with the existing HyGene architecture may afford appropriate mechanisms by which to build a temporally dynamic model of hypothesis generation. In addition, the behavioral dynamics of the Context Activation Model provide heuristic value in considering how WM may handle data acquired over time in hypothesis generation tasks.

Underlying the present work is the assumption that the activations of individual pieces of data in working memory fluctuate dynamically over time prior to and during their use as cues to retrieve hypotheses from long-term memory. It is further posited, as a broad hypothesis for the experiments that follow, that the contributions made by individual pieces of data to the generation process (i.e., their weightings) will be determined by their activations in working memory[4]. Lastly, it is assumed that working memory limitations will place an upper bound on the amount of data that can actively contribute to hypothesis generation processes.

[4] There are additional factors that are likely to contribute to data weightings in working memory as well (e.g., diagnosticity, utility) that but these factors must be the focus of independent investigation and are set aside for the time being.

Chapter 3: Experiment 1- Data Serial Position & Hypothesis Generation

Order effects are pervasive in investigations of memory and decision making (Hogarth & Einhorn, 1992; Murdock, 1962; Page & Norris, 1998; Weiss & Anderson, 1969). Such effects have even been obtained in hypothesis generation tasks specifically. Sprenger (2007) found that people tended to weight cues more heavily that were presented later in a sequence rather than early cues.

The generalized order effect paradigm was developed by Norman Anderson (1965; 1973) and couched within the algebra of information integration theory to derive weight estimates for individual pieces of information presented in impression formation tasks (e.g., adjectives describing a person). This procedure involved embedding a fixed list of information with a critical piece of information at various serial positions. The differences in the critical information serial position thus defined the independent variable, and given that all other information was held constant between conditions, the differences in final judgment were attributable to the serial position of the critical data. The present experiment represents an adaptation of this paradigm to assess the impact of data serial position on hypothesis generation and probability judgment.

Method

Participants

One hundred and twenty one participants from a large Midwestern university participated in this experiment for course credit.

Design (1-Way Within-Subjects Design)

The design of Experiment 1 was a one-way within-subjects design with data order as the independent variable. The ecology for this experiment as defined by the conditional probabilities between the various hypotheses and data is shown in Table 1. Each of the values appearing in this table represents the probability that the data will be positive (e.g., fever) where the complementary probability represents the probability of the data being negative (e.g. normal temperature). As you can see, the only diagnostic piece of data is D1 whereas the remaining cues, D2-D4, are non-diagnostic.

	Symptoms				
		D1	D2	D3	D4
Diseases	H1: Metalytis	0.8	0.6	0.6	0.6
	H2: Zymosis	0.2	0.6	0.6	0.6
	H3: Gwaronia	0.2	0.6	0.6	0.6

Table 1: Disease x Symptom ecology of Experiment 1

Table 2 displays the four data orders. Each of these orders was identical (D2 → D3 → D4) except for the position of the D1 data within them. An additional baseline condition, in which D1 was presented in isolation, was implemented as well. All participants received and judged all data orders. However, given that carry over effects are expected, this will be considered a between subjects variable in terms of the planned analysis in which only the first order received by the participant (i.e., first trial) is to be considered uncontaminated from preceding trials.

	→ Presentation Position →			
	1	2	3	4
Order 1	**D1**	D2	D3	D4
Order 2	D2	**D1**	D3	D4
Order 3	D2	D3	**D1**	D4
Order 4	D2	D3	D4	**D1**
Baseline	**D1 only**			

Table 2: Data presentation orders and baseline control used in Experiment 1

Procedure

The procedure began with an exemplar training phase in which a series of hypothetical pre-diagnosed patients was presented to the participant in order for them to learn, through experience, the contingencies between the hypotheses and data. Each of these patients was represented by a diagnosis at the top of the screen and a series of test results (i.e., symptoms) pertaining to the columns of D1, D2, D3, & D4[5]. The specific results of these tests respected the probabilities in Table 1 from patient to patient. Thus, over the course of training the participants encoded this statistical ecology in long term memory. Prior to beginning the exemplar training the participants were informed that they had an opportunity to earn a $5.00 gift card to Wal-Mart if they performed well enough in the task. If the participant scored greater than 60% on a diagnosis test (described below) then they were awarded the gift card at the end of the experiment.

A short diagnosis test phase directly followed exemplar training. This test was included to allow discrimination of participants that adequately learned the contingencies between the data and the hypotheses in the training phase[6]. The participants were presented with the

[5] Examples of the displays used in this experiment appear in Appendix A. These examples demonstrate the exemplar training screens used as well as those used at elicitation. These displays are very similar to those used in Experiments 2 & 3 as well.

[6] Previous investigations in our lab utilizing exemplar training tasks have demonstrated variation in conclusions drawn from results conditionalized on such learning data against entire non-conditionalized data set. Therefore including this learning test allows us a

symptomologies of a series of 12 patients (4 of each disease). The data of each of the patients was presented simultaneously on a single screen. The participants' task in this phase was to correctly diagnose the patient(s) with the disease of greatest posterior probability given the presenting symptoms. No feedback on this test performance was provided. This was followed by a distracter phase in order to clear working memory of information processed during the diagnosis test phase. The distracter task consisted of a series of fifteen arithmetic equations for which the correctness or incorrectness was to be reported (e.g. 15/3+2 = 7? Correct or Incorrect?). This distracter task was self paced.

The elicitation phase then proceeded. First, the diagnosis task was described to the participants as follows: "You will now be presented with additional patients that need to be diagnosed. Each symptom of the patient will be presented one at a time. Following the last symptom you will be asked to diagnose the patient based on their symptoms. Keep in mind that sometimes the symptoms will help you narrow down the list of likely diagnoses to a single disease and other times the symptoms may not help you narrow down the list of likely diagnoses at all. It is up to you to determine if the patient is likely to be suffering from 1 disease, 2 diseases, or all 3 diseases. When you input your response make sure that you respond with the most likely disease 1st. You will then be asked if you think there is another likely disease. If you think so then you will enter the

check on the presence of such discrepancies in addition to obtaining data that may inform how greater or lesser learning influences the generation process.

next most likely disease 2^{nd}. If you do not think there is another likely disease then just hit the Spacebar. You will then have the option to enter a 3^{rd} disease or hit the Spacebar in the same manner. To input the diseases you will use the first letter of the disease, just as you have been during the training and previous test."

The participant was then presented with the first patient and triggered the onset of the stream of data themselves when they were ready. Each datum of each patient was presented individually for 1.5 seconds. The order in which the data were presented was determined by the order condition as shown in Table 2. Following the presentation of the last datum the participant responded to three sets of prompts: the diagnosis prompts (as previously described in the instructions to the participants), a single probability judgment prompt of their highest ranked diagnosis, and a thought listing pertaining to the probability judgment.

The probability judgment was elicited with the following prompt: "If you were presented 100 patients with the symptoms of the patient you just observed how many would have [INSERT HIGHEST RANKED DISEASE]?" The thought listing pertaining to the probability judgment was then solicited with the following prompt: "Now try and remember all of the ideas you had while thinking about the response you just made. Please list these thoughts." The participant was then presented with the remaining orders in the same manner with distracter tasks intervening between each trial.

Hypotheses & Predictions

A recency effect was predicted on the grounds that more recent cues would be more active in working memory and contribute to the hypothesis generation process to a greater degree than less recent cues. Given that the activation of the diagnostic cue (D1) in working memory at the time of generation was predicted to increase in correspondence with its serial position, increases in the generation of Metalytis were predicted to be observed with increases in the serial position occupied by D1. Correspondingly, decreases in the generation of the alternatives to Metalytis were expected with increases in the serial position of D1. Bolstering this prediction is the finding of recency in hypothesis generation by Sprenger (2007) discussed above. Previous work has demonstrated the dilution effect whereby supplemental non-diagnostic information serves to reduce probability judgments (Smith, Stasson, & Hawkes, 1998 ; Troutman & Shanteau, 1977; Zukier, 1982). It was predicted that the presence of supplemental non-diagnostic data would serve to dilute the contribution of the diagnostic cue to the generation process thereby decreasing the generation of Metalytis relative to the baseline condition. Additionally, probability judgments assigned to Metalytis were expected to be lower for these conditions with multiple cues in comparison to the baseline condition.

Results

Carry-over effects were evident in the data as evidenced by a significant interaction between order condition and trial, $\chi^2(3) = 10.3$, $p < 0.05$. In light of this, only the data from the first trial was subjected to analyses as it was assumed that this was the only uncontaminated trial for each subject. Nominal logistic regression was carried out over the generation data to examine the effect of data serial position on the generation of Metalytis, the disease with the greatest posterior probability given the data[7]. These tests were conducted with the baseline condition omitted and therefore incorporated orders 1 thru 4. As can be seen in Figure 4 , the results demonstrated a significant trend for the output of Metalytis (in any output position) with increases in the serial position of the diagnostic data, $\chi^2(1) = 5.15$, $p < 0.05$. Additionally, there was a significant trend for Metalytis being the first hypothesis output, $\chi^2(1) = 6.72$, $p < 0.01$ as displayed in Figure 5. Interestingly, however, there was no trend for the generation of either alternative hypothesis, $\chi^2(1) = 0.18$, $p = 0.667$. As a side note, comparisons of the generation of Metalytis as first in the output sequence vs. the generation of Metalytis in any position in the output sequence revealed differences across orders 1 – 3, z = -2.01, $p < 0.05$, z = -

[7] For experiments 1 & 2 analyses were conducted on both the entire dataset (including participants scoring poorly on the diagnosis test phase of the procedure) and the conditionalized data set in which such participants were excluded. The primary analyses reported are of the entire data set. Unless otherwise noted the reader can assume equivalent results from the concomitant analyses conditionalized on only those scoring above 50% on the learning test.

2.63, $p < 0.05$, z = -2.08, $p < 0.05$ respectively, and in the baseline condition as well, z = -2.51, $p < 0.05$.

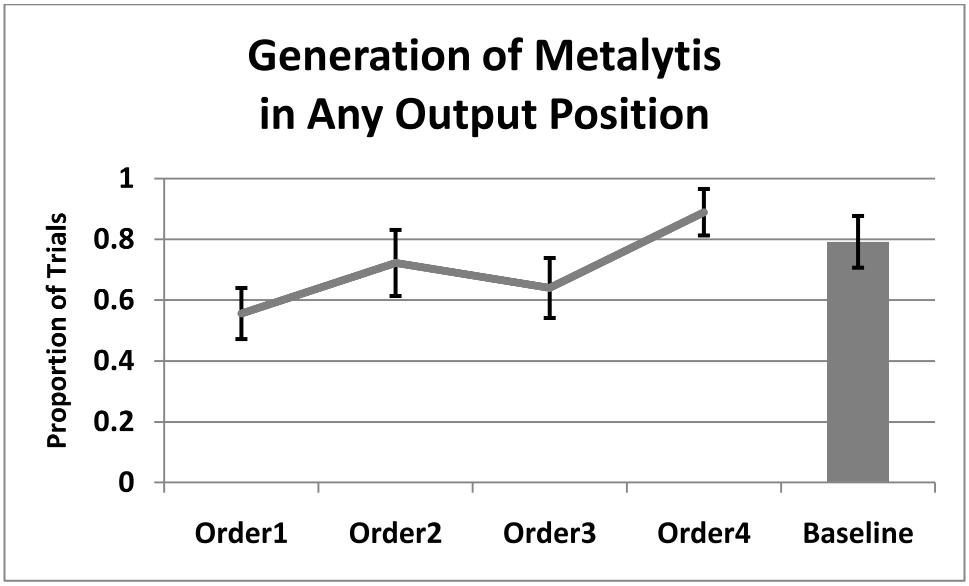


Figure 4: Generation of Metalytis in any output position by order condition in Experiment 1

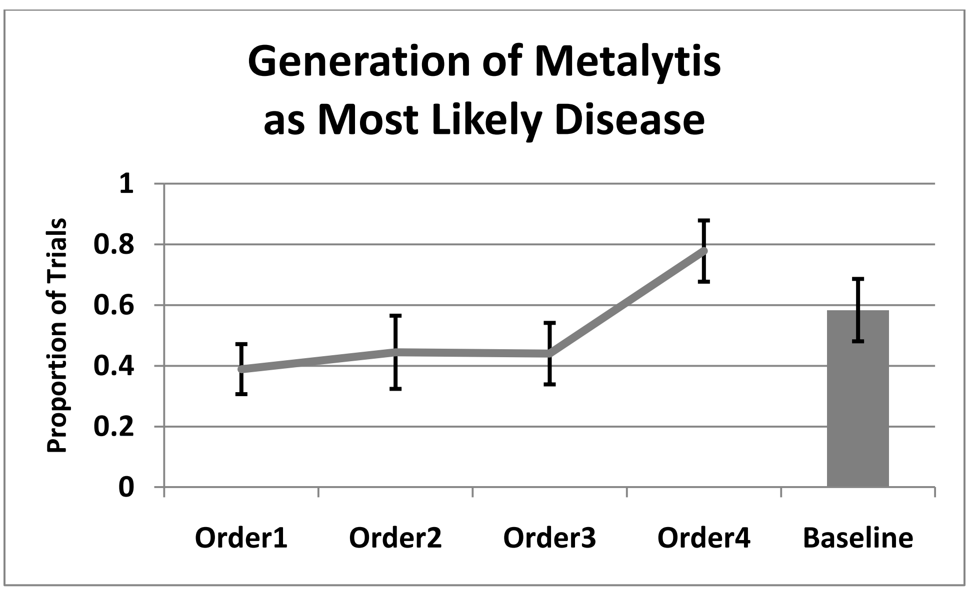


Figure 5: Generation of Metalytis as most likely disease by order condition in Experiment 1

Binomial tests were carried out between each order (1-4) and the baseline condition to test the dilution effect of generation. Only Order 1 demonstrated departure from generation performance in the baseline condition. This was the case for Metalytis generated in any output position, z = -3.49, $p < 0.005$, and for Metalytis generated first in the output sequence, z = -2.37, $p < 0.005$.

The number of hypotheses generated between order conditions did not differ, $F(1,93) = 0.16$, $p = 0.69$, ranging from an average of 1.66 to 1.89 hypotheses. Additionally, as can be seen in Figure 6, the probability judgments (of Metalytis when it was the highest ranked hypothesis) did not differ either, $F(1,41) = 2.43$, $p = 0.127$. Comparisons were carried out between the probability judgments in each order condition and the baseline condition in order to test for a dilution effect in the judgments. There were no deviations from baseline for Order 1, $F(1,25) = 2.27$, $p = 0.144$, order 2, $F(1,18) = 1.54$, $p = 0.23$, order 3, $F(1,22) = 0.01$, $p = 0.905$, or order 4, $F(1,24) = 0.21$, $p = 0.649$.

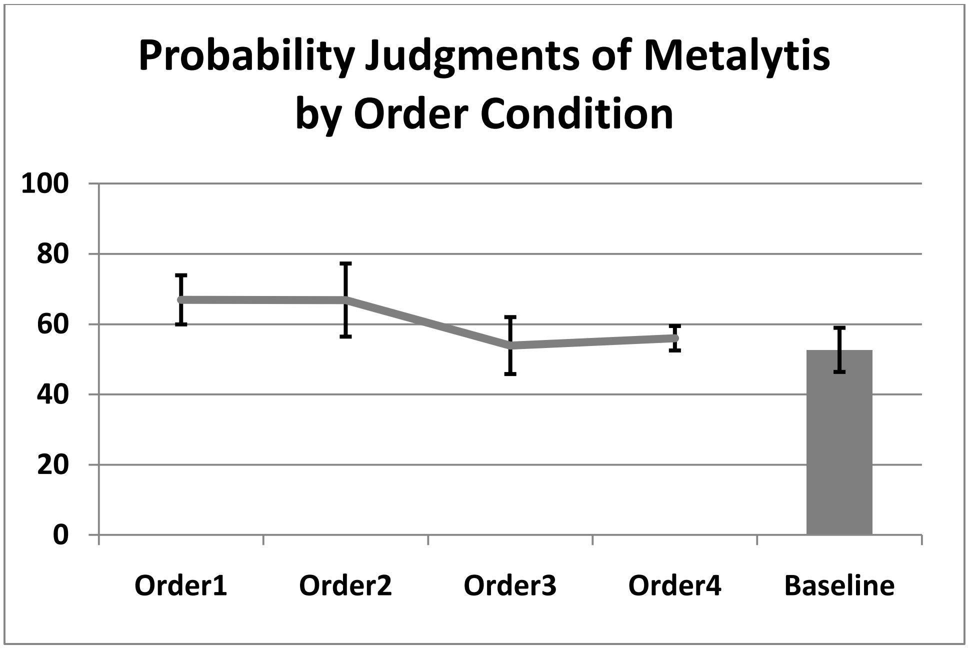


Figure 6: Probability judgments of Metalytis when generated as most likely hypothesis in Experiment 1

Discussion

The primary prediction of the experiment, that generation of the most likely hypothesis would increase in correspondence with increased recency of the diagnostic data, successfully obtained as demonstrated through the trend of Metalytis generation. Additionally this prediction held over the generation of Metalytis both as the most likely disease (first output position) and for the generation of Metalytis generated anywhere in the output sequence. This finding clearly demonstrates that not all available data contribute equally to the hypothesis generation process, that some data are weighted more heavily than others, and that the serial position of a datum can be an important factor governing the weight allocated to it in the generation process. Furthermore, these results are entirely consistent with the notion that the data weightings of the generation process are governed by the amount of working memory activation allocated each datum. This finding is in agreement with a finding in a similar exemplar training paradigm (Sprenger, 2007) in which data order was manipulated and it was found that later cues had the greatest impact on the hypothesis sets generated.

There are, however, two alternative explanations to consider that do not necessarily entail clear-cut activation differences. First, it could be the case that all data in working memory at the time of generation were equally weighted, but that the likelihood of D1 dropping out of working memory increased with its distance in time from the generation prompt. Such a process would likely result in a similar trend as that seen in the data when averaged over participants. Future investigations measuring working memory capacity may be helpful in illuminating the veracity of this explanation. Secondly, it could be the case that the participants engaged in spontaneous rounds of generation following each piece of data presented. Because the hypothesis generation performance was only assessed after the final piece of data in the present experiment, such "step by step" generation would result in stronger generation of Metalytis as the diagnostic data is presented closer to the end of the list. For instance, if spontaneous generation is occurring as each piece of data is being presented, then when the diagnostic datum is presented first there remains three more rounds of generation (from non-diagnostic data) that could obscure the generation of that particular round. As the diagnostic data moves closer to the end of the data stream the likelihood that that particular round of generation will be obscured by forthcoming rounds diminishes.

These two alternative accounts of the present data underscore the need for convergent evidence from additional methodologies to discover how working memory activation influences hypothesis generation. The methodologies presented in Chapter 6 provide a platform on which such convergent evidence could be observed when paired with the procedure of the present experiment.

The secondary hypothesis of a dilution effect of hypothesis generation did not obtain as order 1 was the only condition in which Metalytis was generated less than in the baseline condition. This result is somewhat unexpected given that HyGene predicts more precise (i.e., less errorful) generation behavior for the baseline condition in which the diagnostic data appeared in isolation. A simulation of this behavior demonstrates this point. HyGene's episodic memory was endowed with the ecology used in Experiment 1 and the model was presented with 4 different cues for hypothesis generation. The first cue was D1 in isolation and thus represents of the baseline condition of the present experiment. The second cue was a compound cue comprised of D1 & D2, thereby adding a non-diagnostic cue. The third and fourth cues followed this pattern of adding one non-diagnostic cue to the existing compound cue and were, D1,D2, & D3 and D1,D2,D3, & D4 respectively[8].

As can be seen in Figure 7, the generation of H1 decreased with the addition of non-diagnostic data to the compound cue while concomitant increases of the alternative hypotheses were observed. Furthermore, as demonstrated in Figure 8, this increased generation of alternatives to H1 was greater than the decrease in H1 generation causing the overall number of hypotheses generated to increase with the addition of non-diagnostic cues. Although these simulation results do exhibit the predicted dilution of H1 generation it is important to note that the differences of H1 generation rates across cue conditions is slight. In order to provide a better test of this particular prediction it may be necessary to change some characteristic of the simulation (e.g, ecology) so that the prediction becomes more pronounced than is the case in the present simulation. Such an investigation could be taken up in future research.

[8] For the interested reader, the HyGene parameters used for this simulation were L=0.85,Ac=.1,KMAX=5,Phi=4.

