ABSTRACT

Title of Dissertation: SEQUENTIAL HYPOTHESIS GENERATION

Amber Marie Lehman Sprenger,
Doctor of Philosophy, 2007

Dissertation Directed By: Professor Michael R. Dougherty
Department of Psychology

This paper examined how decision makers generate and evaluate hypotheses when data are presented sequentially. Hypothesis generation occurs in many judgment and decision making tasks, but no research has yet examined the underlying processes of hypothesis generation when data occur sequentially. In a series of three experiments, participants learned the relationship between data and possible causes of the data in a virtual environment. Data were then presented iteratively and participants either generated hypotheses they thought caused the data or rated the probability of possible causes of the data. In a fourth experiment, participants generated hypotheses and made probability judgments based on previously-stored general knowledge. The four experiments examined whether different orders of data led decision makers to consider different sets of hypotheses. Findings revealed that participants weighted data presented later in a sequence more heavily than data presented early in a sequence when responding after each datum was presented. Future experimental directions are detailed and potential assumptions necessary for a model to account for sequential hypothesis generation behavior are discussed.

SEQUENTIAL HYPOTHESIS GENERATION

By

Amber Marie Lehman Sprenger

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2007

Advisory Committee:
Professor Michael R. Dougherty, Chair
Professor Bradley D. Hatfield
Professor Carl W. Lejuez
Professor Kent L. Norman
Professor Thomas S. Wallsten

Acknowledgements

I thank Michael Dougherty for his guidance and support throughout my graduate student career. I also thank Bradley Hatfield, Carl Lejuez, Kent Norman, and Thomas Wallsten for reading and providing thoughtful feedback on my dissertation.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

A fundamental component of human decision making is generating diagnostic hypotheses to explain patterns of data. Hypothesis generation is a pre-decisional process in which people form possible explanations of observations (data). The generation of diagnostic hypotheses is an important component of many real-world tasks such as diagnosing patients (Botti & Reeve, 2003; Elstein & Schwarz, 2006; Elstein, Shulman, & Sprafka, 1978; Vermande, van den Bercken, & De Bruyn, 1996; Weber, et al., 1993), determining the cause of car failure (Mehle, 1982), and interpreting patterns of scientific data (Fischhoff, 1977). Despite the centrality of hypothesis generation for understanding judgment and decision making, little research has been directed at understanding the underlying processes of hypothesis generation. Most judgment and decision making research has examined how people formulate beliefs about and test pre-specified hypotheses—hypotheses provided to the decision maker by the researcher. In contrast, for most real-world applications, the decision maker is enticed (and indeed required) to generate the to-be-judged hypotheses.

For any given observed pattern of data, or symptoms, there exists a set of possible explanations for those data. The exhaustive set of possibilities is defined external to the human decision maker, and is determined by the statistical relationships between the data and the possible true causal explanations. The external possibilities can be contrasted with the set of possible explanations, or hypotheses, that the decision maker entertains. Thus, in this paper the term hypothesis is used to refer to the mental event that is used as a best-guess explanation of the pattern of data. Hypotheses serve several functions. First, hypotheses explain existing patterns of data, and therefore serve a summary function. Hypotheses also can help structure information search and interpretation. Finally, hypotheses serve as input into judgment processes. For example, given that a decision maker is entertaining a set of possible hypotheses, the perceived likelihood of any hypothesis will be determined by comparing its support with the support for the set of alternative hypotheses under consideration.

This paper examines how hypothesis generation is affected by the order that information is observed. Order effects occur when people view the same information in different orders and then generate different sets of hypotheses. When information presented early in the sequence influences the set of hypotheses generated more than information presented later in the sequence, primacy effects result. Alternatively, when information presented later in a sequence influences the set of hypotheses generated more than information presented early in a sequence, recency effects result. Consider two doctors. One learns that a patient has shortness of breath, coughing, chest pain, and fatigue. The second doctor learns that a patient has fatigue, chest pain, coughing, and shortness of breath. Although the two doctors observe patients with the same symptoms, the physicians may generate different diagnoses due to the different orders that the symptoms were presented. Perhaps the hypothesis "pneumonia" is generated most often when the symptom "fatigue" occurs early in the sequence; and perhaps the hypothesis "bronchitis" is generated most often when coughing occurs early in the sequence. Weber et al. (1993) noted that the order in which information is presented has been found to affect performance in a variety of cognitive tasks from free recall to belief updating, but that no research has examined how information order affects hypothesis generation. However, real-world hypothesis generation involves gathering information over time rather than all at once. Even when all data are available at the same time, data enter the cognitive system sequentially via the sensory systems.

Studying hypothesis generation has important consequences for the ultimate accuracy of decisions. The processes of hypothesis evaluation and hypothesis testing are highly contingent on the processes of hypothesis generation and, as a consequence, errors in hypothesis generation cascade into errors in probability judgment and information search (Thomas, Dougherty, Sprenger, & Harbison, 2007). Koehler (1994) pointed out that there are differences in the way confidence estimates are made when people generate their own hypotheses versus when they are given hypotheses to evaluate. Indeed, he found that participants who generated their own hypotheses expressed less confidence that the hypotheses were true and were consequently less overconfident than those who were presented with a set of hypotheses for evaluation. Hypothesis generation leads people to test more alternatives than when people only evaluate pre-specified hypotheses. Klahr and Dunbar (1988) found that spending time generating hypotheses before designing experiments and testing hypotheses had important effects on the hypothesis-testing process. Participants who generated hypotheses before testing them solved tasks more quickly and correctly than participants who did not generate hypotheses. Participants who did not generate a set of alternative hypotheses took longer to abandon hypotheses that had been refuted by experimentation. Further, Farris and Revlin (1989) argued that the confirmation bias in hypothesis testing found by Wason (1960) could be due to the hypothesis-generation process. Normally when testing hypotheses, people generate hypotheses and then compare hypotheses with their competitors. However, when the task involves only hypothesis testing, participants consider each hypothesis separately and try to disconfirm one hypothesis at a time, leading to confirmation bias. Finally, Fisher (1987) noted that the failure to generate a correct hypothesis in tasks such as mechanical troubleshooting, analysis of aerial reconnaissance photography, military intelligence, and scientific inquiry can have deleterious consequences. Thus examining factors influencing hypothesis generation has important consequences for how people test and evaluate hypotheses.

More specifically, it is important to study how cue order affects hypothesis generation. This research is necessary because in most real-world situations cues reveal themselves sequentially over time, rather than all at once. Further, hypothesis generation underlies hypothesis testing, which is an inherently dynamic process. For example, imagine that one passively observes some new data and wishes to find the correct hypothesis to explain the data. Observing new data prompts one to generate hypotheses. Then, one attempts to find the correct hypothesis by searching for new information that rules out some hypotheses currently being considered. Then, new information prompts one to evaluate previously considered hypotheses and to again generate new hypotheses. Thus, hypothesis generation and hypothesis testing form a dynamic cycle. As such, understanding the dynamic aspects of hypothesis generation will increase understanding of hypothesis testing.

A Model of Sequential Hypothesis Generation

Thomas et al. (2006) developed HyGene, a process model of hypothesis generation and evaluation. HyGene was based on three principles: 1) Information in the environment serves as retrieval cues to prompt the retrieval of hypotheses from LTM; 2) Working memory capacity and task characteristics constrain the number of alternative hypotheses that can be actively considered; and 3) hypotheses maintained in WM serve as input for probability assessment and guide information search and hypothesis testing. The first principle suggests that hypothesis generation is a general case of cued recall, except that in hypothesis generation a *set* of possible diagnostic hypotheses is usually generated whereas in cued recall tasks only *one* item is usually recalled. The second principle suggests that the number of hypotheses one can entertain is constrained by WM limitations and task constraints such as time pressure and divided attention. The third principle suggests that hypothesis generation drives probability judgment and information search.

Figure 1 presents an overview of the processes proposed in HyGene (Thomas et al., 2007). I will briefly describe the original model, and then will specify the sequential version currently being developed (Dougherty, Harbison, Sprenger & Thomas, 2007). For a more mechanistic account of HyGene, see Thomas et al. (2007).