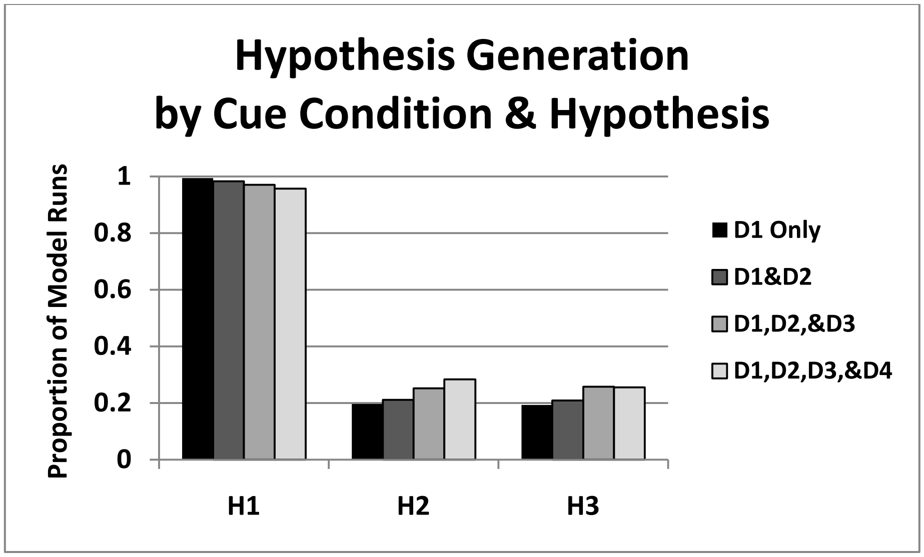


Figure 7: Simulation results for hypothesis generation of HyGene by hypothesis and cue condition

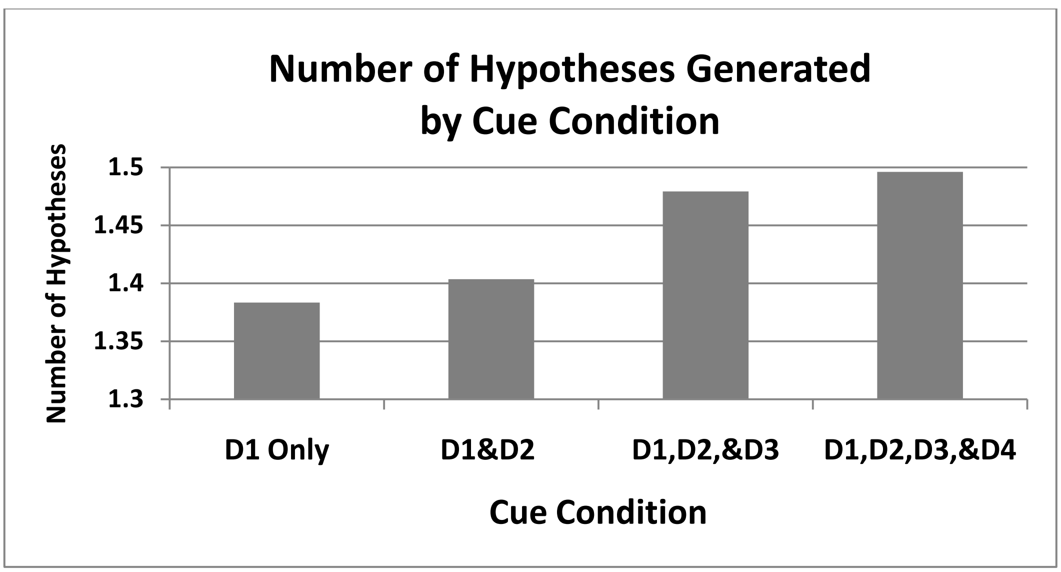


Figure 8: Simulation results for total number of hypotheses generated by HyGene across cue conditions

The probability judgments observed in the present experiments did not differ across order conditions. Because the probability judgments were taken on only the highest ranked hypothesis and analyzed only with respect to the runs on which Metalytis was ranked first, the conditions under which the probability judgments were analyzed were constrained. It should be noted of course that the focus of the present experiment was to address generation behavior. An independent experiment utilizing a similar paradigm to explicitly examine probability judgment behavior would be useful and informative. Lastly, such a future investigation might utilize a different type of prompt for the elicitation of the judgments as the prompt used in the present investigation ("Out of 100 patients like this patient…") might have engaged frequency judgment mechanisms rather than the comparative mechanisms assumed to underlie judgment in HyGene (Spenger & Dougherty, 2006). Future investigation, therefore, may benefit from a different prompting than that used in the present experiment.

The goal of Experiment 1 was to determine how relative data serial order contributes to hypothesis generation processes. It was predicted that data presented later in the sequence would be more active in memory and would thereby contribute more so to the generation process. Such an account predicts a recency profile for the generation of hypotheses from long-term memory. This effect obtained in the data as the generation of the most likely hypothesis (Metalytis) increased with the serial position of the diagnostic data. Despite this positive result it is not entirely clear if the grounds on which the prediction were made are entirely discernable in the present experiment as the aforementioned alternative explanations would predict the same result. Converging evidence for the hypothesis that data activation governs the contribution of individual data to the generation processes should be sought.

Chapter 4: Experiment 2- Data Activation & Hypothesis Generation

As mentioned in the introduction, the Context Activation Model of memory (Davelaar et al., 2005) successfully predicts a unique effect of presentation rate in which a shift in recall performance, from a recency profile to a primacy profile, is observed under high presentation rates (See Figure 3 above). As alluded to in Chapter 2, this particular finding provided strong evidence for the dual store account of list recall memory as single store models cannot readily produce this effect. Of importance for the present experiment is the fact that this particular demonstration supports the notion that presentation rate alters the activation of items in working memory. Insofar as activation of data in working memory governs hypothesis generation, it can be expected that generation should be sensitive to presentation rate differences as well. The current experiment manipulated the presentation rate of a stream of data prior to a hypothesis generation prompt. Obtaining differences in generation behavior between presentation rate conditions in the present experiment would support the hypothesis that activation in working memory drives the weightings allocated to individual pieces of data in cuing hypotheses from long-term memory.

Method

Participants

One hundred and fifty six participants from a large Midwestern university participated in this experiment for course credit.

Design (1-Way Between-Subjects Design)

The independent variable was the presentation rate of the data presented; there were two levels. In one condition each piece of data was presented for 300 ms (Fast condition) and in the other condition each piece of data was be presented for 1500 (Slow condition) ms. Two additional conditions were included as controls. It is possible that the initial data in the Fast condition may act as a forward mask on the following data thereby confounding an interpretation of the activation account with a spurious masking effect. The first control condition was set up to guard against this account. This condition was identical to the Fast condition, but included a mask prior to the first piece of data. If the initially presented datum does in fact act as mask on the forthcoming data in the fast condition, then masking the first piece of data should preclude its acquisition thereby disallowing its use in the generation process. The second control condition was included to investigate a potential effect of the discrepancy between the total task durations between the Fast and Slow conditions. This condition was also identical to the Fast condition except that a retention interval was placed between the offset of the last datum and the hypothesis-generation prompt thereby equalizing the total trial time with the Slow condition. Thus, there were four conditions in all: Fast, Slow, Fast with Mask (FastMask), and Fast with Retention Interval (FastRI).

Procedure

As in Experiment 1, the procedure began with exemplar training on the ecology displayed in Table 3. This was followed by a test to verify learning and a distracter task (arithmetic verification task) prior to elicitation. The experiment was again cast in terms of medical diagnosis. At elicitation the data were presented in the left to right order in which they appear in Table 3. Therefore, the participants always received diagnostic data for Metalytis early in the sequence and diagnostic data for Zymosis late in the sequence.

	D1	**D2**	**D3**	**D4**	**D5**
H1: Metalytis	0.8	0.7	0.5	0.3	0.3
H2: Zymosis	0.3	0.3	0.5	0.7	0.8

Table 3: Disease x Symptom ecology of Experiment 2

The participants were warned that the data would appear briefly and were instructed to trigger the onset of the data by pressing the Spacebar. They were then presented with hypothetical trial data presentations (in which the data were replaced with numbers) in order for them to become acquainted with the presentation rates of the to-be-presented data. This was done so that the participants were not confused or caught off guard by the high presentation rate in the fast condition. When the participants pressed the Spacebar, a fixation point blinked in the center of the screen three times. Each piece of data (D1-D5) then appeared serially at this central fixation point in accordance with the present presentation rate condition. Generation was then elicited with the following prompt, “Enter the disease that this patient is most likely suffering from.” As the subjects were familiar with inputting a set response for each disease during exemplar training they were guided to use these responses (e.g., “Press the letter corresponding to the disease.”). A probability judgment and thought listing were then elicited with the following prompts; “Out of 100 patients with the same results as this patient, how many would have the same disease that you selected?”, and “Now try and remember all of the ideas you had while thinking about the response you just made. Please list these thoughts.”

Hypotheses & Predictions

It was predicted that early data would be more active in working memory at the time of generation under each of the fast presentation rate conditions. This would in turn allow the early data to contribute more so to the generation process and would be evidenced by greater generation of Metalytis. Alternatively, the opposite was predicted in the slow presentation rate condition where greater generation of Zymosis was predicted on account of the greater relative activations of the data appearing later in the sequence.

Results

Because the comparisons of interest in the present experiment are between the Slow condition and each of the conditions in which a fast presentation rate was used, comparisons were performed to assess these divergences for each fast condition individually through ChiSquare tests. As can be seen in Figure 9, results indicate that the rate of generation of Metalytis in the Slow condition only differed in the FastMask condition, $\chi^2(1) = 5.66$, $p < 0.05$, whereas comparisons of the Slow condition with the

Fast condition and the FastRI condition did not differ, $\chi^2(1) = 0.85$, $p = 0.36$ and $\chi^2(1) = 2.66$, $p = 0.1$ respectively[9].

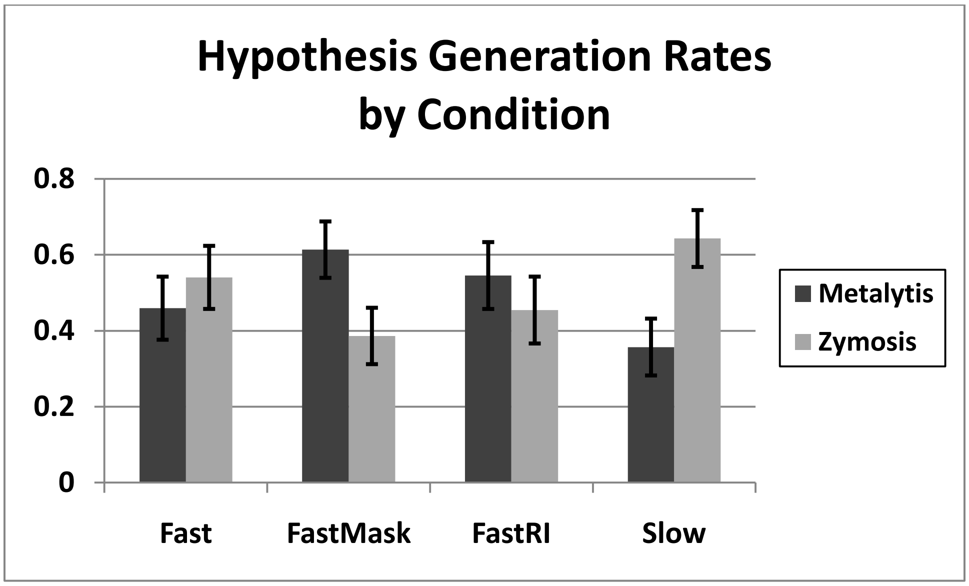


Figure 9: Hypothesis generation results of Experiment 2

Binomial tests were also carried out within condition to assess divergences from chance exhibited within each condition. Results indicate that none of the Fast conditions deviated from chance performance; Fast, z = -0.49, $p = 0.31$, FastMask, z = 1.51, $p = 0.07$, FastRI, z = 0.48, $p = 0.32$. Generation performance in the slow condition, on the other hand, did

[9] This pattern of results is maintained following the application of a learning criterion (50% correct in diagnosis test phase) and consequent purging of participants' data not meeting this criterion. However, the difference between the Slow and the FastMask conditions is attenuated to a marginal difference, $\chi^2(1) = 3.17$, $p = 0.075$.

deviate significantly from chance, z = -1.85, $p < 0.05$, as there was greater generation of Zymosis over Metalytis in this condition as predicted[10].

Because probability judgments were taken on only the most likely hypothesis generated by the participant per trial, results were analyzed conditional on both hypotheses. That is, there were a set of comparisons for those that generated Metalytis and a set of comparisons for those that generated Zymosis. Within these conditionalized groups, comparisons were made between each Fast condition and the Slow condition. These results are displayed in Figure 10 below. Within the Metalytis group there were no significant differences between any of the Fast conditions and the Slow condition. The Fast condition diverged the greatest and achieved a marginal difference $F(1,30) = 3.53$, $p = 0.07$, whereas the FastMask and FastRI conditions were equivalent to the Slow condition, $F(1,39) = 0.03$, $p = 0.87$ and $F(1,33) = 0.24$, $p = 0.63$. The judgments of the Zymosis group were uniform as well. Comparisons in this group showed no differences between the Fast, FastMask, or FastRI conditions with the Slow condition, $F(1,32) = 0.83$, $p = 0.37$, $F(1,41) = 0.17$, $p = 0.69$, and $F(1,39) = 2.01$, $p = 0.16$ respectively.

[10] This result becomes marginal following application of the 50% learning criteria as well, z = -1.3, $p = 0.09$.

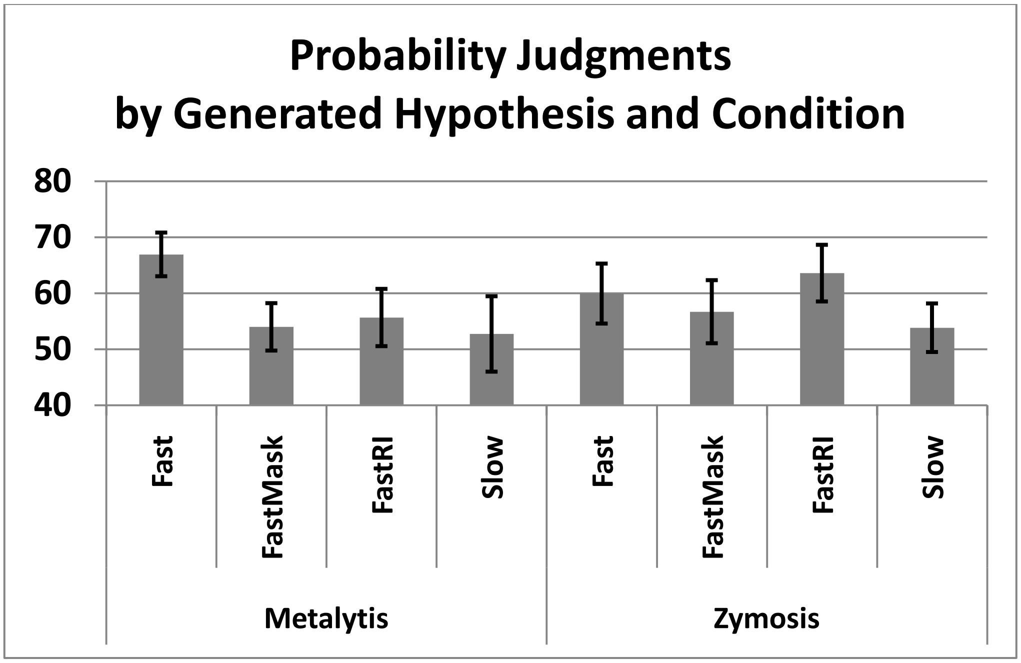


Figure 10: Probability judgment results of Experiment 2

Discussion

Overall the results of the experiment suggest that the manipulation of presentation rate did successfully influenced data activation in working memory and furthermore that this activation difference influenced hypothesis generation in the predicted manner. Although only one of the three Fast conditions demonstrated a difference, it is important to note that it was the FastMask condition in which this result obtained. Had only the Fast condition obtained the result it would appear as if a masking effect may have been present. However, the particular pattern of results that obtained inoculates the experiment from this argument while at the same time demonstrating the predicted effect of presentation rate.

The lack of consistency across the Fast conditions is still a concern however. It is possible that the presentation rate used in the Fast conditions was straddling a tenuous border over which data activations at the time of the prompt did not consistently favor any serial position(s) across or within trials. Such inconsistencies in the activation profile associated with each piece of data would result in chance performance. Davelaar's (2005) original finding of a shift from recency to primacy with increased presentation rate used a higher rate (100ms) than the rate used in the present experiment (300ms). It is possible that greater uniformity between the Fast conditions would be observed by increasing the presentation rate to something closer to the 100 ms used in the original cued recall experiment.

The results of Experiment 2 support the notion that data activation in working memory drives their contributions to the hypothesis generation process. The results of the Slow condition are entirely in line with this position, however, stronger support for this position would have been obtained had the predictions for each of the Fast conditions obtained. Future research examining this point should be pursued.

Chapter 5: Experiment 3- Data Maintenance & Data Consistency

The belief adjustment model (Hogarth & Einhorn, 1992) predicts the manifestation of order effects along three specific task characteristics; information complexity, length of data series, and response mode. Although the framework of the belief adjustment model is not entirely germane to the present investigation[11], the distinction of response mode is likely to have dramatic effects on hypothesis generation as data unfold over time. The two response modes are step-by-step (SbS), in which a response is elicited following each piece of incoming data, and end-of-sequence (EoS), in which only one response is elicited after all data have been received.

An experiment manipulating this response mode variable in a hypothesis generation task was conducted by Sprenger (2007) in which people hypothesized which psychology courses were being described by various keywords. In one task of this experiment the participants were provided with cue orders differing in the placement of diagnostic information. More specifically, in one condition the initial data were

[11] The belief adjustment model, as currently formulated, does not adequately generalize to hypothesis generation tasks. Whereas hypothesis generation tasks rely on the participant to generate alternatives from long-term memory they necessarily engender the possibility of entertaining over 2 hypotheses. In contrast, the belief adjustment model assumes that only 1 or 2 hypotheses are being estimated or evaluated which have been provided by the experimenter or task at hand. Furthermore, the present experiment doesn't manipulate the order in which cues are received which is the paradigm structure on which the model is based. Lastly, the distinction between mixed and consistent evidence espoused by the model doesn't make sense in a hypothesis generation task as evidence can only be judged as consistent or inconsistent with a hypothesis once it has been generated.

diagnostic with non-diagnostic data being presented later. In a complementary condition the opposite was the case as non-diagnostic data was followed by diagnostic data. Following the last piece of data those in the SbS conditions exhibited a recency effect wherein participants in the "late diagnostic data" condition were more likely to generate correct hypotheses than those in the "early diagnostic data" condition. The EoS condition, on the other hand, did not reveal any order effects.

The present experiment compared these response modes to examine differences between data maintenance prior to generation (EoS mode) and generation that does not encourage the maintenance of multiple data (SbS mode). Considered in another light, SbS responding can be thought of as an anchoring and adjustment process where the set of hypotheses generated in response to the first piece of data supply the set of beliefs in which forthcoming data are interpreted. The EoS condition, on the other hand, does not engender such belief anchoring as generation is not prompted until all data have been observed. As such, the SbS conditions will provide the opportunity to test for the propensity to re-generate a new set of hypotheses (and discard an existing set) in the face of inconsistent data.

Method

Participants

One hundred and fifty seven participants from a large Midwestern university participated in this experiment for course credit.

Design (2 Response Mode x 2 Ecology Between-Subjects)

As previously mentioned, the first independent variable was the timing of the generation & judgment promptings provided to the participant as dictated by the response mode condition. This factor was manipulated within-subject. The second independent variable, manipulated between-subjects, was the consistency of the 2nd piece of data (D2) with the most likely hypotheses likely to be entertained by the participant given the 1st piece of data. Table 4 displays the ecologies of the present experiment wherein the data are arranged in the temporal order in which they were presented in the experiment (i.e., D1 → D2 → D3 → D4). The only difference between the ecologies was the conditional probability of D2 under H1 in which this probability was 0.9 in Ecology A and 0.1 in Ecology B. Given that D1 should prompt the generation of H1 and H2, this manipulation of the ecology can be realized to govern the consistency of D2 with the hypothesis(es) currently under consideration following D1.

A	**D1**	**D2**	**D3**	**D4**
H1: Metalytis	0.9	**0.9**	0.5	0.5
H2: Zymosis	0.7	0.1	0.4	0.4
H3: Gwaronia	0.2	0.8	0.8	0.8

B	**D1**	**D2**	**D3**	**D4**
H1: Metalytis	0.9	**0.1**	0.5	0.5
H2: Zymosis	0.7	0.1	0.4	0.4
H3: Gwaronia	0.2	0.8	0.8	0.8

Table 4: Disease x Symptom ecologies of Experiment 3

Procedure

The procedure began like those of the previous experiments: exemplar training to learn the probability distributions, a test to verify learning (for which a $5.00 gift card could be earned for performance greater than 60%), and a distractor task prior to elicitation. The experiment was cast in terms of medical diagnosis where H1, H2, & H3 represented disease states and D1-D4 represented various test results (i.e., symptoms). The data were presented in the left to right order in which they appear in Table 4. The elicitation prompts for hypothesis generation and thought listing were the same as those used in Experiment 1 & 2. However the probability judgment prompt was slightly different from that of Experiments 1 & 2. The probability judgment prompt used in the present experiment was as follows: “How likely is it that the patient has [INSERT HIGHEST RANKED DISEASE]? (Keep in mind that an answer of 0 means that there is NO CHANCE that the patient has [INSERT HIGHEST RANKED DISEASE] and that 100 means that you are ABSOLUTELY CERTAIN that the patient has [INSERT HIGHEST RANKED DISEASE].) Type in your answer from 1 to 100 and press Enter to continue.” Note that in the SbS mode the thought listing was only carried out following the final (4^{th}) probability judgment.

Hypotheses & Predictions

Given that the step-by-step response mode affords the elicitation of generation following every piece of presented data, the predictions for this mode cover two time frames. Following the presentation of D2 persons in Ecology A were predicted to generate H1 to a greater extent than those in Ecology B who were expected to purge H1 from their hypothesis set in response to its inconsistency with D2. Furthermore, it was predicted that those in Ecology B would generate H3 to a greater extent than those in Ecology A as they would recue memory following the purging of working memory. It was expected that those in Ecology A would not recue long term memory as the consistency of D2 with the most likely hypothesis, H1, would engender a belief updating process while maintaining H1. Following D4, final generation performance in both ecologies was predicted to resemble the results following D2 as none of the forthcoming data rule out the hypotheses being considered.