1) First, the generation process is initiated (top left box in Figure 1). It is assumed that data from the environment serve as retrieval cues that prompt the retrieval of diagnostic hypotheses from memory. In the case of a clinician, she may

Figure1

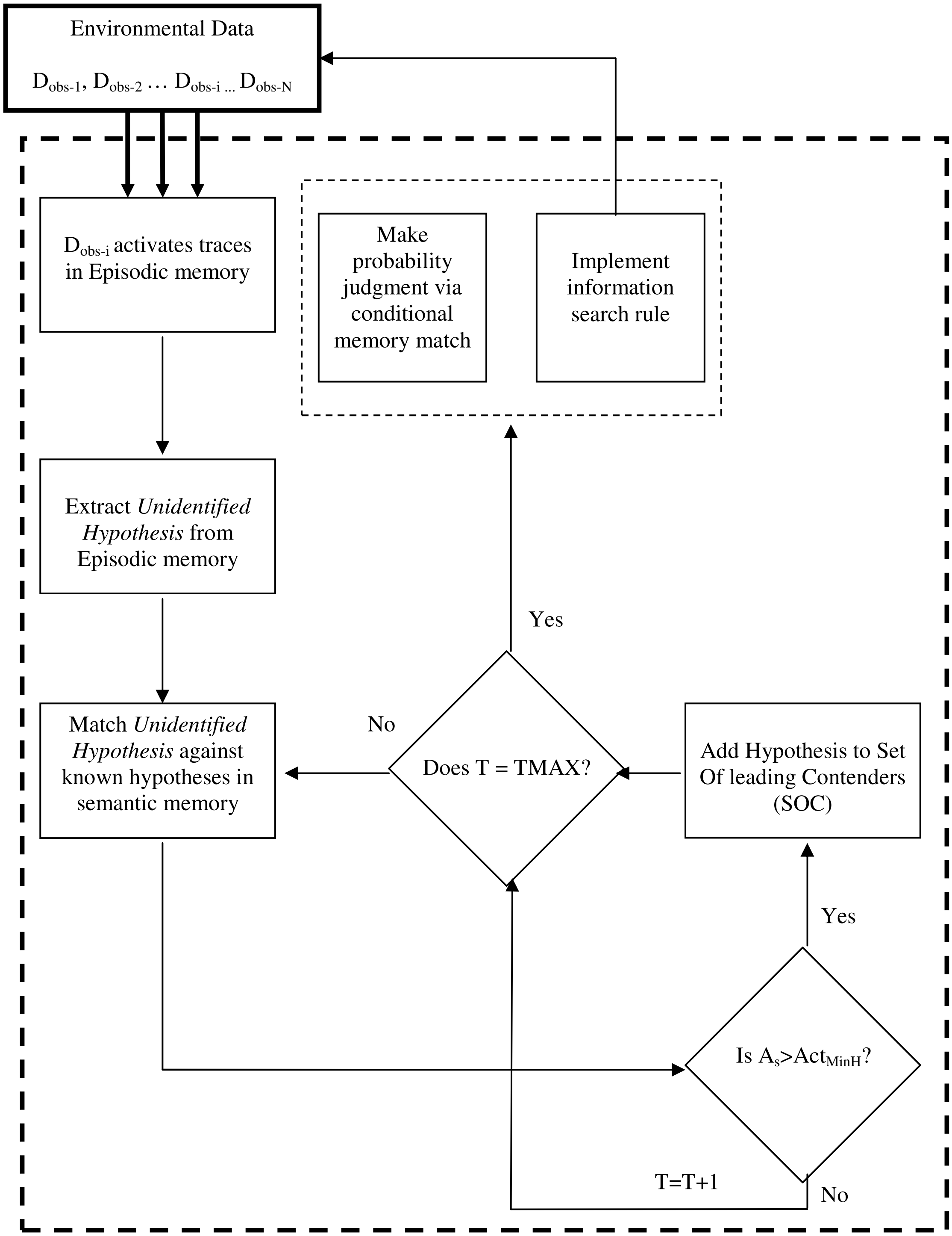


Figure 1 illustrates the processes involved in the HyGene model.

begin generating hypotheses as soon as new symptoms are described or new test results are received.

2) Observing data activates traces in episodic memory that represent past patients who have exhibited symptoms similar to the observed data.