The general prediction for the end-of-sequence response mode was that recency would be demonstrated in both ecologies as the most recent data should contribute more strongly to the generation process. Therefore, greater generation of H3 relative to the alternative hypotheses was expected in both ecologies. This reveals a prediction between response mode conditions for final generation within Ecology A as the EoS condition was predicted to favor H3 while the SbS was expected to favor H1.

Results

The main dependent variable analyzed for this experiment was the generation behavior as converted into rank scores. As the participants were instructed to enter their hypotheses in accordance with their subjective ranks the following conversion was employed. The first hypothesis reported received a rank score of 3. If a second and third hypothesis were generated they received a rank of 2 and 1 respectively. When a hypothesis was not generated it was given a rank score of 0. These rank scores were then treated as interval data in the analyses that follow.

Multiple tests were run to assess the effect of ecology following the presentation of D2 (in the SbS response mode condition). Firstly, the between subjects effect was tested on the overall ranks of H1 following D2.

As can be seen in Figure 11, H1 was ranked significantly higher in Ecology A relative to Ecology B, $F(1,76) = 6.96$, $p < 0.05$. A further test of the effect of ecology was carried out within subject by means of a difference score between H1 ranking following D1 and H1 ranking following D2. This was taken as D1 ranking subtracted from D2 ranking so that a positive adjustment in the ranking following D2 would be reflected by positive sign. As can be seen on the left hand side of Figure 12, this within subject analysis did not approach significance, $F(1,76) = 0.84$, $p = 0.362$. However, when conditionalized on those participants that generated H1 following D1 we see an effect of ecology emerge following D2 where those in Ecology B that had generated H1 previously ranked H1 significantly lower following D2, $F(1,50) = 6.61$, $p < 0.05$, as demonstrated on the right hand side of Figure 12. Similar analyses were carried out to assess differences in the rates of H3 generation following D2. The between subjects analysis did not reveal any difference between the rankings of H3 between D1 and D2 elicitations, $F(1,76) = 0.141$, $p = 0.709$. Additionally, the same within subjects difference analysis as above did not detect any difference either, $F(1,76) = 0.13$, $p = 0.719$, nor did a conditional analysis in which participants not generating H3 following D1 were assessed for their difference in rankings following D2, $F(1,41) = 0.867$, $p = 0.357$.

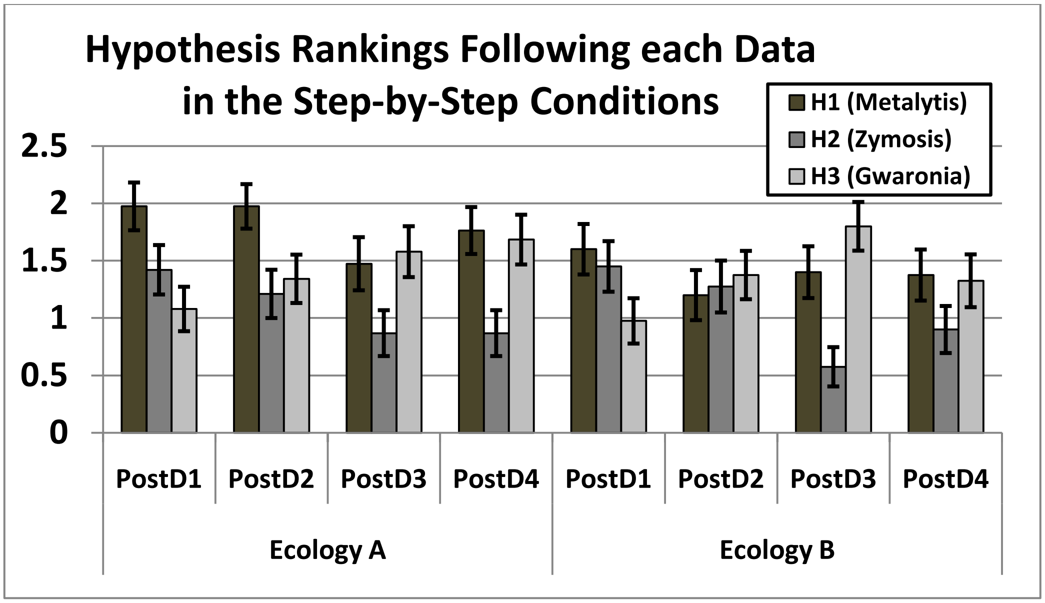


Figure 11: Hypothesis rankings in Experiment 3 by ecology conditions following each piece of presented data in the SbS conditions

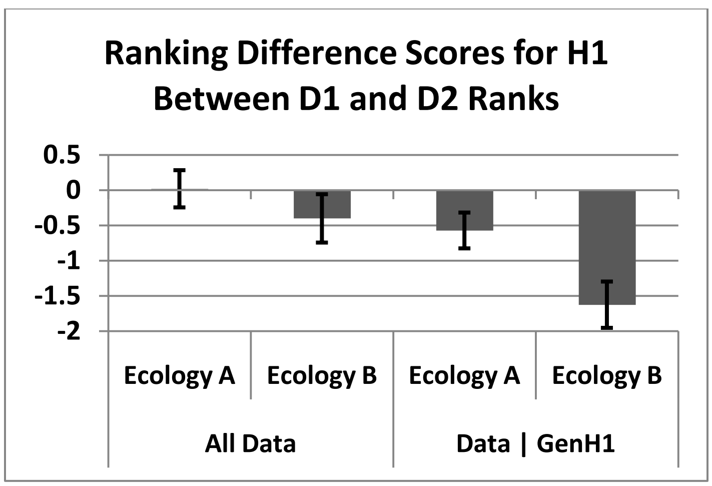


Figure 12: Ranking difference scores of Hypothesis 1 in Experiment 3. All participants (left hand side) vs. only participants that generated H1 following D1 (right hand side)

To test the hypothesis that generation performance following D4 would resemble generation performance following D2 differences between the rankings of each hypothesis following D2 were compared with their rankings following D4. No differences were detected in the rankings of H1, H2, or H3 within Ecology A, $F(1,37) = 0.776$, $p = 0.384$, $F(1,37) = 1.437$, $p = 0.238$, and $F(1,37) = 1.778$, $p = 0.191$, nor were any differences evident within Ecology B, $F(1,39) = 0.443$, $p = 0.51$, $F(1,39) = 1.489$, $p = 0.23$, and $F(1,39) = 0.03$, $p = 0.864$.

To test the prediction that H3 would be generally favored in the EoS conditions, comparisons of hypothesis rankings were carried out within each ecology condition. As shown on the right side of Figure 13, a marginal difference in the hypothesis rankings was revealed in Ecology A, $F(2,39) = 3.19$, $p = 0.052$, whereas no difference was found in Ecology B, $F(2,36) = 0.68$, $p = 0.513$. Furthermore, Ecology A comparisons between individual hypothesis rankings revealed that H3 that was not favored. Rather it was H1 that was the favored hypothesis as H1 differed significantly from both H2, $F(1,40) = 5.09$, $p < 0.05$, and H3, $F(1,40) = 5.84$, $p < 0.05$, while H2 and H3 did not differ from one another, $F(1,40) = 0.00$, $p = 1.00$.

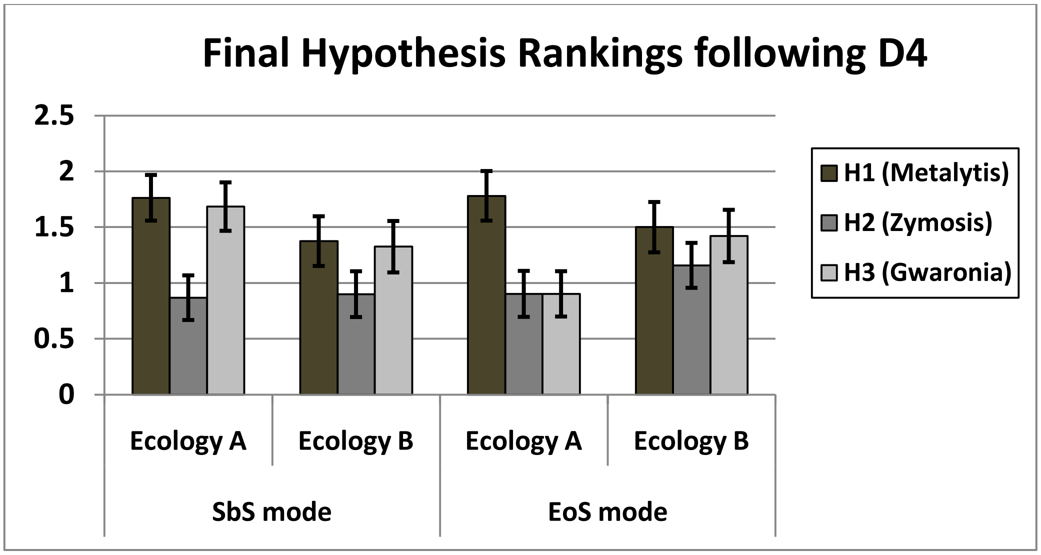


Figure 13: Final hypothesis rankings following the last piece of data (D4) in Experiment 3 by response mode & ecology conditions

The last prediction to be tested was the interaction between response mode and ecology following D4. It was predicted that within Ecology A H3 would dominate for the EoS condition whereas H1 would dominate in the SbS condition. Although a response mode by ecology interaction was detected, $F(1,153) = 11.512$, $p < 0.001$, it did not manifest in the predicted manner. Rather, this interaction was due to H1 being favored in the EoS condition (as reported above) while H1 and H3 vied for dominance in the SbS condition as they both differed from H2, $F(1,37) = 7.43$, $p < 0.01$ and $F(1,37) = 6.51$, $p < 0.05$, but did not differ from one another, $F(1,37) = 0.045$, $p = 0.834$.

Analysis of the number of hypotheses generated across response mode and ecology conditions resulted in neither main effects of response mode, $F(1,153) = 0.5931$, $p = 0.4424$, or ecology, $F(1,153) = 0.2133$, $p =$

0.645, but did reveal a significant interaction, $F(1,153) = 9.987$, $p < 0.005$, as displayed in Figure 14.

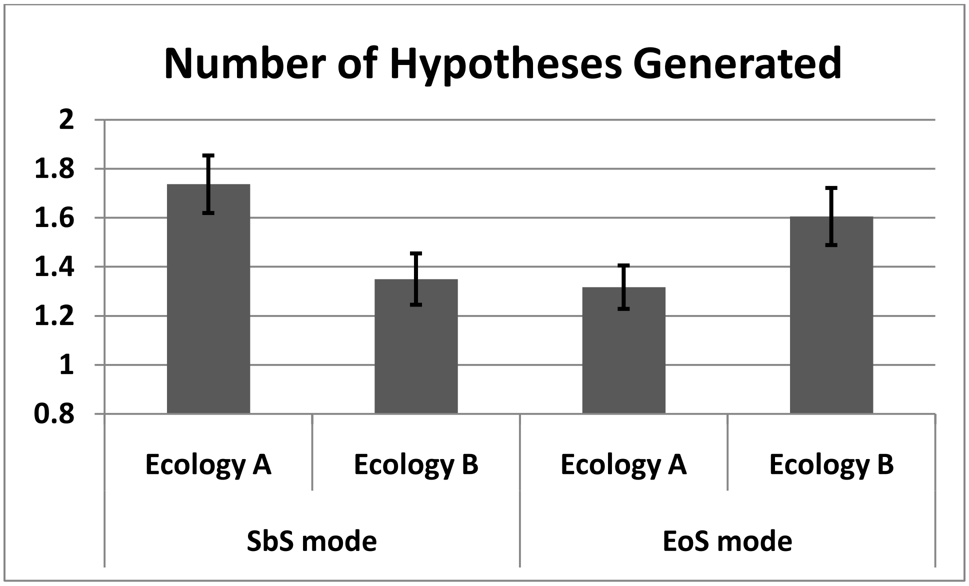


Figure 14: Number of hypotheses generated by response mode and ecology conditions

Probability judgments were not analyzed for this experiment as low N (due to differences in highest ranked hypotheses across conditions) made tests unfeasible.

<u>Discussion</u>

The present experiment has provided an interesting window into two distinct processing dynamics. Additionally the generation behavior was quite rich as the participants not only reported their generated hypotheses, but did so in rank order. The first dynamic under investigation was how inconsistent data influenced the hypothesis set currently under consideration. In the step-by-step conditions we found clear evidence that people do purge working memory in response to the inconsistency of a newly received cue as those that had generated the most likely hypothesis in the initial round of generation were less likely to rank it following the inconsistent cue. This can be viewed as consistent with an extension of the consistency checking mechanism currently employed in HyGene. The present data suggests the hypotheses currently under consideration are checked against the newly acquired data and are purged in accordance with its (in)consistency[12]. This is slightly different from, although entirely compatible with, the operation of the current consistency checking operating within HyGene over a single round of hypothesis generation. The consistency checking operation within the current version of HyGene checks each hypothesis retrieved into working memory for its consistency with the data used as a cue to its retrieval as the set of leading contenders is populated. The consistency checking mechanism exposed in the present experiment, however, suggests that people check the consistency of *newly* acquired data against hypotheses generated from *previous* rounds of generation as well. If the previously generated hypotheses fall below some

[12] Additionally, it is quite likely that the purging of hypotheses from working memory is contingent in large part on the diagnosticity of the newly acquired data in ruling out one or more of the hypotheses currently being considered.

criteria of agreement with the newly acquired data they are purged from working memory.

There was no evidence, however, for the re-cueing of LTM following D2 in either ecology as the rankings of H3 did not differ from those following D1[13]. This could have been due to the fact that the generation rankings (and rates) of H3 following D1 were somewhat high in comparison to what was expected. Given that H3 was only 0.2 related to D1 it was expected that the generation and ranking of this hypothesis as likely would be extremely low in response to D1. If this had been the case then we may have had a better opportunity to observe differences between these successive rounds of hypothesis generation. Perhaps a follow up experiment in which the D1-H3 association is changed to 0.0 would offer a better test of the discrete recuing of LTM following inconsistent data.

As predicted there were no changes in hypothesis rankings between the D2 and D4 elicitations suggesting that hypothesis sets remain relatively stable in the face of consistent data. However this conclusion may be premature. Examination of the PostD3 elicitations in Figure 11 suggests that there may have been more movement in the rankings between D2 and D4 elicitations than is captured by the present analyses. This is particularly evident in Ecology B as the rankings of H1 & H3 diverge from H2 following D3 only to re-converge following D4. The factors governing this peculiar fluctuation in the hypothesis rankings are not entirely clear. It is

[13] No difference was detected for analyses of the rates of generation of H3 following D1 vs. D2 either.

possible that the generation behavior following D4 was qualitatively different from the initial 3 rounds of generation due to some subtle demand characteristic. As there were four tests presented in the training phase, the participants may have assumed (correctly) that the round of generation following the 4^{th} bit of data was their last opportunity to diagnose the patient at hand. This may have prompted an additional search of memory, either for their previous beliefs of the patient's disease state or for the previous symptoms that the patient had previously presented with. It does appear that perhaps the participants engaged in such a recall of previous belief states as the generation post D3 is highly in line with generation behavior implied by Bayesian diagnosticity whereas the generation behavior following D4, with the same exact diagnosticity and directionality as D3, reverted towards earlier belief states, namely towards H1.

The prediction for the generation behavior in the end-of-sequence conditions was that recency would obtain, which would be evidenced by higher rankings for H3 as the last two cues in the presentation sequence clearly implicate this disease as most likely. This result, however, did not obtain. Rather it was a clear preference for H1 that obtained in Ecology A and indifference between hypotheses that obtained in Ecology B. Far from a recency effect being evidenced in the data, these results suggest something of a primacy effect. The results of Ecology A make it somewhat difficult to argue convincingly for a primacy *effect* in the data as the Bayesian preference in this condition is in favor of H1 and the preference for H1 may thus not represent a primacy effect insofar as an *effect* is equated with bias[14]. That is, if all data contributed equally to the generation process then a preference for H1 would still result. However, the results from Ecology B clearly support the presence of a primacy effect as the lack of a clear preference for H3 demonstrates severe departure from the Bayesian prescription and the equal contribution of all available data to the generation process[15].

Given that the EoS conditions of the present experiment are procedurally very similar to the elicitations used in Experiment 1 it becomes important to reconcile the manifestation of the present primacy effect with the recency effect obtained in Experiment 1. First, it is possible that there is something inherently different in the processing of inconsistent data that causes the primacy effect to emerge in the present study. In Experiment 1, although most of the presented data were non-diagnostic, they were not inconsistent with the most likely hypothesis. In the Ecology B condition of the present experiment however, some of the data presented in the data stream were in conflict with the most likely hypothesis at various points. Whereas D2 was inconsistent with the most likely hypothesis suggested by D1 it was also the case that D1, could be seen in retrospect to be inconsistent with D2-D4. It is possible that the mere presence of evidence pointing in different directions altered the generation process. Additionally,

[14] The posterior probabilities of H1, H2, & H3 in Ecology A are 0.64, 0.04, & 0.32 respectively.
[15] The posterior probabilities of H1, H2, & H3 in Ecology B are 0.17, 0.08, & 0.75 respectively.

unlike Experiment 1, every piece of data presented in the present experiment was diagnostic of one of the hypotheses. Of course this aspect of the ecology was explicitly controlled out of Experiment 1. It is difficult to ascertain why exactly the two different usages of the presented data appear between Experiments 1 and 3.

The EoS condition primacy effect is also important to compare to the Sprenger (2007) results discussed above in which the SbS conditions revealed recency and no order effects were revealed in the EoS conditions. In Sprenger's experiment the participants were presented with a wide array of data from which to generate hypotheses, nine pieces in total. Participants in the present experiment, on the other hand, were only provided with four cues. It's possible that people's strategies for cue usage would differ between these conditions. Whereas the present experiment provided enough data to fill working memory to capacity (or barely breach it), Sprenger's experiment provided an abundance of data that necessarily breached the working memory capacities of the participants. It is possible that the larger pool of data engendered a larger pool of strategies to be employed than in the present study. The final generation behavior collapsed across participants would then represent mixtures of the deployed hypothesis generation strategies. Understanding the strategies that people employ and the retrieval plans that developed under such conditions (Fisher, 1987; Gillund & Shiffrin, 1984; Raajmakers & Shiffrin, 1981) as well as how these processes contrast with situations in which less cues are available is a crucial aspect of dynamic memory retrieval in need of better understanding. Further experimentation should address these issues and the additional methodologies presented in the next section may be of use in this endeavor.

Of additional interest in the present experiment is the fact that the EoS results strongly resemble the SbS results following D2 within both ecology conditions. This could be taken to suggest that those in the EoS condition were utilizing the initial cues while potentially excluding the later cues altogether. Fisher & Gettys (1987) suggested that people generally tend to use only a subset of the pool of data provided. They estimated that people generally only use two cues when three cues are available and three cues when four cues are available. Such estimates are in line with this view of the picture evoked from this comparison between the EoS and SbS conditions. Again, the methodologies presented in the next section may be beneficial in determining how data usage differed between these conditions.

The present experiment set out to test how inconsistent data altered the generation process when generation was elicited after every piece of data and to test how generation behavior differed between this step-by-step elicitation and elicitation occurring only after all the data had been presented. It was hypothesized that when the participants received data that was inconsistent with the hypotheses being currently maintained that they would purge working memory of those hypotheses and recue long-term memory. Clear evidence was found for the purging of inconsistent hypotheses from working memory, however there was no evidence for the recuing of long-term memory. It was additionally hypothesized that differences would emerge between the step-by-step and end-of-sequence conditions. It was hypothesized that the EoS conditions would demonstrate recency effects. On the contrary, it was primacy that was revealed in these conditions. The exact reasons for the disjunction between the primacy findings in the present experiment with those of the recency effects in Experiment 1 are not immediately clear, but this apparent dissociation represents a potentially interesting problem and a unique challenge to account for computationally.

Chapter 6: Non-Invasive Methods for Assessing the Contents of Working Memory

Past research on hypothesis generation behavior has relied almost exclusively on the overt reporting of beliefs and/or evaluations of those beliefs as the sole dependent measures observed. Although this methodology has been very informative in guiding our understandings of hypothesis generation behavior it is likely too limited a method to inform fine grained details of hypothesis generation dynamics. It is my goal to understand how subtle fluctuations in the activations of items in working memory influence the hypothesis generation process as it unfolds over time. Overt reports do not adequately support this endeavor. This is due to the fact that the prompting of overt reports, as well as the reporting itself, is likely disruptive to the processes unfolding at the time the prompt is delivered. Such prompts interfere greatly with the contents of working memory during their service to the hypothesis generation task at hand. Furthermore, one cannot be sure of the influence of the prompt on the report thereby elicited. For instance, a prompt for overt report is likely to act as two distinct requests in various instances. If the participant has already engaged in a round of hypothesis generation prior to the prompt, then it can be taken as a request for processing at that point in time to be severed in service of the readout of the current contents of working memory. If, on the other hand, hypothesis generation has not been initiated at the time of the prompt then it is taken as a request to engage in hypothesis generation. It is likely that such prompting procedures largely result in mixtures of these two behaviors that are then indistinguishable in the data. With these considerations in mind it can be seen that overt promptings necessarily distort a potentially important aspect of the time course of the generation process and may in some cases introduce an artificiality that would not otherwise be imposed in naturalistic settings.