3) Episodic traces that are activated above a threshold (A_c) produce an unidentified hypothesis that resembles those hypotheses that are most commonly and strongly associated with the data.

4) The unidentified hypothesis is then matched against known hypotheses (e.g., diseases prototypes) in semantic memory to determine the possible hypotheses for explaining the initial observed data, symptoms that co-occur with the observed data, and potential treatments that have been associated with the observed data in the past. The probability of sampling a given hypothesis in semantic memory is proportional to its similarity to the unidentified hypothesis.5) Hypotheses in semantic memory are generated and placed in working memory if they are sufficiently activated by the unidentified hypothesis. To enter working memory, the activation of the hypothesis in semantic memory (A_s) must exceed the activation of the least-activated hypothesis in working memory (Act_{MinH}). Working memory is limited in capacity. Once WM capacity is reached, new hypotheses that enter working memory *replace* old hypotheses that have the lowest activation.

6) Consistency checking takes place. In consistency checking, retrieved hypotheses are compared with observed data to ensure that they are consistent with those data. Hypotheses that are inconsistent with observed data are eliminated from working memory.

7) At some point during the generation process, people stop searching for new hypotheses. It is assumed that the generation process terminates when the number of successive retrieval failures (T) is higher than a criterion, TMAX.

8) After generation has ended, hypotheses in working memory can be used either for assessing the likelihood of the generated hypotheses and/or for organizing the search for information in the environment. Judgments of probability are assumed to derive from a comparison between the memory-support for the hypothesis in question with the memory-support for all other hypotheses in WM.

HyGene was developed to account for hypothesis generation when a set of cues are learned simultaneously. However, in most real-world situations people observe cues sequentially over time. This real-world constraint raises the theoretical question; what modifications are necessary for HyGene to account for sequential hypothesis generation? Dougherty, Harbison, Sprenger, & Thomas (2007) are currently developing a *sequential* sampling model of hypothesis generation. This model, Sequential HyGene, assumes that both cues present in the environment as well as previously observed cues are activated and used to probe LTM. However, the capacity to maintain cues is limited, and consequently people consider only a subset of cues during hypothesis generation. Fisher (1987) noted that when more than two data have been observed, the process of hypothesis generation may become more complex because of limited processing capacity of working memory. Working memory constrains the number of cues that can be active at one time. Gettys et al. (1978) provided estimates of the number of data that decision makers use to retrieve hypotheses in problems containing more than two data. It was estimated that participants retrieve hypotheses in response to two data in three- data problems and approximately three data in response to six-data problems (as reported by Fisher, 1987). In addition to working memory limiting the number of data simultaneously active during hypothesis retrieval, some previously-observed cues may be forgotten as time passes. Thus, both working memory constraints and forgetting lead participants to use a subset of all possible cues in hypothesis generation, and this is especially true when the number of observed cues is large.

Because a limited number of cues can be simultaneously maintained in working memory, the activation of each cue changes as new cues are observed. Which cues are remembered and used to probe memory will likely depend in part on the order they were presented.

The Sequential HyGene model assumes that both primacy and recency affect which data will be used to probe memory. The model activates cues based on the retrieval function specified in equation 1:

$$Activation_i = 1 - \left(\gamma^{\alpha^{i-1}} \times \varphi^{\beta^{N-i}}\right) \qquad (1)$$

The activation of the i[th] cue is a function of proactive interference (the γ term) and retroactive interference (the φ term). N represents the total number of cues presented so far. Alpha and beta are attentional weighting parameters. The parameters γ, φ, α, β are all constrained to range between 0 and 1. Note that the cue activation function is not intended to provide a *psychologically plausible* model of cue retention, but rather it merely provides a means to model the retention of cues. The shape of the cue-activation curve depends on γ, φ, α, β, and list length. Figure 2 presents the