In order to untangle the fine grained information processing dynamics at the heart of naturalistic decision making tasks, a less intrusive method of assessing the contents of working memory is needed. Such a method would benefit from four key characteristics. Firstly, the technique would be non-invasive in that it would obviate significant amounts of perturbation to the processing at hand or the current contents of working memory. Secondly, it would be possible to deploy the measure on-line, within and throughout the task. Thirdly, the measure would be item specific in that it would allow measurement at the level of individual items (e.g., data or hypotheses) rather than simply the engagement of generalized processing as is measured in much of modern neuroscience (e.g., fMRI, EEG, MEG). Lastly, the method would be of superior usefulness if it were sensitive not only to the current processing of individual items, but to their current levels of activation in memory. At present there exists no methodology satisfying these criteria. Modern neuroscience techniques afford some of these characteristics, but they are not item specific and as such will be of limited usefulness in addressing how individual items are employed in the hypothesis generation process over time.

By exploiting automaticities in visual search and modifying paradigms from the literature on visual attention it may be possible to develop two methodologies possessing these crucial characteristics. The first methodology involves gauging early oculomotor behavior (via eye tracking) following the onset of visual search displays to assess the active contents of working memory. The second methodology modifies the standard rapid serial visual presentation procedure of attentional blink paradigms and measures deficits in attentional performance as an index of working memory activation.

Eye Movements

It has been demonstrated that the contents of WM can effectively guide visual search/attention towards items presented in a search array in visual search tasks (Downing, 2000; Huang & Pashler, 2008). Furthermore, the work of Soto et al. (2005) provides strong evidence that this bias of visual attention is automatic in nature. Soto demonstrated that proportionally less initial fixations fell on the search target when the working memory item appeared in the array. Critically, this pattern held even when the subjects were told beforehand that the prime they had been provided to hold in working memory would never match the target position and thus would always hurt their visual search performance (i.e., search efficiency). Additionally, Soto et. al. (2007) demonstrated that these effects manifest when verbal primes are used in lieu of visual primes and Moores et al. (2003) demonstrated that semantic relation between WM contents and items in a search array (rather than exact matches) drew first fixations and greater fixation durations.

Together this suggests that what is most active in WM at any given instant should draw initial visual attention, thereby providing a "snapshot" of the most highly activated item in WM (amongst the items provided in the search array). By presenting search arrays to participants at various points throughout a task it should be possible to assess which items are highly active in WM (e.g., observed data & generated hypotheses). I refer to this procedure of using initial eye movements to assess the activation of items in working memory as the Memory Activation Sensitive Saccades (MASS) procedure.

Attentional Blink

The second methodology utilizes a modified attentional blink paradigm (see Dux & Marois, 2009 for a review) to assess the activation of particular items through time. Pashler and Shiu (1999) demonstrated that when the contents of WM match an item presented in a rapid serial visual presentation (RSVP) sequence it causes attention to "blink" over an item following shortly after as evidenced by an inability to report a target following the matching item. By applying the same logic as above, attentional blink trials may be used to measure the active contents of working memory. That is, by deploying specially tailored RSVPs, in which an item matches a piece of data or hypothesis within a generation task, it may be possible to infer the activation of those items in working memory through the attentional blinks observed. I refer to the utilization of attentional blink sensitivity for inferring the activation of working memory contents as Memory Activation Sensitive Attentional Blink (MAS-AB).

A Hypothetical Example

Figure 15 represents a hypothetical schematic of how these measures of working memory content, either MASS or MAS-AB, would be deployed over the course of a simple hypothesis generation task. The y-axis of the figure represents activation in working memory and the x-axis represents time.

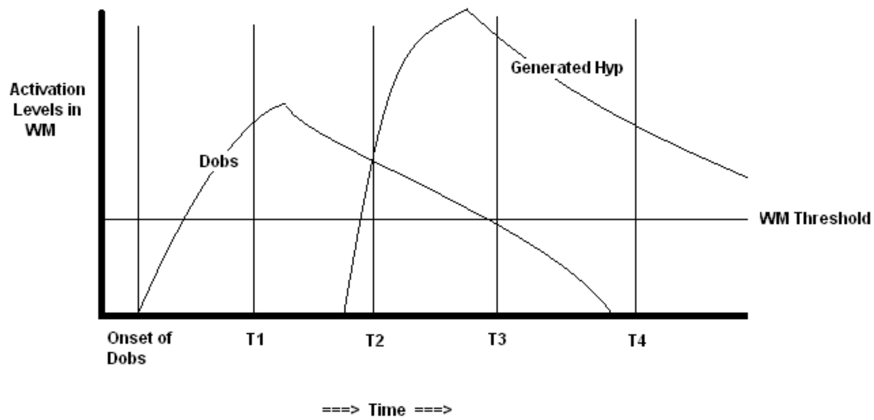


Figure 15: Hypothetical activation trajectories of observed data (Dobs) and a generated hypothesis through time. T1, T2, T3, & T4 represent MASS or MAS-AB measurement points.

The trial is initiated with the onset of the observed data (Dobs). Following the onset of the data you see its activation start to rise. At time T1 the first attentional probe is displayed on the screen (in the case of the MASS procedure this would be an array of $\approx$ 4 items lasting for about 500 ms. In the case of the MAS-AB procedure this would be the presentation of an RSVP stream and the acquisition of a response lasting about 1 second total). Given that nothing else should be active in working memory at this time we would expect that attentional biases should only manifest in reaction to the Dobs. However, as we move across time to the next deployment of the measure(s) we see that a hypothesis has now been generated. At this point we should begin to see attentional biases for both the Dobs as well as the hypothesis that has been generated in response to the data. Furthermore, depending on how the attentional probes are constructed (by probing both items at once) we may be able to directly detect the competition for attentional control exerted by each item and if these measures are activation sensitive we may be able to see this competition play out in proportion to each items' relative activation. Moving to the third attentional probe at time T3 we see that the Dobs has now dropped out of working memory as the hypothesis gains activation strength. Attentional performance would reflect this by demonstrating a strong preference for the hypothesis under consideration while no longer orienting towards the data. At T4 the contents of working memory have not changed, but the activation of the hypothesis has diminished as would be evidenced by decreased rates of attentional deployment towards it in comparison to the rates observed at T3. If eye movements are sensitive to WM activation then the oculomotor behavior we observe in hypothesis generation tasks utilizing this paradigm should facilitate the inference of rudimentary forms of the activation trajectories displayed in Figure 15.

Tests of Attentional Sensitivities to Memory Activation

The work mentioned above highlights an important connection between the active contents of working memory and attentional processes. However, this work does not tell us if early oculomotor behavior and attentional blink are differentially sensitive to the relative activations of items in working memory. If these methods are not sensitive to the relative activations of items in working memory this will limit (although not eliminate) their usefulness for assessing the working memory processing dynamics. For instance, given the work of Moores et al. (2003), a lack of activation sensitivity would make it more difficult to determine if saccadic preference is truly indicative of a hypothesis having been generated or merely reflect its semantic relation to active data. Therefore, before utilizing the above methodologies in our tasks of interest it will be useful to gain insight into the level of sensitivity they afford.

Initial tests of eye movement and attentional blink sensitivity to memory activation are described in the forthcoming sections. These tests were carried out by fusing standard paradigms from the memory and attention literatures. One of the tasks in these experiments was a standard free recall task in which a list of items (e.g., words) was studied and subsequently recalled in no prescribed order. In the present experiments, however, attentional tasks were placed in between the study and recall phase of each trial. In the case of testing eye movements the attentional measure comprised a visual search task performed within a search array containing multiple items. In the case of testing the attentional blink, the attentional measure was carried out within an RSVP stream of pictures. In both cases a member of the list studied for free recall on that trial reappeared in pictorial form in the attentional measure.

The serial position of the reappearing item was manipulated along a recency gradient. That is, only the last 4 items of the study list reappeared (individually across trials) in the attentional measures. The Context Activation Model (Davelaar, 2005), predicts differential activation of these recency items which largely accounts for its ability to reproduce standard recency effects in immediate free recall. Therefore, by manipulating the reappearing item's serial position we effectively manipulated the amount of WM activation associated with different items appearing in the attentional tasks. The goal of the present experiment was to translate these differences in memory activation into differences in obligatory visual search. This would be evidenced by recency gradients in attentional bias (in the case of eye movements) and attentional deficits (in the case of attentional blink) along the recency portion of the serial position curve.

Experiment 4a: Eye Movement Sensitivity to Working Memory Activation

Method

Participants

Sixteen participants from a large Midwestern university participated in this experiment for course credit.

Design (2 Target Present/Absent x 4 WM Probe Recency Within-Subjects)

The design of the present experiment served to structure the contents appearing in the search arrays upon which the participants performed the visual search task on each trial. The participant's search task was to report the presence or absence of a search target in the search array. Therefore, the first independent variable was the presence or absence of the target in the search array. Fifty percent of the trials contained the target. Furthermore, one of the items in the search array was always pulled from a list of words presented for memorization prior to the search task. The second independent variable was the serial position occupied by this reappearing item in the memory list. This reappearing item was manipulated to be from either the very last, 2nd to last, 3rd to last, or 4th to last items from the memory list. The implementation of these manipulations will be further elaborated below in the explication of the experiment's procedure.

Procedure

Prior to beginning the experiment the participant was placed in an eye-tracking apparatus (Arrington ViewPoint) that recorded several parameters of the participants' eye movements throughout the experiment. Each trial began with the presentation of a study list, consisting of sixteen words (at a rate of 1.25 sec./word), in the center of the display that were to be memorized for later free recall. Directly following the last list word, a visual search array appeared containing 4 icons. The search task was to report either the presence or absence of a target item. The target item was always the same picture of a jacket. The presence/absence of the search target was randomized within the other independent variable condition (WM probe serial position) so that there were an equal proportion of target present and target absent trials within each WM probe condition. Figure 16 displays an example of the search arrays that appeared in the experiment. This figure represents a target present trial as the jacket target icon can be seen on the left hand side. Following the search response, the participant was provided eight opportunities to recall words from the study list. The recall task was encouraged as the primary task. The participants were also instructed to report their search response as quickly as possible while maintaining accuracy.

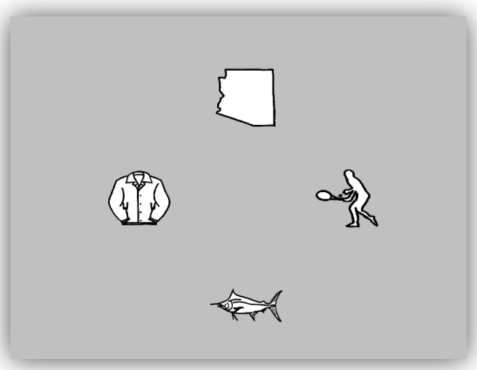


Figure 16: Example of visual search array used in Experiment 4a

One of the items appearing in the search array was an icon representing one of the study list items. This reappearing list item was randomly drawn from 1 of the last four serial positions (13th, 14th, 15th, or 16th). There were 128 trials total for each participant which allowed 32 trials per WM probe condition. The shape of the array was randomly displayed in either a square or a diamond configuration and the target randomly appeared in one of four orientations (0°, 90°, 180°, or 270°) on each target-present trial in order to encourage overt search. Each item in the array appeared at 10.3° of visual angle from the center of the display and subtended approximately 4.8° x 4.8°.

The word lists were constructed using a subset of categorized nouns from Murdock (1976). Eight categories were used for the study lists. To achieve the list length of sixteen words two randomized lists were created by pulling one word from each category without replacement and then concatenating these two lists. This ensured that the first and last 8 words in each list were locally semantically unassociated. Distractor items were taken from an additional 4 categories. Distractor items were drawn independently from these lists such that no array contained multiple distractors from the same category.

Hypotheses & Predictions

The focal prediction for this experiment was that greater attention would be allocated to WM probes appearing in the search array with increases in the serial position occupied by the WM probe in the free recall list. Specifically it was predicted that a greater proportion of early eye movements would be directed towards the WM probe as its serial position increased (i.e., when it had appeared more recently).

Results

Prior to analysis, calibration criteria for the eye tracker were applied. The participants were required to be within 1.15 degrees from 8 points around a circular arrangement on the screen for each of four calibrations throughout the experiment following every other trial block. Only the data from 11 subjects meeting these criteria was subjected to analysis. Furthermore, two exclusion criteria were applied to each trial. A trial was discarded if the participant's gaze was not contained in a central 3.8° x 3.8° region of interest (ROI) upon the onset of the search array and/or if their search response was inaccurate. As a result, 12% of data was excluded from the analyses[16]. To assess which items were initially visited by the eyes the primary DV was whether or not the list item probe ROI was the

[16] These exclusions were primarily due to the gaze criterion as search was 96% accurate throughout the entirety of the experiment across participants.

first ROI entered amongst all of the array items' ROIs. Each ROI covered a 7° x 7° area centered on each item in the array.

Figure 17 displays the free recall data alongside the list item probe first ROI engagement plotted by serial position. As can be seen in the recall data there was a marginal trend of primacy over the first five serial positions, $F(1,50) = 2.87$, $p = 0.09$, and a substantial trend of recency over the last five serial positions $F(1,50) = 19.00$, $p < 0.0001$. Crucially, the focal prediction of the experiment was borne out as there was a recency effect for eye movements. Logistic regression revealed that eye movements were more likely to enter the list probe ROI first as the serial position of the list item increased as indicated by a significant trend of WM probe recency, $\chi^2(1) = 8.843$, $p < 0.005$. Additionally, there was a significant within-subject correlation between item recall and list item saccadic engagement over the last four serial positions $r(9) = 0.55$, $p < 0.05$[17]. However, the omnibus test including target presence/absence as a factor indicated an unexpected main effect of target presence/absence, $\chi^2(1) = 4.299$, $p < 0.05$, in addition to the main effect of WM probe recency, $\chi^2(3) = 9.179$, $p < 0.05$, despite a non-significant interaction, $\chi^2(3) = 5.937$, $p = 0.115$, as depicted in Figure 18.

[17] Fisher's R to Z transformation was applied to each participant's Pearson correlation. The average of these transformed correlations was then transformed back to the Pearson metric and subjected to the inferential test reported above.

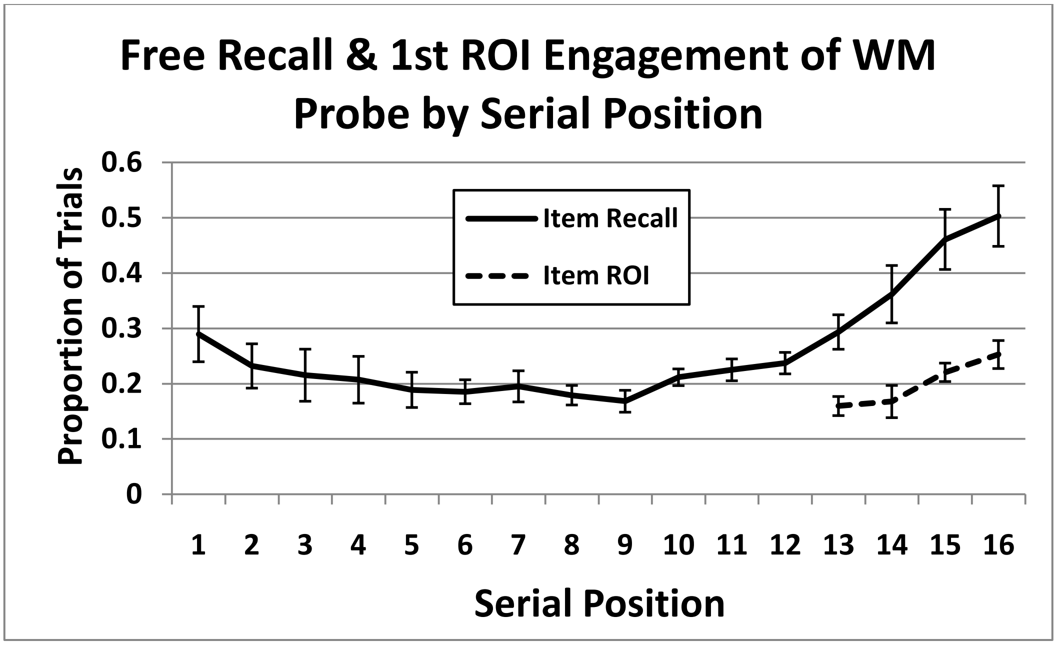


Figure 17: Free recall data plotted alongside the proportion of trials in which the WM probe was the first ROI entered within the search array in Experiment 4a

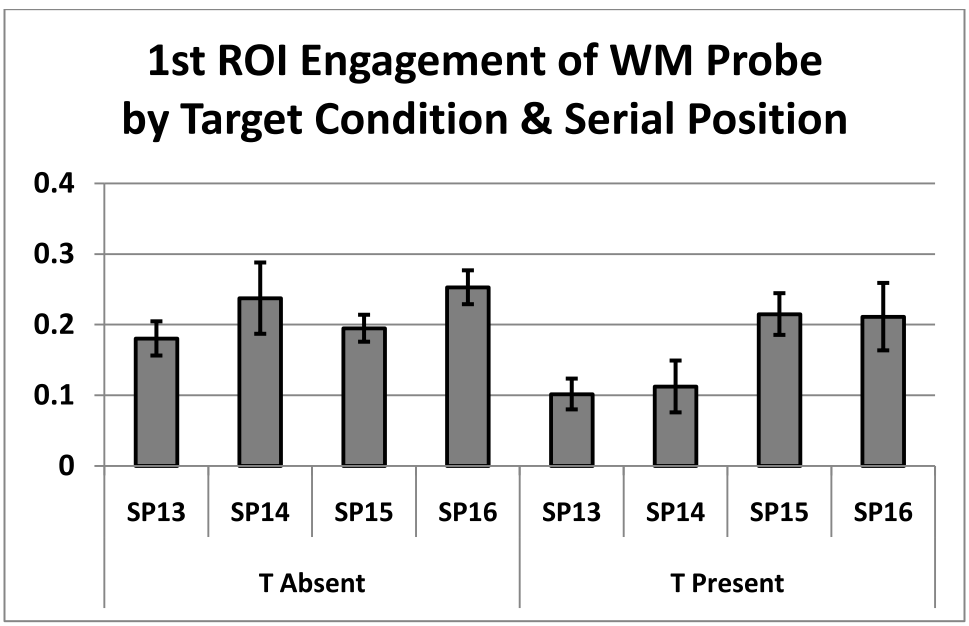


Figure 18: Proportion of trials in which the WM probe ROI was the first ROI entered within the search array plotted by target presence/absence and serial position of the WM probe in Experiment 4a

Discussion

The results revealed that participants were more likely to engage the WM probe as the serial position of the WM probe increased in the study list. In addition there was a significant within subject correlation between free recall of the last four items and engagement of the list probe. Together these results strongly suggest a link between the activation of individual items in working memory and biases in the execution of early oculomotor behavior. However, examination of the effect of the target presence/absence suggests that this relationship is not clear cut as there was a main effect of target presence/absence on the 1^{st} ROI engagement of the WM probe. Specifically, the effect of WM probe recency was pronounced in the target present condition, but did not manifest in the target absent condition. As no effect of target condition was anticipated, this finding did not accord with the expectations of the present experiment.

It should perhaps be noted, however, that it was hypothesized that the data in the target absent condition might suffer from less error as there would be less attentional competition in the search task and thereby facilitate the engagement of the WM probe. In contrast to this hypothesis, it was exactly the condition in which competition was promoted between the target and WM probe that allowed the recency effect to manifest. Furthermore, analysis of the rates of WM probe 1^{st} ROI engagement within the target present condition arose due to the low rates of engagement within the 13th & 14th serial positions as these rates were pushed down below chance level of 25%.

The fact that these items were engaged less than chance implicates a strategic usage of the list items during the search task. Past research in this area has demonstrated that people can use the contents of working memory in a flexible manner to strategically avoid the engagement of items matching the contents of working memory (Woodman & Luck, 2007). The results of the present experiment suggest that the participants learned that the list items would reappear in the search task and, given that the list probe would never be the target, the participants used this contingency to their advantage. That is, the participants adopted a strategy of not looking at the reappearing WM probes since engaging those items would only serve to slow search for the target. Interestingly and importantly, however, the data suggests that this strategy was only permitted to operate for items associated with low WM activation.

An important alternative explanation to the present data warrants discussion. It could be posited that a positive association between the recency of the WM probe and eye movements could be purely due to the fact that later items are more likely to occupy working memory at the time of search whereas earlier items would be more likely to have dropped out of WM by this time. This explanation suggests that the finding of an effect of WM probe recency would not implicate eye movement sensitivity to item-specific memory activation levels. That is, due to the working memory dynamics during list study, the present experiment might simply boil down to a partial replication of previously demonstrated effects (cf. Soto et. al. 2005). The manner in which the present effect manifested itself, however, suggests that this account of the data is unlikely. As mentioned above, it was due to the flexible and strategic use of the 13^{th} and 14^{th} items during search that allowed the differences along the recency gradient to emerge. Such strategic use of these earlier items would of course require that they reside in working memory during search to support the inhibition of attention towards the matching WM probes appearing in the arrays. As this is a difficult alternative account to counter under most conditions, future research employing related paradigms must guard against this alternative explanation of the data.