Figure 2

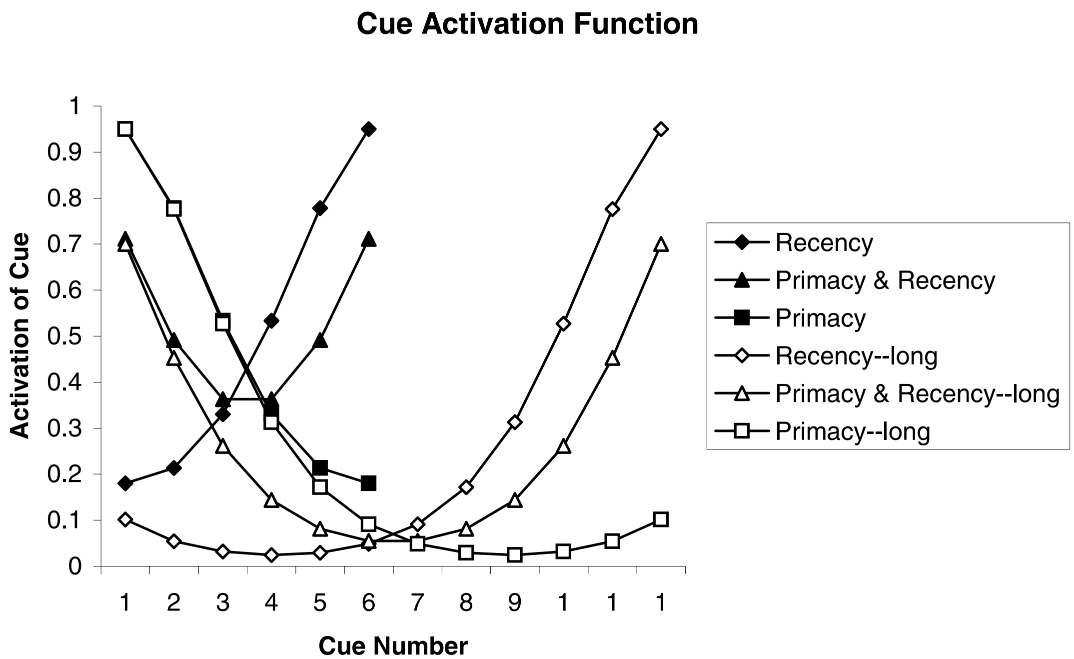


Figure 2 presents the cue activation function for the Sequential HyGene model.

cue activation function for three different parameter settings[1] and two different list lengths (6 and 12). As one can see, in the "recency" curves, participants weight later cues more than early cues. Here, proactive interference (γ) is set to a high value (.9) and retroactive interference (φ) is set to a low value (.05). In the "primacy" curves, participants weight early cues more than later cues. Here, proactive interference (γ) is set to a low value (.05), and retroactive interference (φ) is set to a high value (.9). In the "primacy and recency" curve, participants weight early and late cues more than mid-sequence cues. Here, proactive and retroactive interference are both set to low values (.3). Figure 2 also demonstrates the effect of list length on the cue activation function. As more cues are presented, more cues are "forgotten" (given low attention weightings). The effect of the attention parameters, α and β , interacts with the values of the proactive and retroactive interference parameters.

Cues experienced over time are integrated into a single composite probe, where the degree to which each cue contributes to the composite probe is weighted by its activation in working memory. After the consolidated data probe is created, the steps assumed in Sequential HyGene model continue as previously described and as depicted in Figure 1.

What are the implications of the cue-maintenance process assumed by Sequential HyGene? The particular data that are activated directly affect which hypotheses are generated. For instance, if data from the beginning of the sequence have been forgotten, then the consolidated data probe will have more features

[1] In the "recency" curve, gamma=.9, alpha=.5, theta=.05, beta=.5. In the "primacy and recency" curve, gamma=.3, alpha=.5, theta=.3, beta=.5. In the "primacy" curve, gamma=.05, alpha=.5, theta=.9, beta=.5.

consistent with recent cues. Then, episodic memory traces consistent with the most recent data will tend to be activated since they will be most similar to the consolidated data probe. Consequently, the unidentified hypothesis will be most similar to recent data. Then, when semantic memory is probed with the unidentified hypothesis, hypotheses consistent with recent cues will tend to be retrieved.

One of the main goals of the current study was to gain insight about how data are activated and weighted during hypothesis generation in order to provide an empirical basis for the development and testing of the Sequential HyGene model.

Order Effects in Other Cognitive Tasks

That most real-world hypothesis generation involves cues unfolding over time raises an empirical question: To what degree are people sensitive to cue order in hypothesis generation? Although research has not yet directly examined how observing the same data in different orders affects the set of hypotheses generated, order effects have been found in other areas of cognitive processing, such as in memory and belief updating.

Memory

Memory research has found primacy and recency effects in immediate free-recall tasks (Page & Norris, 1998; Rundus, 1971) and in delayed-recall tasks (Davelaar, Goshen-Gottenstein, Ashkenazi, Haarmann, & Usher, 2005; Greene, 1986; Koppenaal & Glanzer, 1990; Poltrock &MacLeod, 1977). If these serial position effects generalize to hypothesis generation, one would expect that hypothesis generation would be most affected by data occurring early and late in the sequence.

Page and Norris (1998) argued, “To a large degree, interest in serial recall stems from a conviction that it involves a system whose operation underlies performance in a great variety of cognitive tasks” (p. 761). However, without direct experimental evidence it is unclear whether memory research findings will generalize to hypothesis generation. Several differences between hypothesis generation and memory recall tasks exist that do not allow one to make direct generalizations from one domain to the other. First, in memory experiments, participants recall all items at one time after learning all items on a list. In hypothesis generation tasks, participants may recall and use cues after each cue is presented during the act of generation. By doing so, cues' memory representations may be strengthened when they are used during intermediate generation cycles, ultimately affecting the expected serial memory function. Secondly, whereas in typical memory tasks items are usually unrelated to each other, in generation tasks cues are usually semantically or associatively related to each other in the context one is considering (i.e., runny nose and tiredness are associated in the flu disease context). Glanzer and Schwartz (1971) found that participants recalled lists containing pairs of weak associates better than lists of unrelated words. Thirdly, In recall, typical words tend to be recalled more easily. However in generation, it is not always possible a priori to determine which cues are typical, because typicality depends on the context one is considering. Pei and Tuttle (1999) pointed out that recall tasks use experiments with well-structured taxonomic categories, whereas hypothesis generation experiments use categories created during generation that are often ad hoc and goal-derived. They noted, “For ad hoc categories, people's typicality perceptions (how well an item satisfies the ideal or goal for which the category is created) and frequency of instantiation (how often an item is encountered as a member of the category) for category members are context specific." (p. 238). Finally, whereas in typical memory paradigms the goal is to recall items, in hypothesis generation the goal is to use data in the generation task itself. Anderson and Hubert (1963) found evidence that the memory processes underlying person-impression judgment differs from the memory processes underlying recall tasks. Similarly, it may be the case that the memory processes involved in hypothesis generation differ from those in recall tasks, suggesting that one cannot directly use the memory literature to make predictions about serial position curves for cue-use in hypothesis generation tasks.

Belief Updating

In examining how beliefs are updated over time, researchers have found that beliefs differ depending upon the order that cues were presented. Both primacy and recency effects have been observed by different researchers (for a review, see Hogarth & Einhorn, 1992). Note, though, that findings from belief updating research may not directly generalize to hypothesis generation. In belief updating research, participants must both remember cues and evaluate hypotheses, but usually only two contrasting hypotheses are considered. WM demands are higher during hypothesis generation tasks than in typical belief updating tasks, because in generation tasks participants must store and consider multiple cues and multiple hypotheses at the same time.

Hypothesis Generation

No research has yet examined whether hypothesis generation is affected by cue order. Fisher (1987) conducted a study examining how data presented serially are differentially weighted in a hypothesis generation task. Participants completed hypothesis generation problems in which they read six pieces of data and then listed the hypothesis they thought was correct given the data presented. Fisher then gave participants a surprise recall test, in which participants were asked to recall the data that were presented. Fisher theorized that participants would best recall data that they had used most during generation. He found that participants recalled the first piece of data and the most diagnostic piece of data better than any other data. This experiment suggests that primacy effects occur in generation, and that people tend to remember and use diagnostic data more than non-diagnostic data. Note, though, that this study did not *manipulate* the order that cues were presented to determine the effects of order on generation; only one cue-order was used. Fisher had participants generate only one hypothesis, rather than a set of hypotheses. Further, Fisher focused on recall functions of the cue words, and not on whether generated sets of hypotheses were more consistent with early cues, later cues, or both. Gettys and Fisher (1979) also conducted a study in which hypothesis generation was examined as cues were presented sequentially. They found that participants tended to generate new hypotheses most often when new evidence did not support the previous set of hypotheses considered. Gettys and Fisher also did not manipulate the order that cues were presented to determine the effects of order on generation; only one cue-order was used. The experiments presented in this paper extend these previous findings by examining whether hypothesis sets generated after observing a sequence of cues are more influenced by early, middle, or late cues.