The present experiment demonstrates a clear link between an item's level of activation in working memory and the likelihood that it will be the first item engaged with the eyes in a visual search task. The intriguing manner in which this relationship manifested suggests, however, that the association between working memory activation and early oculomotor deployment is more complicated than anticipated. The strategic use of the list items was inadvertently incentivized in the present experiment. As the target would never coincide with the list items, flexibly inhibiting attention away from the reappearing list items was always to the benefit of the participant in determining the presence or absence of the search target. It will be important to discover the range of conditions and task characteristics under which this link between working memory activation and eye movements persists. Changing the search task to one of target discrimination (e.g., report the direction of slant of a vertical line), for instance, may greatly influence the way search is carried out. This is especially likely when the incentivization of the present experiment is removed by allowing the target to coincide with the spatial location of the WM probe. Much future investigation is needed to better understand this subtle and complex relationship.

Furthermore, several questions present themselves in moving forward to utilize early oculomotor behavior as an index of working memory activation in higher level decision making tasks. When deploying the MASS procedure during a hypothesis generation task, no explicit search task will be present. Rather, the arrays of items will flash on the screen with no reason provided to the participant. If the movement of the eyes towards WM matching items is automatic in such a context then such a procedure should succeed in measuring the active contents of working memory. If, however, the participant elects to use the contents of the hypothesis generation task in some strategic manner during the MASS trials then such measurement will be severely compromised. It is my opinion that the adoption of such strategies should not occur under such conditions as there should not be any incentive present to foster this. The present experiment, however, highlights the importance of considering the presence of such strategies when the MASS is deployed within higher level tasks.

Overall, however, the present results suggest that eye movements are sensitive to working memory activation and that the MASS procedure may afford sensitivity to this item-specific memory activation.

Experiment 4b: Attentional Blink Sensitivity to Working Memory Activation

An experiment nearly identical to Experiment 4a was carried out in order to assess the efficacy of attentional blink sensitivity for measuring item-specific memory activation. The list study phase and recall phases were exactly the same, but rather than a visual search task amongst an array of items appearing on the screen together, the visual search task was carried out over a series of items that appeared one after the other in a rapid serial visual presentation. This task is elaborated below.

Method

Participants

Nineteen participants from a large Midwestern university participated in this experiment for course credit.

Design & Procedure (3 RSVP Probe Position x 5 WM Probe Recency Within-Subjects)

In the rapid serial visual presentation search task eight pictures were presented to the participant in succession for 75 ms each with an inter stimulus interval of 45 ms between each item. Figure 19 depicts an example of the RSVP procedure as deployed in the experiment. Just as in the previous experiment, one of the items appearing in the RSVP stream was a reappearing item drawn from one of the last 4 positions in the study list. Additionally, there was a condition which was devoid of a WM probe in which none of the items in the RSVP were taken from the study list. This was included as a baseline condition that would allow the assessment of the upper bound of perceptual accuracy afforded in the search task in the absence of any attentional blink effects.

In Figure 19 the reappearing item is the first item on the sequence (i.e., the dolphin). The WM probe was manipulated to appear in one of the first 3 positions in the RSVP stream. This was done so that the participant wouldn't develop a clear expectation of observing the WM probe in any single position of the RSVP stream while at the same time positioning the WM probe near the beginning of the stream to allow it to tap the contents of WM[18]. The distance (i.e., time) between the WM probe and the search

[18] It is possible that the distractor items appearing in the RSVP stream may interfere with the contents of working memory as the RSVP unfolds. This conclusion was suggested by pilot work.

target was held constant across the entire experiment. The search target always appeared at a lag of 2 items following the reappearing WM probe as depicted in Figure 19. The target was randomly selected as one of the five shapes depicted in the figure (diamond, circle, start, square, or triangle). The participant's search task on every trial was to accurately report the identity of the target following the offset of the RSVP stream. Stickers with these shapes had been placed on 5 of the keyboard keys and the participants pressed these keys to indicate their search response.

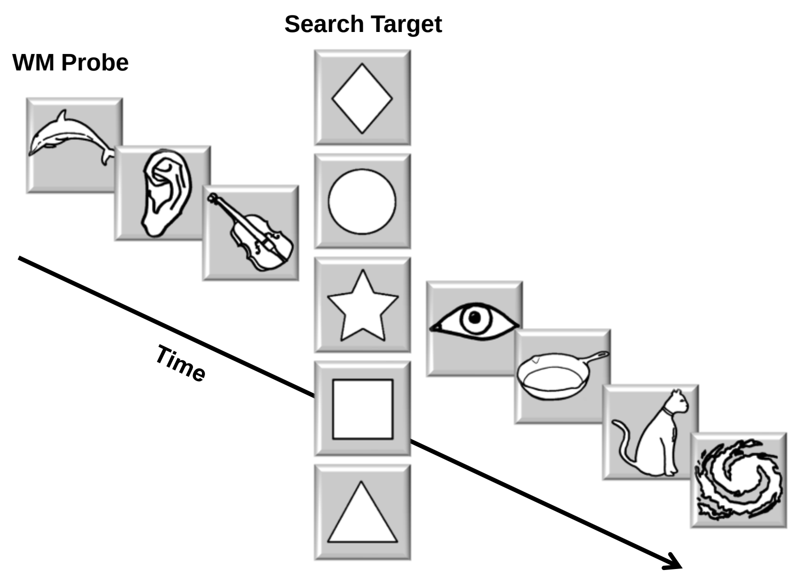


Figure 19: Example of RSVP stream used in Experiment 4b

Hypotheses & Predictions

The hypothesis for this experiment was that greater attentional blink would be observed with increasing recency of the WM probe. This would be evidenced by decreased accuracy in reporting the search target with increasing recency of the WM probe.

Results

Analysis of the free recall data, as displayed in Figure 20, demonstrated a substantial trend of primacy over the first five serial positions, $F(1,90) = 28.43$, $p < 0.0001$, and a substantial trend of recency over the last five serial positions $F(1,90) = 29.93$, $p < 0.0001$. Critically, the focal prediction of the experiment was confirmed. A significant trend of serial position for search accuracy was revealed, $\chi^2(1) = 5.44$, $p < 0.05$[19].

[19] Analyses of this trend, as well as those forthcoming, excluded the baseline condition. However, a MANOVA taking difference scores (subtracting each subject's search accuracy average by serial position from their baseline average) revealed a significant effect of WM probe recency as well, $F(3,16) = 3.28$, $p < 0.05$. Logistic regression excluding the baseline condition was used as the primary analysis as there were not enough observations per condition to calculate reliable difference scores once the data was broken out by RSVP position (2 observations per condition per subject).

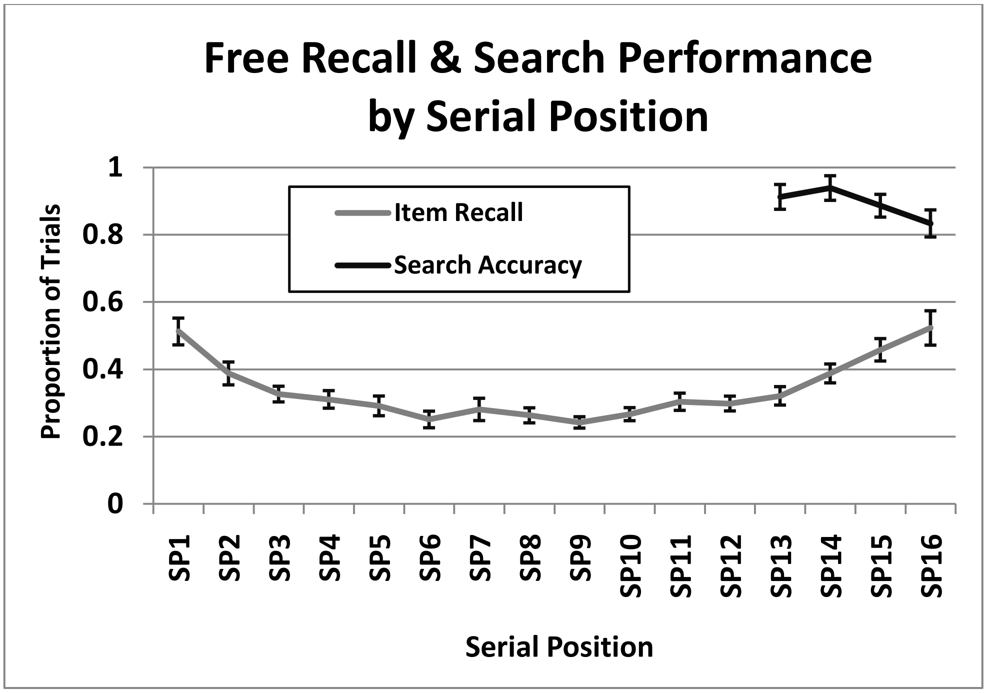


Figure 20: Free recall data plotted alongside the accuracy in the search task by serial position in Experiment 4b

Interestingly, however, when The RSVP position of the WM probe was included in the logistic model, this effect of serial position became marginal, $\chi^2(1) = 3.32$, $p = 0.06$. Furthermore there was a significant trend for the RSVP position variable itself, $\chi^2(1) = 5.36$, $p < 0.05$, as search accuracy was less accurate with the earlier positioning of the WM probe in the RSVP stream. This relationship is depicted in Figure 21. Additionally, the interaction between WM probe recency and RSVP position trends was not significant, $\chi^2(1) = 0.02$, $p = 0.898$.

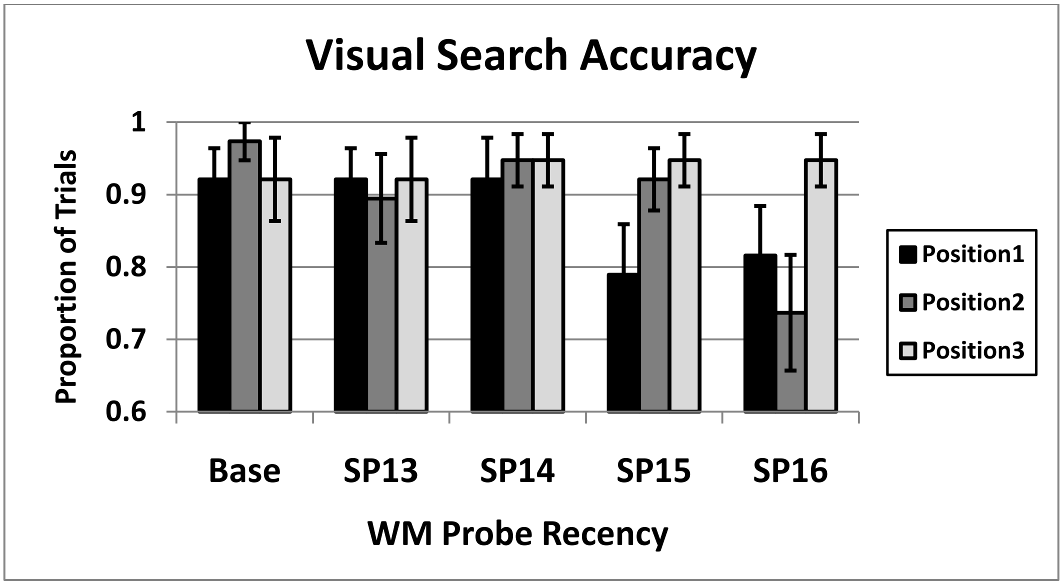


Figure 21: Visual search accuracy plotted by serial position of WM probe in study list and position of the WM probe in RSVP stream (Position1, 2, or3) in Experiment 4b

Discussion

As predicted, a trend of declining visual search accuracy in the RSVP search task obtained with increases in the list serial position from which the WM probe was drawn. At first blush this seems to suggest that the present experiment has been successful in translating the predicted differences in the working memory activations of the list items into differences in visual search performance. Closer inspection of the data, however, revealed that there is reason to approach this conclusion with some skepticism.

The unexpected effect of RSVP position demonstrated that as the WM probe appeared in later RSVP positions, its effectiveness at producing the attentional blink diminished. That is, when the probe appeared as the first RSVP item, attentional blink was observed over the 15^{th} and 16^{th} study list serial positions, but when the probe was the second RSVP item it was only able to elicit the blink for the 16^{th} item, and when the probe was in the 3^{rd} position it uniformly failed to elicit the attentional blink. This discrete patterning of the attentional blink deficit makes it difficult to argue that the measure is sensitive to working memory activation. It was my expectation that the differences in delay between the study list and the WM probe in the various RSVP conditions would be slight and thus not result in detectable differences in search performance. Additionally, I assumed that any interference to the contents of WM from RSVP items preceding the probe would be negligible. The particular pattern in the data, however, suggests that it is reasonable to assume that both factors may have played a role in the overall search performance results. Furthermore it can be inferred from Figure 21 that the smooth trend for WM probe serial position results from averaging over the step functions of visual search performance in RSVP position 1 & 2 conditions.

Given these considerations it appears that the attentional blink is likely of limited use for deployment in higher level cognitive tasks as a measure of working memory content. Some limitations were expected from the outset regarding the efficacy of informing fine grained working memory dynamics over time. For instance, given that the RSVP search task requires an overt response it could not be expected to be deployed multiple times within a single trial of a hypothesis generation task in the same manner potentially supported by the MASS procedure.

Despite the limitations of the MAS-AB procedure, however, it could still assist in the investigation of working memory dynamics through time. The data of the present experiment as well as Pashler and Shiu (1999) clearly demonstrate that the attentional blink provides a measure of whether an item is in or out of working memory. Although fine grained activation dynamics are the goal of my present and future investigations, being able to assess the maintenance vs. non-maintenance of items in working memory can substantially inform current theory. Lastly, there is one major advantage of the MAS-AB procedure over that of the MASS procedure. It doesn't require an eye tracker. Eye tracking is often cost and time prohibitive in comparison to experiments only requiring a standard PC. This advantage alone is reason enough to consider its use in future investigations of working memory dynamics over time in higher level decision making tasks.

Chapter 7: Deploying the MASS Procedure in a Hypothesis Generation Task

Two experiments were carried out in which the MASS procedure was deployed during a hypothesis generation task. In Experiment 5a the efficacy of the MASS procedure to detect the generation of hypotheses into working memory was assessed. That is, the first experiment served as a further validation of the MASS procedure in a higher level decision making task. The second experiment, Experiment 5b, deployed the MASS procedure in a synonymous hypothesis generation task, but was used to investigate a question of theoretical importance for understanding the temporal dynamics of hypothesis generation processes. The question underlying this experiment concerned how working memory resources are allocated throughout time during a hypothesis generation task.

Experiment 5a: Deploying the MASS Procedure - Hypothesis Generation

Method

Participants

Ninety one participants from a large Midwestern university participated in this experiment for course credit.

Design & Procedure (2 Data x 2 MASS Array Onset Position Within-Subjects)

The procedure of the experiment was conceptually very similar to Experiments 1, 2, & 3 in that it began with an exemplar training task, a test to verify learning, and ended with a hypothesis generation task. However, despite this conceptual overlap, the experiment was substantially different in appearance. It is assumed that the MASS procedure requires pictorial/object stimuli for its successful utilization. It is assumed that if word stimuli were used (as in Exps. 1, 2, & 3) that participants would not be able to resolve the details of the words to a strong enough degree to drive early oculomotor behavior. Therefore colored disks were used to represent the hypotheses and data in the present experiment so that during the MASS trials, the hypotheses and data would be clearly represented and quickly resolvable to peripheral vision.

Eleven colors were used in the present experiment whose RGB codes appear in Appendix B. For each participant each of these colors was randomly assigned to represent a hypothesis, a data level, or was assigned as a distractor in the search array and therefore had no meaning in the context of the exemplar training or hypothesis generation tasks. Table 5 presents the hypothesis by data ecology used in the present experiment. Unlike the representations of the ecologies of the previous experiments, this table displays the data levels for the negative/absent levels of the data in addition to the positive/present states. This is meant to underscore the fact that every possible data (and hypothesis) state in the present experiment was associated with a unique color.

		Symptoms			
		D1+	D1-	D2+	D2-
Diseases	H1	0.8	0.2	0.8	0.2
	H2	0.5	0.5	0.8	0.2
	H3	0.2	0.8	0.8	0.2

Table 5: Hypothesis x Data ecology of Experiments 5a & 5b

Changing the representations of the hypothesis and data states to colors of course required changes in how the exemplar training paradigm was used in the present experiment. Rather than couching the task as one of medical diagnosis, the task was described as a cause-effect learning task in which some colors would represent causes (i.e., hypotheses) and some colors would represent the effects of those causes (i.e., data)[20]. The instructions used in the present experiment are provided in Appendix C. The exemplars consisted of a hypothesis color and two data level colors. As displayed in figure 22, a cause was denoted by two arrows pointing away from it towards each of the effects appearing on that exemplar. The cardinal direction in which these associations were indicated (up, down,

[20] Although causes & effects are not semantically interchangeable with hypotheses & data this description afforded the essential knowledge structures assumed to underlie the hypothesis generation processes under investigation.

left, or right) as well as the spatial locations in which the items appeared on the screen (amongst twelve options) were randomly selected for each exemplar. This was done so that there would be no consistent spatial associations for any of the hypotheses or data in the experiment thereby obviating any spatial expectations for the hypotheses or the data during MASS elicitation.

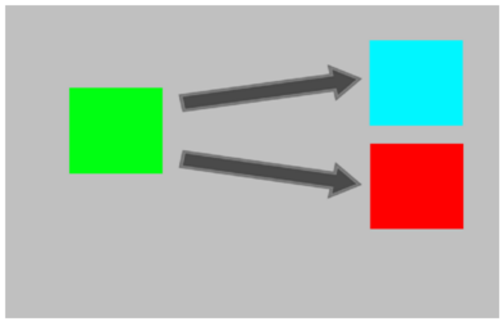


Figure 22: Example exemplar used in Experiment 5a

Following the instructions the exemplar training procedure commenced. The participants were presented with 150 exemplars (50 per hypothesis) at a rate of four seconds per exemplar. Following training a learning test was issued in which the participants were provided with each of the four data states and were asked to identify the most likely cause to generate each of the four effects. The eye tracker (Arrington Viewpoint) was then moved into place and set up for the participant who was then calibrated in the eye tracker.

The elicitation phase then ensued. Each trial in the elicitation phase was carried out as depicted in Figure 23. The participants were presented with a piece of data, a MASS array, a hypothesis generation prompt, an overt hypothesis generation task in which a key press was elicited, and a probability judgment task. The independent variables were the piece of data presented at the beginning of each trial and the timing of the MASS array onset within the elicitation sequence. The presented data was manipulated to be either D1+ or D1- and, as indicated in Figure 23, the MASS array was manipulated to appear in one of two possible positions in the elicitation sequence. Note that these locations on the timeline bracket the prompting of hypothesis generation. For this reason the condition in which the MASS array appears in the first position is referred to as the PrePrompt condition and the condition in which the MASS array appears following the hypothesis generation prompt is referred to as the PostPrompt condition. Each participant was presented with four elicitation trials in which these two conditions (data & array onset position) were factorially manipulated.

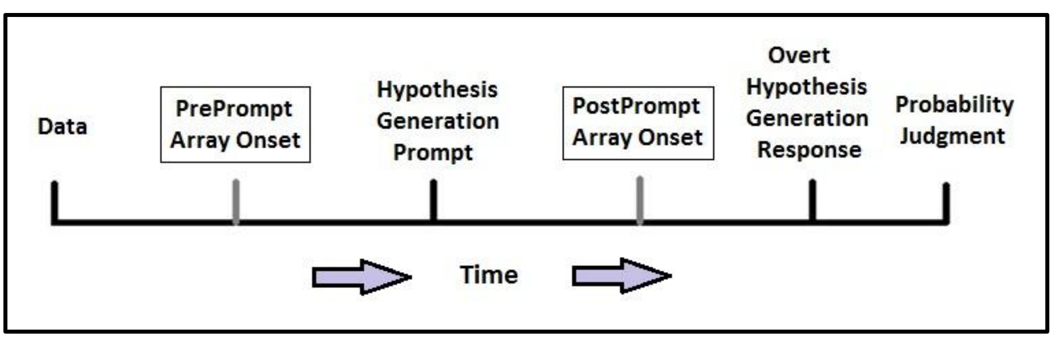


Figure 23: Time course of elicitation in Experiments 5a & 5b

The MASS arrays appeared in star shaped pattern and contained the same five elements, randomly assigned to one of the five positions, on each trial[21]. These five items were H1, H2, H3, and two distractor colors that had not appeared in the experiment prior to the MASS arrays. Figure 24 displays an example of the MASS arrays used in this experiment. Directly prior to the onset of the MASS array a fixation cross (i.e., +) quickly flashed in the center of the screen to draw the participant's eyes to the center of the screen before the MASS array appeared.

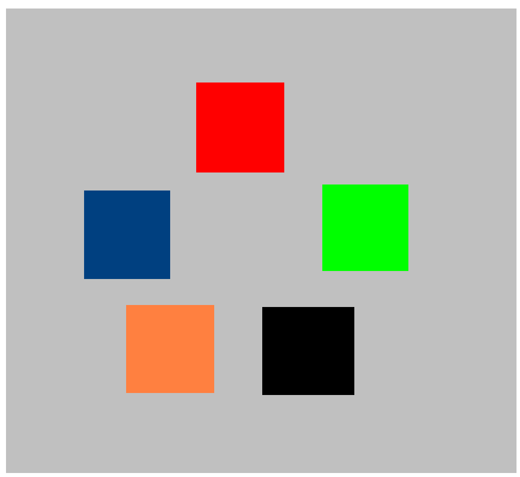


Figure 24: Example MASS array used in Experiment 5a

Each trial finished with an overt selection of the most likely hypothesis in which the participant responded with a key press and a probability judgment concerning the hypothesis they considered most

[21] The center of each array item was equally distant from the center of the screen and the distance between the center of each item and its neighboring items was equal as well. That is, each item was equally spaced around an imaginary circle centered on the middle of the display.

likely[22]. As in the previous experiments a series of ten arithmetic verification trials was carried out as a distractor task between each trial.

Hypotheses & Predictions

The present experiment was a test of the ability of the MASS procedure to detect hypothesis generation. The hypothesis was that more initial eye movements would be directed towards the most likely hypothesis in the PostPrompt condition as compared to the PrePrompt condition.

Results

Prior to analysis the same calibration criteria implemented in Experiment 4a were applied. Fifty-eight the ninety-one participants met this criteria and only data from these participants was subjected to analysis. An additional exclusion criterion was applied for each trial. A trial was discarded if the participant's gaze was not contained in a central 3.8° x 3.8° region of interest (ROI) upon the onset of the MASS array. The application of this criterion excluded 43 trials, leaving 189 trials subject to analysis.

The data was collapsed over the data conditions (D1+, D1-) in order to assess the effect of the MASS array onset position. As D1+ and D1-

[22] The probability judgment prompt was similar to that used in Experiments 1 and 2. The prompt read as follows: "If you viewed 100 cases with this effect, how many would have the cause you selected on the previous screen?"

were associated with H1 and H3 to the same extent, these conditions were collapsed and analyzed with respect to the high probability and low probability hypotheses given the data that had been presented on that trial. For instance, if D1+ was presented then the high probability hypothesis was H1, but if D1- had been presented then the high probability hypothesis was H3. H2 was always the medium probability hypothesis as it was related to both states of the data with a probability of 0.5.

Figure 25 demonstrates the rate at which each item's ROI was entered first following the onset of the MASS arrays broken out by PrePrompt and PostPrompt conditions[23]. As can be seen there was no effect of Prompt Condition for the engagement of any of the items in the array. The high probability hypothesis was no more likely to be engaged following the generation prompt than it was prior to the prompt, $\chi^2(1) = 0.425$, $p = 0.514$.

[23] In this figure, and those that follow for this experiment, the rates of distractor engagement are the rates for looking at an individual distractor in the array. These rates were obtained by dividing the rate of looking at either distractor by two as two distractors appeared in each array.

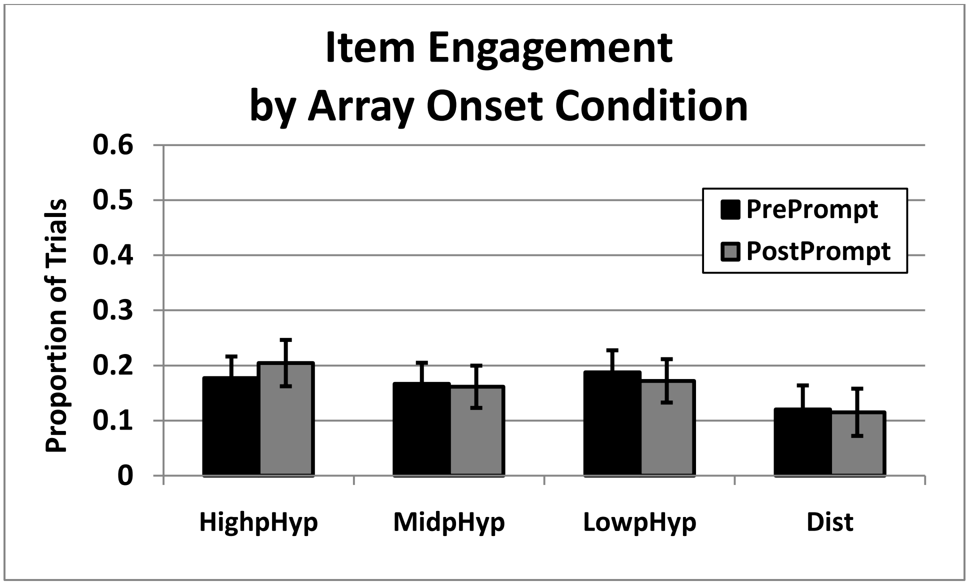


Figure 25: Item engagement by Array Onset condition and Array Item type in Experiment 5a

Further inspection of the data, however, revealed a large bias in the eye movement data. Figure 26 plots the proportion of trials on which each ROI was the first ROI entered within the MASS arrays. ROI1 was the ROI appearing on the right hand side of the star array and the remaining ROIs were numbered consecutively in a clockwise manner. As can be seen, ROI 5 appearing in the upper middle of the star-shaped array was favored more than twice as much as any other ROI.

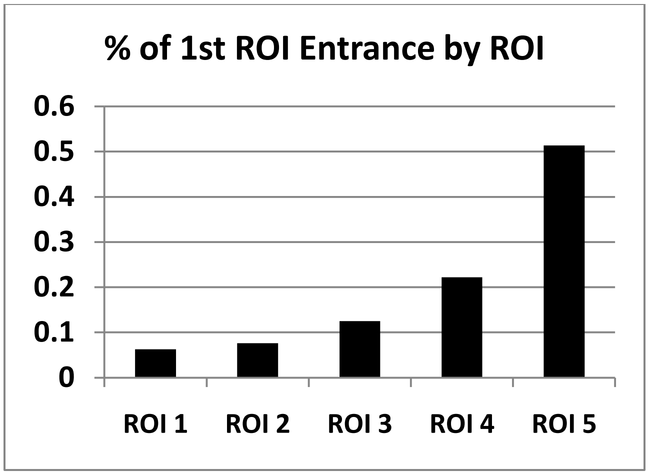


Figure 26: 1st ROI entrance rates over all trials by ROI in Experiment 5a

Figure 27 plots the rate at which each item's ROI was entered first following the onset of the MASS arrays by array onset condition when the trials in which ROI 5 was entered first are removed from the data set. Although the effect of array onset location remains non significant, $\chi^2(1) = 1.47$, $p = 0.225$, this conditionalization does move the mean rates in the predicted direction as a 13% differences emerges between the PrePrompt and PostPrompt conditions for the high probability hypothesis. Further conditionalizing on only the participants demonstrating good learning (in the learning test prior to the MASS trials) further magnifies this effect as demonstrated in Figure 28[24]. When conditionalized on this subset of participants a significant result obtains, $\chi^2(1) = 5.419$, $p < 0.05$, as there were significantly more eye movements directed towards the high

[24] The criterion for good learning was whether or not the participant was able to accurately identify the hypothesis that was most likely given D1+ and D1- in the learning check.

probability hypothesis in the PostPrompt condition as compared to the PrePrompt condition[25].

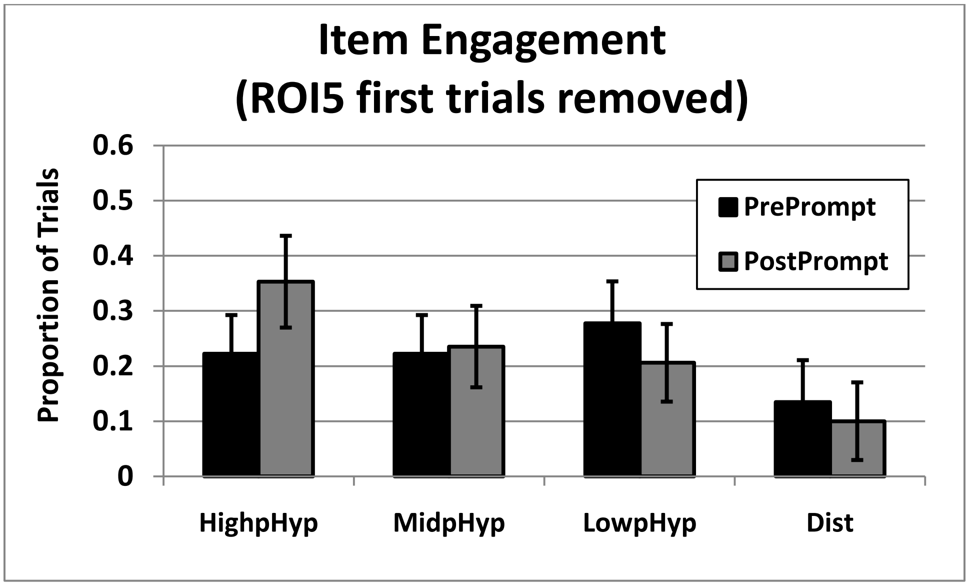


Figure 27: Item engagement by Array Onset condition and Array Item type in the absence of trials in which ROI 5 was the first ROI engaged in Experiment 5a

[25] Analysis of the participants not meeting the learning criterion revealed no effect of Prompt onset position, $\chi^2(1) = 2.76$, $p = 0.097$. Although this could be considered a marginal effect, it is important to note that the means are in the opposite direction as those for the learners as they were 0.5 in the PrePrompt condition and 0.125 in the PostPrompt condition.

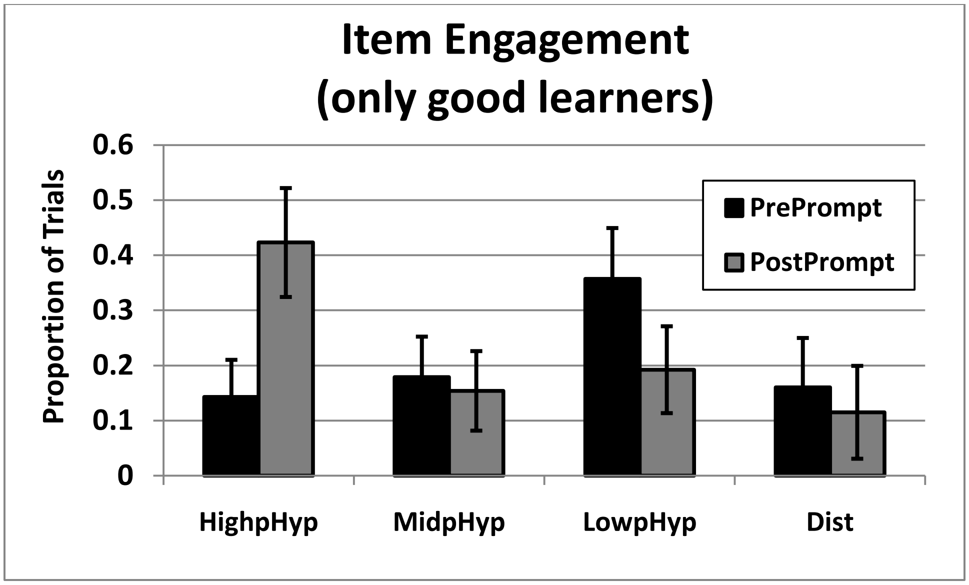


Figure 28: Item engagement by Array Onset condition and Array Item type for only good learners in the absence of trials in which ROI 5 was the first ROI engaged in Experiment 5a

Discussion

The present experiment sought to provide an initial low-bar test of the ability of the MASS procedure to detect the movement of information into working memory. The results demonstrate that the MASS procedure successfully passed this test. As predicted, more initial eye movements were directed towards the high probability hypothesis following the generation prompt.

It is important to realize that this procedure is very much in its infancy. The two conditionalizations performed on the data in the present experiment suggest important ways to improve the deployment of the MASS in future experiments. Firstly, the fact that the top-middle ROI was substantially favored suggests a potential limitation with the arrays used in the present experiment. Future deployments of the MASS procedure would likely benefit from the deployment of different array types. A potential solution would be to use randomly selected square and diamond arrays as was done in Experiment 4a. This solution does however come with a cost in that the number of items competing with one another in the array is then limited to four rather than five as in the present experiment.

The fact that the significant results obtained only when poor learners were excluded from the data set suggests that improvements to the learning in the exemplar training task could be beneficial to the use of the MASS procedure in this context. The exemplars used in the present experiment featured one cause pointing towards two effects. This could be simplified by showing the cause-effect relationships as dyads rather than triads. Such an adjustment may enhance the learning taking place on each exemplar. Additional remedies improving the learning taking place during the exemplar training phase should also be considered.

The overall sensitivity of the measure may appear weak in light of the results of the present experiment. However, it is important to note that the problems with the implementation of the present experiment may have been substantial and yet the measure was still able to capture the movement of information into working memory. The eventual level of sensitivity that may be afforded by the MASS procedure is not at all clear from the present observations. Future implementations of this procedure should utilize the above remedies in pursuit of enhancing the sensitivity and efficacy of the measure as well as the general paradigm.

Experiment 5b: Deploying the MASS Procedure - Informational Trade-offs in WM

The present experiment utilized the MASS procedure to assess differences in working memory contents prior to and following hypothesis generation. As such this experiment serves as an initial investigation of informational tradeoffs occurring in working memory allocation over the course of a hypothesis generation task. In addition to replicating the previous experiment, the present experiment sought to determine if the data probe utilized in the initial stages of retrieval is still maintained following the generation of hypotheses.

Method

Participants

Twenty nine participants from a large Midwestern university participated in this experiment for course credit.

Design & Procedure (2 Data x 2 MASS Array Onset Position x 4 MASS Array Content Competitions Within-Subjects)

The procedure of the experiment was nearly synonymous with that of Experiment 5a with a few notable exceptions[26]. The exemplar training phase was simplified by displaying cause-effect dyads rather than cause-effect triads. The display of the dyads was slightly modified from the exemplar displays in the previous experiment. Whereas the triads could appear in one of twelve spatial locations in Experiment 5a, the dyads were constrained to appear in the middle of the screen (pointing either vertically or horizontally). The orientation of the dyads was randomized for each exemplar so as to preclude spatial expectations for the data or hypotheses. Figure 29 displays the orientations of the cause-effect dyads.

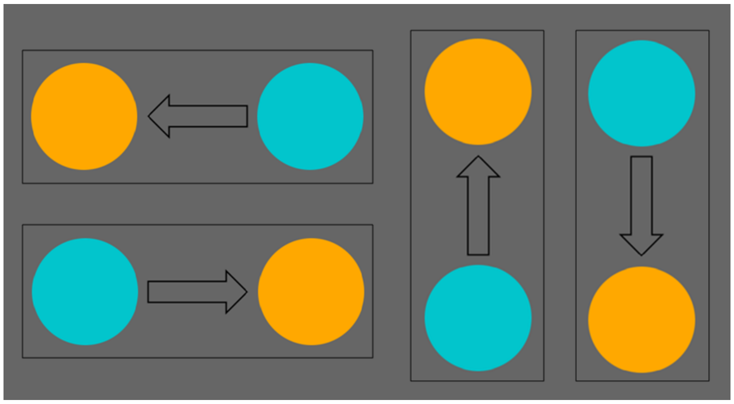


Figure 29: Example exemplars from Experiment 5b

[26] Slightly different shades of color were used in this experiment in an effort to better control for luminance differences between the colors. The RGB codes of these colors appear in Appendix D.

Since dyads were used, the exemplars were presented at a faster rate than in Experiment 5a (two seconds, as opposed to four seconds, per exemplar). Three hundred exemplars (100 per hypothesis) were presented in the exemplar training phase. The same Hypothesis x Data ecology used in the previous experiment was used in the present experiment (see Table 5) and the same learning test followed exemplar training. Directly after the learning test the eye tracking equipment was set up and the participant was calibrated in the eye tracker (SR Research Eye Link 1000).

The elicitation phase then began. The elicitation sequence on each trail was exactly the same as that used in the previous experiment (see Figure 23). The data presented to the participant at the beginning of each trial, either D1+ or D1-, and the MASS array onset positions, either PrePrompt or PostPrompt, were the same as in the previous experiment. Whereas the contents populating the MASS arrays were always the same in the previous experiment, the MASS array contents were manipulated in the present experiment to examine how different competitions amongst the array items influenced eye movement behavior. The MASS array content competition conditions are displayed in Table 6 below. The present MASS arrays only contained 4 items (as opposed to the 5 item arrays in Experiment 5a) and appeared randomly in either square or diamond configurations. These measures were taken in an effort to obviate the previously observed first ROI bias. Examples of the MASS arrays utilized in the present experiment appear in Figure 30. Each participant completed one trial in each condition. Accordingly, each participant completed 16 elicitation trials with an arithmetic distracter task intervening between each trial.

	Array Contents			
	1	**2**	**3**	**4**
Condition 1	H1	D2+	Dist	Dist
Condition 2	H1	D1+	Dist	Dist
Condition 3	D1+	D2+	Dist	Dist
Condition 4	H1	H2	H3	D1+

Table 6: MASS array content competition conditions of Experiment 5b

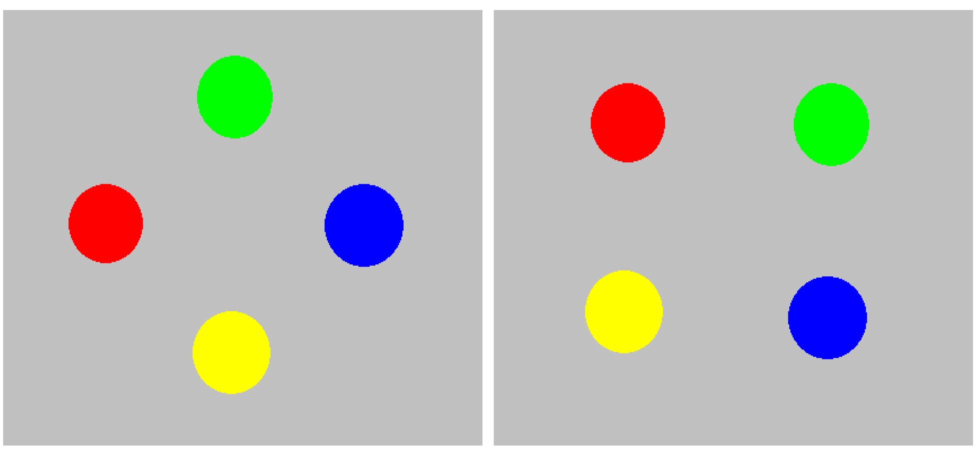


Figure 30: Example MASS arrays used in Experiment 5b

As previously discussed, the work of Moores et al. (2003) demonstrated that first saccades can be drawn to items that are merely semantically related to the content of working memory. This represents a potential problem for the use of visual-search methodologies with respect to paradigms utilizing “micro-worlds” defined directly by semantic relations between data and hypotheses. In order to use visual attention to assess working memory activation in this context we must be able to parse out the contribution of semantic relatedness to the overall attentional performance observed.

The Competition conditions of the present experiment were designed with this consideration in mind. Condition 1 provided a baseline competition in order to define the rate of attention drawn to semantically related data in the search array. As D2+ was not been presented in the elicitation sequence prior to its appearance in the search array and it maintained an equal semantic relation to H1 as D1+ (which will have been presented at the beginning of the trial). This can be seen in the Hypothesis x Data ecology of the present experiment where the conditional probabilities of D1+ & D2+ are both 0.8 under H1.

Condition 2 establishes a direct competition between the presented data and the most likely hypothesis (within the D1+ data condition). Item specific activation of D1+ will be assessed by the extent to which positive deviation from the baseline established in condition 1 is observed. Condition 3 then sets up a competition between semantic relation and presented data in the absence of competition from the likely generated hypothesis (within the D1+ data condition). Lastly, condition 4 provides a relatively global competition between all of the possible hypotheses and D1+. This competition thus allows replication of Experiment 5a.

Hypotheses & Predictions

The general prediction is that the observed data will be expunged from working memory following hypothesis generation as working memory resources will be reallocated to the maintenance of generated hypotheses at that time. This prediction will be assessed by comparisons between array content conditions 1 and 2 within the PostPrompt and D1+ data conditions. Additionally it was hypothesized that the results from Experiment 5a would be replicated within array content condition 4 in which all 3 hypotheses appeared in the MASS array.

Results

Of initial interest is that the effect of MASS array onset position observed in Experiment 5a was replicated within the Competition 4 condition. As demonstrated in Figure 31 significantly more initial eye movements[27] were directed towards the High Probability Hypothesis

[27] Due to differences in the data structure of the Eye Link eye tracker data the present analyses examined the first ROI fixated on each trial. Although this differs slightly from the dependent variable used in Experiments 4a & 5a (i.e., first ROI entered regardless of the registration of fixation), it is assumed that it is highly unlikely that the two measures would diverge considerably as most ROI entered are assumed have been fixated in the

following the hypothesis generation prompt, $\chi^2(1) = 5.52$, $p < 0.05$. Note that this result obtains in the present experiment without recourse to conditionalization of the dataset. Examination of the distribution of first ROI fixated (Figure 32) suggests that the modification to the MASS arrays deployed in the present experiment attenuated, but did not fully alleviate, the "up-left bias" observed in Experiment 5a.