Questions of Interest, Predictions, and Hypotheses

A series of experiments were conducted to examine how the order that data are presented affects hypothesis generation. To empirically test hypothesis generation behavior, a virtual environment was developed in which participants imagined being detectives. For "job training" participants learned the relationship between robbers (possible causes) and evidence (data) in crimes committed in the past year in a hypothetical town called "Crimeville". After learning robber-evidence relationships, participants were told that a new crime occurred. Then, pieces of evidence were presented iteratively and participants generated robbers they thought caused the robbery (Experiments 1&3) and/or rated the probability of a given robber (Experiments 2&3). Experiment 4 was slightly different. Rather than learning relationships between cues and possible causes in the laboratory, participants generated hypotheses based on previously stored general knowledge. Also, in Experiment 4 many possible causes existed. Participants observed words from psychology course descriptions and guessed which course(s) was being described. The experiments tested: a) how the order in which data are presented affects hypothesis generation and probability judgment, b) whether response mode affects the type of order effects that obtain c) how probability judgments are affected by the hypothesis set under consideration, and d) which cues participants tend to recall after generating hypotheses. The empirical work both informs the development of Sequential HyGene and itself provides an important addition to the literature. The experiments address the empirical question of whether people are sensitive to cue order in their hypothesis generation. The experiments also provide an empirical database to aid in addressing the theoretical question of how to modify HyGene to account for order effects in hypothesis generation.

Order Effects

The main question of interest in these experiments was whether primacy or recency effects would obtain. Because the experiments were the first to examine order effects on hypothesis generation, it was not clear which types of order effects would most likely to obtain.

Number of Hypotheses Generated

Several studies have found that participants tend to generate around 4 hypotheses (Weber et al., 1993; Joseph & Patel, 1990; Barrows et al., 1982; Elstein et al., 1978; Mehle, 1982). Weber et al. argued that problem solvers may limit the size of their hypothesis set to the number of hypotheses they can use in later problem solving stages that are constrained by cognitive capacity limitations (such as working memory capacity). Weber et al. (1993) presented doctors with case studies of patients and asked them to generate diagnoses for those case studies. Weber et al. found that different doctors tended to generate different numbers of hypotheses (and the number of hypotheses a given doctor generated was consistent across case studies), but that there were no other factors that affected how many total hypotheses were generated. Although the amount of case information available did not affect the number of hypotheses generated, Weber et al. argued that more information may provide cues that trigger associated hypotheses not otherwise generated, but at the same time more information tends to constrain the set of explanations that are plausible given the set of symptoms. Thus, the doctors did not generate more or fewer hypotheses when more information was presented, because the added information led them to generate more similar hypotheses and fewer dissimilar hypotheses. In fact, Weber et al. found that the set of hypotheses doctors considered were more homogenous when given more information and were more heterogeneous when given less information. Thus, it was hypothesized that as participants received more and more cues (information) the total number of hypotheses they considered would not change.

Base Rates

In the judgment domain, research has found that while people are not always sensitive to base-rates, they are able to integrate this knowledge into their judgments when it is acquired through experience (Medin & Edelson, 1988; Carroll & Siegler, 1977; Christensen-Szalanski & Beach, 1982; Christensen-Szalanski &Bushyhead, 1981). Hypothesis generation research has also found that when base rate information is learned through experience, generation is sensitive to base rate information. For example, Dougherty and Hunter (2003a) used a learning task and found that participants generated nearly twice as many high base-rate hypotheses compared to low base-rate hypotheses. Similarly, Weber et al. (1993) found that expert physicians nearly always generated the high base-rate disease prior to generating a low-probability (but high-cost) alternative. Similar results were reported by Dougherty et al. (1997) and Gettys et al. (1987). Thus, it was predicted that participants' hypothesis generation would be sensitive to base rates.