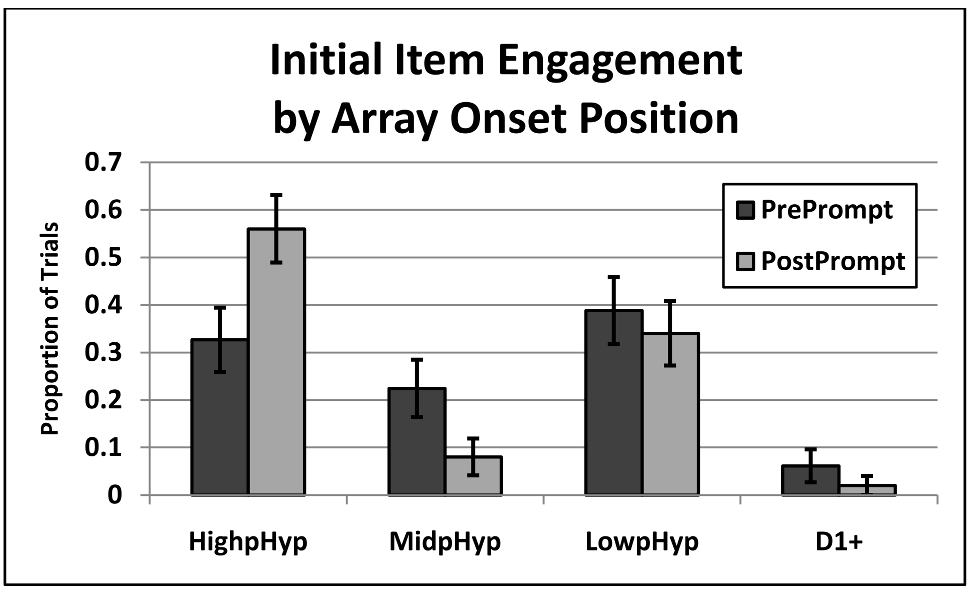


Figure 31: Replication of Experiment 5a. Initial item engagement by array onset condition and array items within competition condition 4 in Experiment 5b.

previous experiments. Due to the sparse arrays utilized in these experiments it would be a rare event for gaze to traverse an ROI in the absence of fixation.

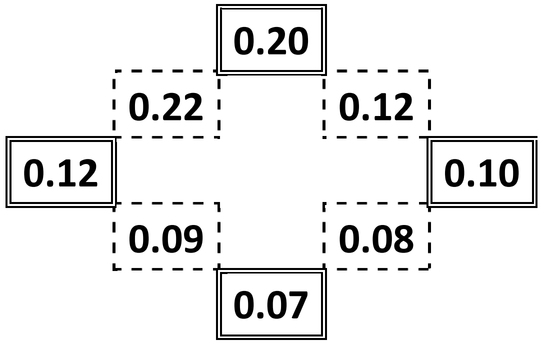


Figure 32: Distribution of 1st fixations within each ROI in Experiment 5b. Dashed boxes = square array ROIs. Double bound boxes = diamond arrays.

In order to investigate the question of whether or not the presented data was maintained following hypothesis generation comparisons between Competition conditions 1 and 2 within the PostPrompt condition were necessary. The first fixation rate of D2+ in competition condition 1 was compared to the first fixation rate of D1+ in competition condition 2 within the D1+ Data condition as displayed in Figure 33. This comparison revealed no difference between these rates, $\chi^2(1) = 0.617$, $p = 0.432$. However, in the absence of competition from H1 in Competition condition 3, D1+ was clearly favored over D2+ as demonstrated in Figure 34. Binomial tests determined that the rate of D1+ engagement significantly differed from chance, z = 3.58, $p < 0.001$, in addition to the rate at which D2+ was engaged, z = 6.77, $p < 0.001$. The rate at which D2+ was engaged did not differ from chance, z = -1.5, $p = 0.096$. These results suggest that the D1+ data was maintained following generation. This same comparison is important to examine in the D1- data condition as this pattern should disappear. As displayed in Figure 35, however, the same pattern manifests as D1+ is engaged above chance, z = 2.45, $p < 0.05$, and above the level of D2+, z = 5.288, $p < 0.001$, which is not engaged above chance, z = -1.36, $p = 0.14$.

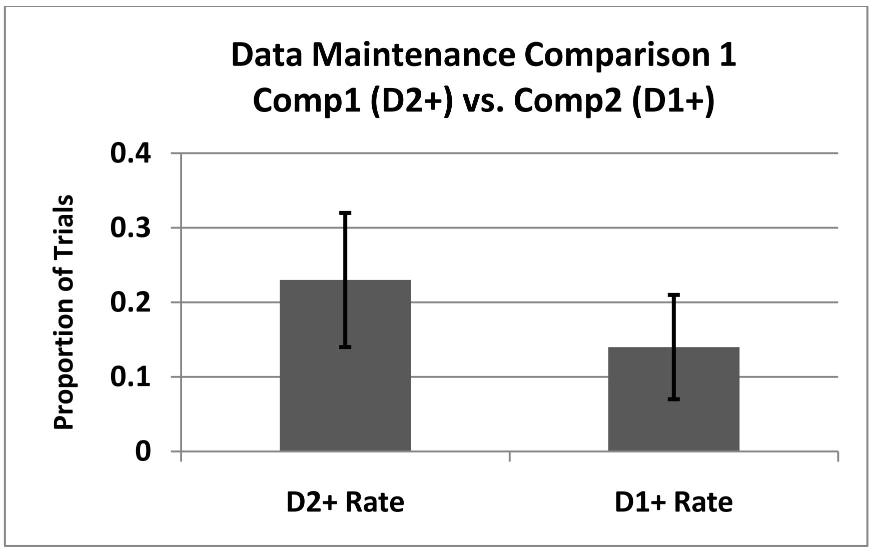


Figure 33: Comparison between rates of first fixation on D2+ in Competition 1 vs. rate of first fixation on D1+ in Competition 2 (within D1+ Data condition) in Experiment 5b

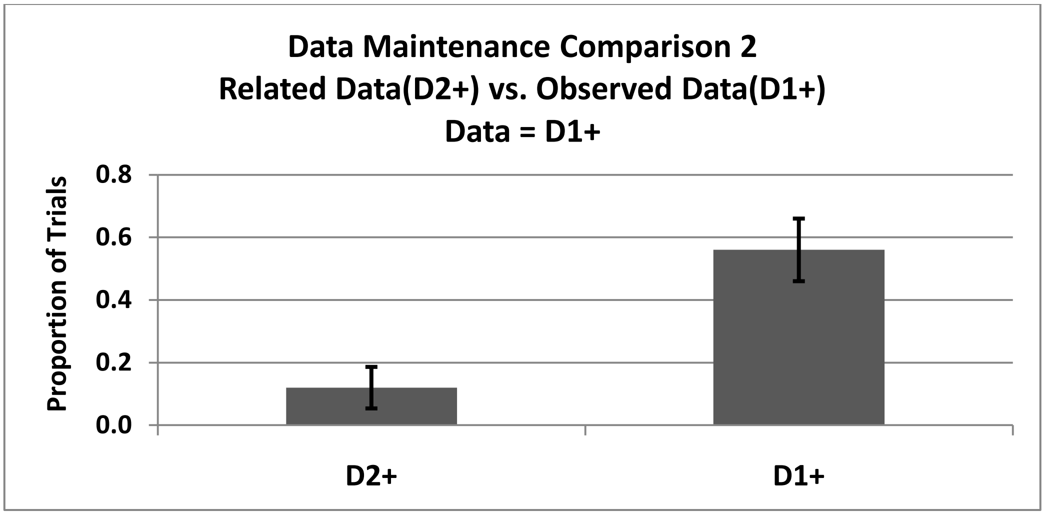


Figure 34: Proportion of trials on which the first item fixated was the observed data (D1+) or data with equal semantic relation to the highest probability hypothesis following hypothesis generation in Experiment 5b

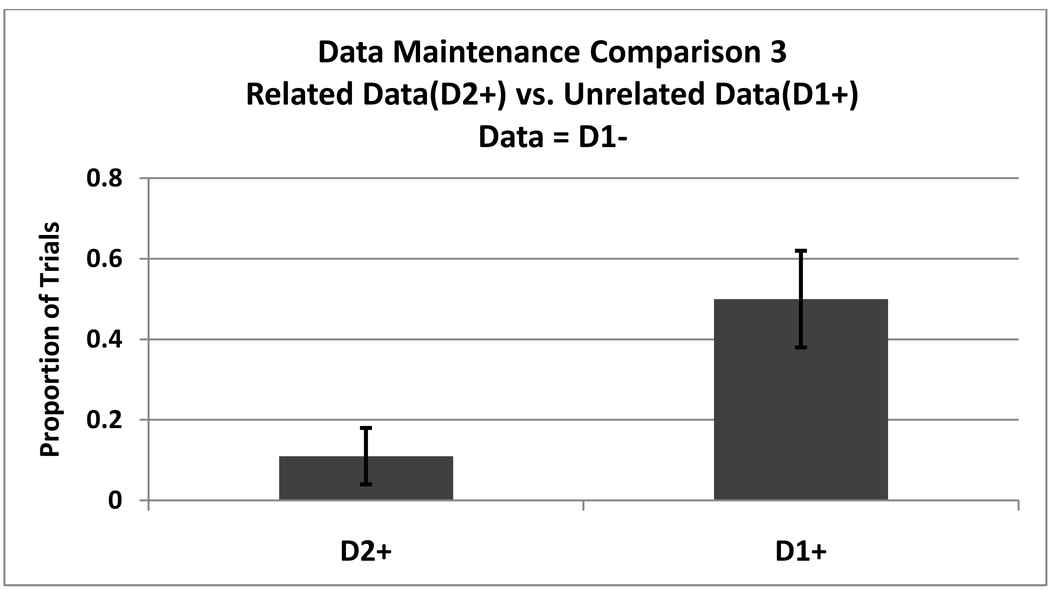


Figure 35: Proportion of trials on which the first item fixated was the observed data (D1+) or data with equal semantic relation to the highest probability hypothesis following hypothesis generation in Experiment 5b

Further comparisons examining D1+ first fixation rate were carried out between PrePrompt and PostPrompt conditions within Competition 2 & 3 and D1+ Data conditions. As can be seen in Figure 34, neither of these comparisons revealed a significant result, $\chi^2(1) = 0.272$, $p = 0.602$ and $\chi^2(1) = 1.973$, $p = 0.16$ respectively.

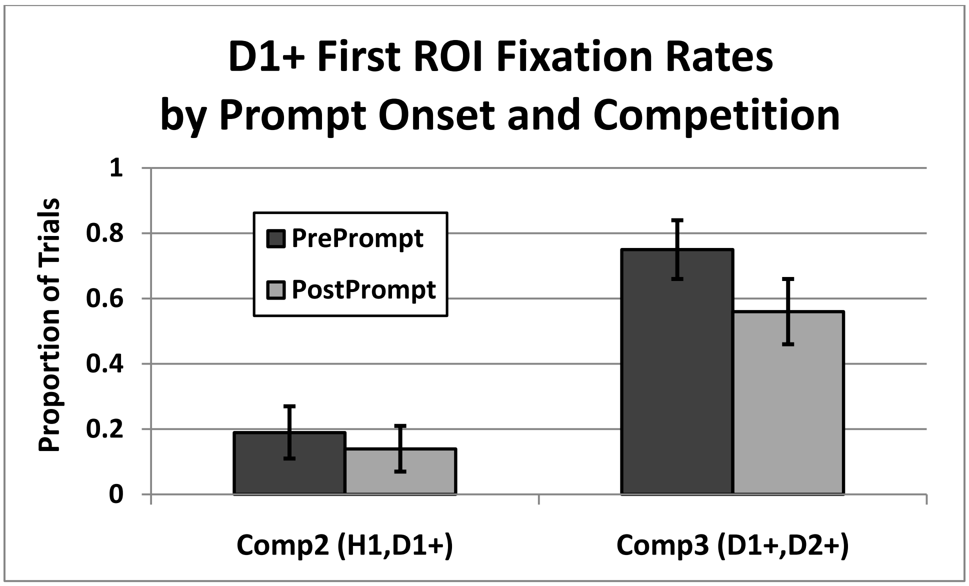


Figure 36: Comparisons between Prompt Onset conditions within two competition conditions for the first ROI fixation rates of D1+ in Experiment 5b

<u>Discussion</u>

There were two goals for the present experiment. The first was to replicate the effect observed in Experiment 5a. This effect, in which early oculomotor behavior more often engaged the high probability hypothesis following the hypothesis generation prompt, was replicated in the present experiment. Furthermore, this effect was replicated in the absence of the conditionalizations utilized in the previous experiment. This successful replication bolsters the validity and utility of the paradigm.

The second goal of this investigation was to ascertain how working memory resources are allocated over time within the hypothesis generation task. Specifically the paradigm addressed whether or not observed data, used as a retrieval cue to generate hypotheses from LTM, would continue to reside in working memory following generation or if it would be purged from working memory in the process of working memory reallocation. The constellation of evidence from the D1+ Data PostPrompt onset conditions suggests that the data was maintained following generation, but that it was not as active in working memory as the H1. It was anticipated that the comparison between D2+ engagement in competition 1 and D1+ engagement in competition 2 would be the crucial comparison for determining if D1+ remained resident in WM following generation. This comparison did not reveal a difference and as a result does not provide any evidence for continued maintenance. The reason for this, however, is likely due to the fact that H1 simply overshadowed the data as it was overwhelmingly favored in these competition conditions 1 and 2 (first ROI fixation rates of 0.45 and 0.73 respectively). Competition 3 allowed for a direct competition between D1+ and D2+ which was equally semantically related to H1. In this condition, in the absence of competition from H1, a clear preference for D1+ over D2+ emerged. This preference indicates that D1+ remained resident in WM following hypothesis generation. However, this conclusion must be approached with caution at the current time as the same pattern was exhibited in the D1- Data condition which should have demonstrated a slight reversal or at least the disappearance of the effect. This mirroring of results between data conditions suggests that the measure is likely too noisy at this time (or with the present amount of data) to make claims about data residing in working memory following hypothesis generation.

The overshadowing effect of H1 in competitions 1 & 2 can perhaps be understood as an offshoot of an effect observed in Moores et al. (2003). In their Experiment 5 it was found that people's initial eye movements would be guided towards array items semantically related to the content of WM, but that this was only the case when the actual item in working memory did not reappear in the array. In such cases when the actual item (e.g., motorcycle) appeared in the array along with the semantic associate (e.g., motorcycle helmet) the engagement rate of the associate was no different than distracters. Although this situation is not entirely synonymous with that of the present experiment, it does demonstrate a similar overshadowing effect. Additional research examining the attentional mechanisms underlying this measurement technique may help untangle precisely how the overshadowing effect in the present experiment operates. A paradigm like that of Experiment 4a in which multiple list items reappear in the search array would be useful in this regard.

Further array content conditions could be useful to future research on data maintenance and use throughout the hypothesis generation process. As a complement to the present conditions it would be useful to have arrays in which the data or hypotheses appear in isolation (as accompanied by only distracter items). These comparisons would be useful as there would not be any competition exerted from other task relevant items in the array and therefore might be more sensitive and more informative in some cases.

First and foremost, the results of Experiments 5a & 5b demonstrate the efficacy of the MASS procedure for the online assessment of the contents of working memory throughout a memory and decision making task. Greater refinement is needed to distill the procedure to its most useful form with improvements to the exemplar training task perhaps being the most important. The results of Experiment 5b suggest that data may have been maintained following generation. Further data collection and experimentation is necessary to determine if this is the case. It is important to note however that even if convincing evidence is found that data are maintained following generation, this will not suggest that data are always maintained following generation, but only that in some cases it can be. The ecologies of these experiments were relatively simple and designed so that the data would suggest only one hypothesis. It is likely that when data implicate a host of hypotheses (and working memory resources are more greatly stained by the population of likely hypotheses into working memory) that data maintenance will be forgone. Additionally, given that data are maintained following generation under some conditions it will be interesting to determine if the data are allowed to consume WM resource during the judgment and decisional tasks of hypothesis generation and information search. Future research utilizing the MASS procedure can address such questions.

Chapter 8: General Discussion

Summary of Results

The primary goal of this work was to provide constraints on the theoretical assumptions to be made in the development of a computational model of hypothesis generation honoring temporal dynamics. Each experiment targeted a specific theoretical question concerning the dynamics of data acquisition and hypothesis generation processes. The constellation of results from these experiments has revealed predicted regularities as well as unexpected behaviors. In this way the present work has been successful in that it will inform the construction of forthcoming computational models of dynamic hypothesis generation.

Experiment 1 investigated a relationship between order effects in cue presentation and their contribution to the cueing of long-term memory in hypothesis generation. Specifically, this experiment tested how differences in the serial order placement of a diagnostic cue affected the successful retrieval of the most likely hypothesis implicated by this, the only, diagnostic cue. It was found that as the serial position of the diagnostic cue increased people were more likely to generate the most likely hypothesis. This finding is in agreement with Sprenger (2007) in suggesting that people more heavily weight recent cues in the retrieval process. It remains to be seen, however, how exactly this pattern of data is being produced. It could be that more recent cues are simply more likely to reside in working memory at the time of the generation prompt and that less recent cues are more likely to have fallen out of working memory by that time. Alternatively, as hypothesized at the outset of Experiment 1, it could be the case that the activation of each cue governed its contribution to the retrieval process. Under this explanation more recent cues are expected to be more active at the time of generation and thus should influence the generation process to a greater degree than earlier cues. The present data of this particular experiment do not allow us to parse these two explanations, but future research can readily address this issue both empirically and computationally.

Experiment 2 tested the same broad hypothesis underlying Experiment 1, that the activation of particular data in working memory at the time of generation would govern their contributions to the hypothesis generation process. The order of the data in this experiment, however, remained constant. The manipulation within this experiment was the presentation rate of the individual data within the sequence. In one set of conditions the rate of presentation was relatively fast (300 ms.) and in another condition the rate was comparatively slower (1200 ms.). As predicted by the Context Activation Model of memory, it was hypothesized that the fast rate would cause the data acquired early in the sequence to reside at higher levels of activation than data acquired later in the sequence when prompted to generate hypotheses. At the slow rate, however, the inverse was expected. Clear evidence was found for this proposition in the slow condition. The results from the fast conditions, although not as unambiguous, supported this hypothesis as well, as the favoring of the later hypothesis inverted in one case as predicted and disappeared in the remaining conditions.

Together the results from Experiments 1 & 2 provide converging support for the proposition that the levels of working memory activation associated with individual pieces of data govern their contributions to the retrieval process underlying hypothesis generation. This provides valuable insight into how the initial retrieval operation of the forthcoming computational model should make use of the various data comprising the compound cue to long-term memory. Once the model is developed it should be easy to test the equal weight model against the proportional activation weighted model. The expectation is that the activation weighted model should provide the best account of the data.

Experiment 3 sought to test two consequences of the hypothesis generation process being carried out over time. In some cases a decision maker may acquire multiple data in quick succession and use this pool of data to initiate hypothesis generation while in other cases data may be acquired in isolation with successive rounds of hypothesis generation occurring between the acquisitions of each datum. Furthermore, as data are acquired over time, they may support or rebut the hypotheses currently under consideration. The results of Experiment 3 indicate that when people have generated a set of hypotheses from a single datum and acquire data inconsistent with a portion of the current set, this portion of the maintained hypotheses is purged from working memory. However, it was expected that inconsistent data would be further used to recue memory in order to repopulate working memory with likely hypotheses. The results of this experiment do not provide evidence for this type of LTM recuing. Interestingly, when the participants generated hypotheses from the pool of data (rather than one at a time) a primacy effect was observed wherein the later data were underweighted relative to the earlier data. This observation runs contrary to the expectation based on the results of Experiments 1 & 2 and Sprenger (2007). This finding in particular poses an intriguing challenge to the forthcoming models of dynamic hypothesis generation as they will need to account for the crucial difference(s) between these experiments. This surprising result is likely to be important to our eventual understanding of the dynamic mechanisms of the model.

A secondary goal of this dissertation was to develop novel non-invasive methodologies capable of assessing the activations of the contents of working memory through time. Two such methodologies were explored in pursuit of this goal, both of which are based on the premise that relatively automatic biases in visual search might foster inference of the active contents of working memory. The first methodology, dubbed the Memory Activation Sensitive Saccade procedure, utilizes early oculomotor behavior in an implicit visual search task as an index of working memory activation.

The second method, dubbed the Memory Activation Sensitive Attentional Blink procedure, utilizes search deficits in an explicit search through a rapid serial visual presentation to infer the contents of working memory. Experiments were carried out in an effort to investigate the level of sensitivity afforded by each technique. As the MASS procedure currently appears to be the more sensitive, flexible, and informative methodology it was selected for initial deployment within a hypothesis generation task.

Experiment 5a demonstrated the successful deployment of the MASS procedure as it successfully detected the generation of a hypothesis into working memory. Experiment 5b went on to use the MASS procedure to examine a question of theoretical interest to the temporal dynamics of hypothesis generation. As little is known about the allocation of working memory throughout a hypothesis generation task, this experiment investigated how working memory resources are allocated following hypothesis generation. The specific question addressed was whether or not the data used as the retrieval cue remains resident in working memory following generation or if working memory resources are reallocated causing the data to be purged from working memory. Conclusions regarding this question remain unclear at this time. The planned analyses revealed a difference indicative of the maintenance of the data following generation, but an additional comparison that should have demonstrated the reverse pattern resulted in the same pattern as that of the planned comparison. This is unfortunate as it appears that either the measure itself or the data collected are simply too noisy at this time to provide a definitive conclusion on this matter.

Practical Guidance

The present results clearly illustrate that people are not invariant to differences in the temporal characteristics of the information they receive. Although it is difficult at this time to pinpoint the contexts under which early or late data may be over weighted, a general awareness of this fact may be of use to decision makers in applied settings. For instance, a physician may reflect on the hypotheses they generated and consider if demonstrate stronger agreement with earlier or later data. This would potentially allow the physician to engage in successive rounds of hypothesis generation in which cues having been identified as underweighted could be more strongly weighted in an effort to generate additional hypotheses. Alternatively the physician could simply try out different orders of cues or engage in generation in response to each cue without the need for reflection following the first round. HyGene asserts that generating more hypotheses will generally be of benefit to probability judgment and information search behavior. Therefore utilizing varying orders of the available data to generate a greater set of potential hypotheses will likely be of benefit to higher level decision making.

Building a Model & Accounting for the Present Data

The present results provide a suitable set of effects for a dynamic model to be built upon and account for. A hybrid model combining the subtle memory activation dynamics of the Context Activation Model (CAM) with the semantic structure and working memory operations (e.g., probability judgment and information search functions) inherent in HyGene should provide a rich architecture capable of accounting for the present findings as well as making a host of testable predictions to guide future research. The development of this hybrid model will be challenging as several issues will need to be addressed during this endeavor.

Two potential hybrid models suggest themselves from their predecessors, each with their own unique challenges for implementation. One possibility would be to import the semantic structure and working memory mechanisms of HyGene into the framework of CAM. Alternatively, it would be possible to import the working memory activation dynamics of CAM into the existing HyGene framework. As CAM and HyGene utilize distinct representations, the focal distinguishing feature that arises between the proposed models is the representation upon which each model would operate. This fundamental difference between the models is important due to the fact that the different representations are endowed with different affordances lending themselves more readily towards some capabilities over others. Likely challenges to the implementation of each of these models are discussed below.

The representation implemented in CAM consists of individual localistic units which together form a matrix of connections between a lexical-semantic system, in which each item is represented individually, and an episodic contextual signal (implemented as an asymmetric random walk) defining the temporal context at any given point in time. Although this representation is well suited for capturing the list recall dynamics that the model was designed to account for, further elaboration will be needed in order to address the cued recall and episodic partitioning assumed by HyGene to underlie hypothesis generation and higher level decision making processes.

Although CAM has been applied to a cued recall task (Davelaar, 2005), this implementation only accounted for cued recall of items resident in working memory during the acquisition of the cue. That is, the cue could only serve to boost the activation of semantically related items already in working memory. This implementation, therefore, doesn't capture the cued retrieval processes inherent in hypothesis generation tasks in which people must generate hypotheses from long-term memory. Addressing the cued recall of hypotheses from LTM will require further elaboration of CAM's representation. Although CAM's representation does honor semantic relatedness, it is likely that the semantic structure is too coarse to capture semantic recall as operative in HyGene. HyGene assumes that a subset of episodic memory is activated in response to a cue and that this subset determines semantic retrieval. This subsetting operation allows sensitivity to graded semantic relations between the contents of memory and the present retrieval cue. Semantic relatedness in CAM, on the other hand, is represented by direct links between units that pass activation when they become active. Furthermore, it is the utilization of the global match and multiple trace assumptions that foster the subsetting function within HyGene. It is presently unclear how such processes should be implemented in an alternate framework that does not share both of these properties.

The alternative model would utilize HyGene as the base and incorporate the activation dynamics of CAM into its working memory construct. HyGene is essentially an extension of the MINERVA II (Hintzman, 1986, 1988) and MINERVA-DM (Dougherty, 1999) frameworks. Therefore HyGene's representation consists of a storehouse of exemplars defined by vectors of individual features. In this representation each new experience with an item gets encoded as a new trace in memory.

At present HyGene receives all available data simultaneously and weights their individual contributions to the hypothesis generation process. These assumptions are of course in need of amendment in order to account for the present effects. The incorporation of CAM's dynamic working memory processes into HyGene's working memory construct should provide the necessary components to achieve this goal. This model would assume that data undergo dynamic fluctuation in working memory as governed by several of CAM's processes such as self recurrent activation, lateral inhibition from other data as a result of competition, and activation decay. The current activation levels of each piece of data can then be used as weights applied to each piece of data in the hypothesis generation process at any point in time. This will require two sets of information to reside within this model's working memory construct as the set of relevant data (RED set) currently active in working memory will co-occupy the space with the set of leading contender hypotheses (SOC).

The challenging aspect of this implementation will be how best to incorporate CAM's activation dynamics into the representation of HyGene. One potential way would assume a separate vector of data weights that updates in accordance with the dynamic processes. The sum of this vector would be assumed to be less than or equal to one in the absence of hypotheses and substantially less when sharing WM resources with generated hypotheses. When only one piece of data is active in working memory it will be weighted strongly (≈ 1) in the generation process, but when other pieces of data enter WM, the activation weight of the first data will necessarily decrease as the total pool of activation will be shared amongst the all data in WM.

It would be interesting to develop both of these hybrid models in order to test their accounts of the present data against one another as well as assess their abilities to account for the extant phenomena for which MINERVA-DM and HyGene were developed to capture. An interesting tradeoff between the modeling accounts may develop here. Many of the predictions of the current work were developed with respect to the operation of CAM in its current form. Therefore it is likely that the first model discussed above should be able to account for the present effects quite well, but may encounter difficulty in capturing the extant phenomena in hypothesis generation, probability judgment, and hypothesis testing. This will provide the true test of the utility of this representation for the purposes of developing a cumulative model honoring temporal dynamics. Conversely, the second model discussed above is not likely to encounter problems accounting for the extant data, but will be challenged by the present effects. Therefore it will be its ability to account for the present effects that will provide a test of its utility for the present purposes. In the end, this competition is likely to boil down to which modeling architecture and representation lends itself more naturally to the assimilation of the companion model's processes and computations.

Snapshots of Working Memory

The secondary goal of the present research was to develop non-invasive methodologies for measuring the active content of working memory. The idea behind these measures is that by exploiting automaticities in visual search, exerted when stimuli in the environment matches the content of working memory, we may be able to infer the contents of the mind by observing attentional performance. Two measures were explored; the MASS procedure in which early oculomotor behavior was the DV of interest and the MAS-AB procedure in which a deficit in attentional performance was the DV.

It was hoped that these measures would benefit from four characteristics that would enhance their usefulness over existing techniques for exploring the movement of information through working memory over time. First, the techniques were hoped to be non-invasive in that they would obviate significant amounts of perturbation to the task processing at hand and the current contents of working memory. Second, the measures were designed to be relatively deployable on-line, within and throughout a task. Such a property would thus allow the observation of changes in information usage throughout an individual trial. Third, the measures are item specific in that they allow measurement at the level of individual items (e.g., data or hypotheses) rather than simply the engagement of generalized processing as is measured in much of modern neuroscience (e.g., fMRI, EEG, MEG). Lastly, the measures would be of superior usefulness if they were sensitive not only to the current processing of individual items, but to their current levels of activation in memory.

Overall the measures encompass these characteristics relatively well. Both the MASS and MAS-AB procedure were designed to be item specific and relatively deployable on-line within a trial. However, the MASS procedure seems to supersede the MAS-AB procedure in the remaining two characteristics. The results of Experiments 4a & 4b suggest that early oculomotor behavior, but not attentional blink performance, is sensitive to the item-specific activation levels in working memory. Furthermore, whereas the MAS-AB procedure requires an overt response (i.e., button press), the MASS procedure does not. This allows the MASS procedure to be deployed relatively quickly and interfere minimally with the primary task at hand. For these reasons it is likely that the MASS procedure will be more useful in addressing fine grained working memory dynamics. As highlighted by the somewhat ambiguous results of Experiment 5b, however, this procedure still stands to benefit from further refinement in order to live up to this promise. Lastly, although the MAS-AB procedure is not as well suited for the present investigations in some respects, it should still be viewed as a valuable tool. The MAS-AB procedure is sensitive to the inclusion or exclusion of items in working memory and does not require state-of-the-art (i.e., prohibitively expensive) equipment for its implementation. As a result, the MAS-AB procedure may be able to inform several pertinent theoretical questions more quickly than the MASS.

As highlighted by the discussion of the characteristics of these measures above, it is clear that these measures provide cognitive researchers with a new type of measure than has previously existed. Although these measures were designed with exploration of dynamic hypothesis generation in mind, there is nothing inherent in the methodologies limiting them to this task domain. Rather, these measures are domain general and as a result can support inquiry within a wide array of tasks within psychology and cognitive science more generally.

References

Anderson, N.H. (1965). Primacy effects in personality impression formation using a generalized order effect paradigm. *Journal of Personality and Social Psychology, 2*(1), 1-9.

Anderson, N.H. (1973). Serial position curves in impression formation. *Journal of Experimental Psychology, 97*(1), 8-12.

Anzai, Y. (1984) Cognitive control of real -time event driven systems. *Cognitive Science*, *8*, 221-254.

Brehmer, B. & Allard, R. (1991). Real-time dynamic decision making: Effects of task complexity and feedback delays. In J. Rasmussen, B. Brehmer, and J. Leplat
(Eds.) *Distributed decision making; Cognitive models for cooperative work*. Chichester: Wiley.

Brown, G.D.A., Neath, I., Chater, N. (2007). A Temporal Ratio Model of Memory. *Psychological Review, (114)*3, 539-576.

Busemeyer, J.R. (1999). Dynamic Decision Making. In *Encyclopedia of the Social and Behavioral Sciences: Methodology, Mathematics, and Computer Science.* Amsterdam: Pergamon.

Davelaar, E. J., Goshen-Gottstein,Y, & Ashkenazi, A. (2005). The demise of short term memory revisited: Empirical and computational investigations of recency effects. *Psychological Review*, *112*(1), 3-42.

Dougherty, M.R.P., Gettys, C.F., & Ogden, E.E. (1999). A memory processes model for judgments of likelihood. *Psychological Review, 106*(1), 180-209.

Dougherty, M.R.P., & Hunter, J.E. (2003)a. Probability judgment and subadditivity: The role of WMC and constraining retrieval. *Memory & Cognition*, *31*(6), 968-982.

Dougherty, M.R.P., & Hunter, J.E. (2003)b. Hypothesis generation, probability judgment, and working memory capacity. *Acta Psychologica*, *113*(3), 263-282.

Dougherty, M. R. P., Thomas, R. P., & Lange, N. (2010). Toward an integrative theory of hypothesis generation, probability judgment, and information search. In *The Psychology of Learning and Motivation* (Vol. 52, pp. 299-342). Elsevier.

Downing, P.E. (2000). Interactions between visual working memory and selective attention. *Psychological Science. 11*(6), 467-473.

Dux, P.E. & Marois, R. (2009). The attentional blink: A review of data and theory. *Attention, Perception, & Psychophysics, 71*(8), 1683-1700.

Fisher (1987). Cue selection in hypothesis generation: Reading habits, consistency checking, and diagnostic scanning. *Organizational Behavior and Human Decision Processes, 40*, 170-192.

Fisher, S.D., Gettys, C.F., Manning, C., Mehle, T. & Baca, S. (1983). Consistency checking in hypothesis generation. *Organizational behavior and human performance, 31*, 233-254.

Gillund, G. & Shiffrin,R.M. (1984). A retrieval model for both recognition and recall. *Psychological Review, 91*(1), 1-67.

Gonzalez, C., Thomas, R.P., Vanyukov, P. (2005). The relationship between cognitive ability and dynamic decision making. *Intelligence, 33,*169-186.

Hintzman, D.L. (1986). “Schema Abstraction” in a multiple-trace memory model. *Psychological Review, 93,*411-428.

Hintzman, D.L. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review, 95*, 528-551.

Hogarth, R.M. & Einhorn, H.J. (1992). Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology, 24*(1), 1-55.

Howard, M.W. & Kahana, M.J. (2001). A distributed representation of temporal context. *Journal of Mathematical Psychology, 46*, 269-299.

Huang, L. & Pashler, H. (2008). Working memory and the guidance of visual attention: Consonance driven orienting. *Psychonomic Bulletin & Review,14*(1), 148-153.

Jagacinski, R. J. & Miller, R. A. (1978) Describing the human operator’s internal model of a dynamic system. *Human Factors, 20,* 425-433.

Kerstholt, J.H. (1994). The effect of time pressure on decision making in a dynamic task environment. *Acta Psychologica, 86,* 89-104.

Kleinmuntz, D. & Thomas, J. (1987). The value of action and inference in dynamic decision making. *Organizational Behavior and Human Decision Processes, 39,* 341-364.

Moores, E., Laiti, L., & Chelazzi, L. (2003). Associative knowledge controls deployment of visual selective attention. *Nature Neuroscience, 6*(2), 182-189.
Murdock, B.B. (1962) The serial position effect of free recall. *Journal of Experimental psychology, 64*(5), 482-488.

Page, M.P.A. & Norris, D. (1998). The primacy model: A new model of immediate serial recall. *Psychological Review, 105*(4), 761-781.

Pashler, H. & Shui, L. (1999). Do images involuntarily trigger search? A test of Pillsbury's hypothesis. *Psychonomic Bulletin & Review, 6*(3), 445-448.

Raajmakers, J.G.W. & Shiffrin, R.M. (1981) Search of Associative Memory. *Psychological Review, 88*(2), 93-134.

Smith, H.D., Stasson, M.F., & Hawkes, W.G. (1998/1999). Dilution in Legal Decision Making: Effect of Non-Diagnostic Information in Relation to Amount of Diagnostic Evidence. *Current Psychology: Developmental, Learning, Personality, Social, 17*(4), 333-345.

Soto, D., Heinke, D., Humphreys, G.W., & Blanco, M. (2005). Early, involuntary top-down guidance of attention from working memory. *Journal of Experimental Psychology: Human Perception and Performance, 31*(2)248-261.

Soto, D. & D., Humphreys (2007). Automatic guidance of attention from verbal working memory. *Journal of Experimental Psychology: Human Perception and Performance,33*(3),730-757.

Sprenger, A.M. (2007). Sequential Hypothesis Generation. *Dissertation Abstracts International: Section B: The Sciences and Engineering, 68*(4-B), 2683.

Sprenger, A. & Dougherty, M.R.P. (2006). Differences between probability and frequency judgments: The role of individual differences in working memory capacity. *Organizational Behavior and Human Decision Processes, 99*(2), 202-211.

Sprenger, A.M., Dougherty, M.R., Atkins, S.M.,Franco-Watkins, A.M., Thomas, R.P.,Lange, N.D., & Abbs, B. (Resubmitted) Implications of cognitive load for hypothesis generation and probability judgment. *Frontiers in Cognitive Science.*

Thomas,R.P, Dougherty, M.R., Sprenger, A.M., & Harbison, J.I. (2008). Diagnostic hypothesis generation and human judgment. *Psychological Review, 115*(1), 155-185.

Toda, M. (1962). The design of a fungus-eater: A model of human behavior in an unsophisticated environment. *Behavioral Science, 7*(2), 164-183.

Troutman, C.M., & Shanteau, J. (1977). Inferences based on nondiagnostic information. *Organizational Behavior and Human Performance, 19,* 43-55.

Weiss, D.J. & Anderson, N.H. (1969). Subjective averaging of length with serial position. *Journal of Experimental Psychology, 82,* 52-63.

Woodman, G.F. & Luck, S.J. (2007). Do the contents of working memory automatically influence attentional selection during visual search? *Journal of Experimental Psychology: Human Perception and Performance, 3*(22), 363-377.

Zukier, H. (1982). The Dilution Effect: The role of correlation and the dispersion of predictor variables in the use of nondiagnostic information. *Journal of Personality and Social Psychology, 43,* 1163-1174.

Appendix A: Examples of the displays used in Experiment 1

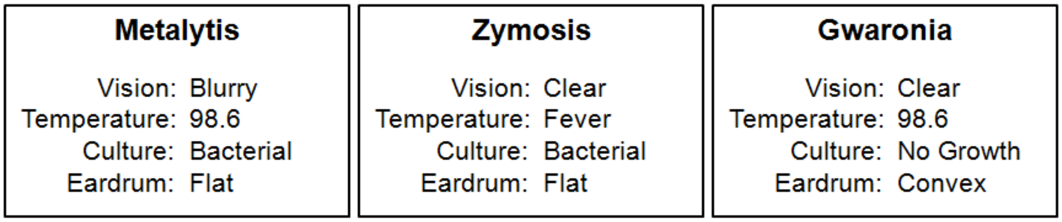


Figure A 1: Examples of exemplar training displays in Experiment 1

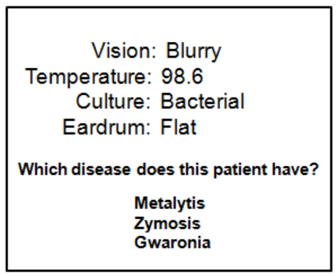


Figure A 2: Example of Learning test display in Experiment 1

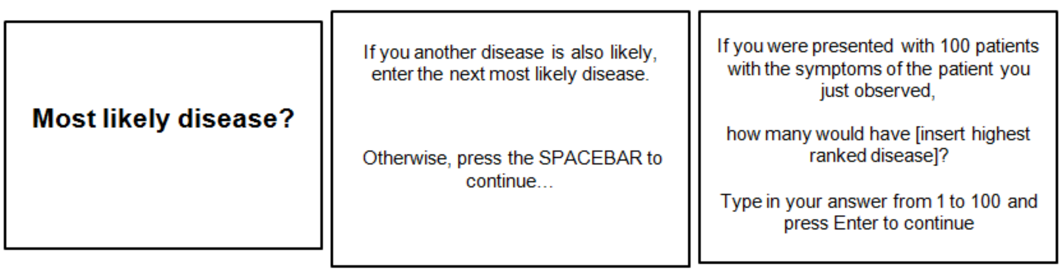


Figure A 3: Examples of Elicitation displays. From left to right: hypothesis generation elicitation screen for most likely disease, hypothesis generation screen for 2nd and 3rd most likely diseases, and probability judgment screen in Experiment 1

Appendix B: Colors Used in Experiment 5a

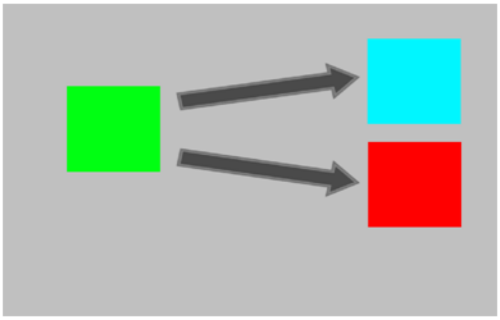
Color	R	G	B
Red	255	0	0
Sky Blue	0	246	255
Forest Green	24	86	60
Green	0	255	18
Brown	111	63	0
Orange	255	168	0
Purple	154	69	234
Blue	12	0	255
Yellow	255	252	0
White	255	255	255
Black	0	0	0

Table B 1: RGB codes of colors used in Experiment 5a

Appendix C: Instructions for Experiments 5a & 5b

In the first part of this session your task is to learn the strength with which various causes are related to various effects. Think about the following as an example of what you will be learning: When medical students learn about various diseases through their experiences with patients, they observe many patients and learn how their symptoms are related to the disease they are suffering from. In this example you can think of a disease as a CAUSE of certain symptoms which you can think of as EFFECTS. When a medical student is learning what symptoms (EFFECTS) are result from which diseases (CAUSES) they must learn the strength of the relationship between each CAUSE and each EFFECT because they do not always appear together. For instance, when someone has the Flu, they will usually have a fever, but not always. Also, there are other illnesses that produce a fever, so a fever doesn't necessarily indicate that the patient has the Flu. Therefore it is important to learn the STRENGTH OF ASSOCIATION between various CAUSES (diseases) and their EFFECTS (symptoms) in order to diagnose patients accurately. This is what you will be learning in the first phase of this session. Your task is to learn the STRENGTH OF ASSOCIATION between CAUSES and EFFECTS so that you will be able to indicate which CAUSE was responsible for EFFECTS that you will observe.

There is, however, one difference between the task of the medical student in the example and the one that you will complete for your learning in this session. Whereas the medical student has the task of learning CAUSE and EFFECT relationships with verbal labels (for example: Flu-Fever), you will be learning CAUSE and EFFECT relationships visually. Specifically, you will learn the relationships between various colors representing EFFECTS and other colors representing CAUSES of those effects.



This figure is an example of the displays that you will be learning these relationships from. The top bar will always represent the CAUSE and the bottom 2 bars will always represent EFFECTS of that cause (Each cause and each effect will have independent colors and each color will always represent the same cause and effect throughout the experiment). You will be presented with many displays like this one in order to learn the STRENGTH OF ASSOCIATION between the various CAUSES and EFFECTS.

Now think back to the example of the medical student learning by observing several patients. This is the same thing you will be doing here. If we were to conceptualize this display in terms of the earlier medical example, the top bar (the CAUSE) would represent the Flu and the bottom 2 bars (the EFFECTS) would represent fever and another symptom like nausea. Over the course of being presented with many displays like this one you will LEARN HOW STRONGLY EACH CAUSE IS ASSOCIATED WITH EACH EFFECT.

You will now begin the training phase in which you will be presented with many of these displays so you can learn the STRENGTH OF ASSOCIATION between each CAUSE (top bar) and each EFFECT (bottom bars). If you have any questions about the learning task please ask the experimenter before beginning.

Appendix D: Colors Used in Experiment 5b

Color	R	G	B
Red	231	30	0
Sky Blue	2	197	204
Forest Green	27	98	27
Green	0	210	13
Brown	180	90	0
Orange	255	122	0
Purple	154	69	234
Blue	0	100	200
Yellow	180	180	0
White	232	232	232

Table D 1: RGB codes of colors used in Experiment 5b