Chapter 2: Experiment 1

Methods

Participants

University of Maryland undergraduate students (n=128) participated in Experiment 1 for course extra credit. Participants were run individually in single sessions lasting approximately 40 minutes.

Materials

Materials included pictures of four robbers (the Silver Swindler, the Purple Pirate, the Black Bandit, and the Red Rogue) and pictures of 11 pieces of evidence (Location: north, south, east, west; Getaway Car: van, sports car; Job: jewelry store, house, bank; Value of stolen goods: high, low). These stimuli, and exact instructions for the entire experiment, are presented in Appendix A.

Design and Procedure

Experiment 1 consisted of two main phases, a learning phase and a generation phase. In the learning phase, participants were instructed to imagine that they were detectives who just moved to the city of "Crimeville". Participants were informed that they were going to review case files to learn about robberies committed in the past year. They were further told that knowing the history of the city's crimes would help them identify and capture criminals in the future. Participants were instructed that all of the robberies in Crimeville were committed by one of four different criminals (the Red Rogue, Silver Swindler, Purple Pirate, and Black Bandit), each of whom always acted alone. During the learning phase, participants viewed 200 robbery "cases".

Each case represented one robbery that had occurred in the past year, and for each case participants viewed five pieces of information about that robbery: the robber who committed the crime, the location of the robbery (north, south, east, or west), the type of robbery (a bank, jewelry store, or home), the value of items stolen (high or low), and the type of getaway car (van or sports car). The information was presented in both picture and word format on each screen. Appendix B presents an example of a single case. Participants pressed the first letter of the name of the robber who committed that robbery to continue from one case to the next case. For instance, if the robbery case was committed by the Black Bandit, participants pressed "B" to continue. This was done to ensure that participants viewed each screen. The learning was self-paced with the restraint that participants were forced to view each case file for a minimum of 2 seconds before the next case would appear.

Table 1 presents the frequency with which each piece of data (i.e. location, getaway car, job, and value of stolen goods) was associated with each possible cause (robber) in the training phase of the task. Note that the most diagnostic cue was the location that robberies occurred. The Black Bandit committed robberies mostly in North Crimeville, the Purple Pirate committed robberies mostly in South Crimeville, the Red Rogue committed crimes mostly in East Crimeville, and the Silver Swindler committed crimes mostly in West Crimeville. For the cue "type of job," three robbers (Black Bandit, Purple Pirate, and Silver Swindler) tended to steal from jewelry stores and one robber (the Red Rogue) tended to steal from houses. For the cue "value," the Red Rogue and Purple Pirate both tended to do jobs in which they stole high value goods whereas the Black Bandit and Silver Swindler did both jobs in which they stole

Table 1

	Location (N,S,E,W)	**Job (J,B,H)**	**Value (H,L)**	**Car (S,V)**	**Robber Base rate**
C1: Black Bandit	60,10,5,5	60,5,15	40,40	40,40	80
C2: Purple Pirate	5,35,5,5	40,5,5	40,10	25,25	50
C3: Red Rogue	5,5,35,5	5,5,40	40,10	25,25	50
C4: Silver Swindler	2,2,2,14	18,1,1	10,10	10,10	20
Total	**(72,52,47,29)**	**(123,16,61)**	**(130,70)**	**(100,100)**	**Total # crimes: 200**

Table 1 depicts the relationship between cues and possible causes (i.e. pieces of evidence and robbers) for Experiments 1 and 2. N=north, S=south, E=east, W=west, J=jewelry store, B=bank, H=house, H=high value, L=low value, S=sports car, V=van.

high value goods and jobs in which they stole low value goods. Finally, all robbers drove away from robberies in sports cars and vans equally often. For each case presentation the four data were selected randomly from the distribution presented in Table 1. Note that the robbers were completely counterbalanced.

In the second phase of the experiment, participants were told a new robbery had occurred and that they had been called in to figure out who had committed the robbery. Participants were further instructed that they would be given pieces of evidence (i.e. location, job, getaway car, and value cues) about this new crime scene and would be asked to generate a list of their top suspects based on the robbers' crime histories and the pieces of evidence available from the crime scene. They were told that their list of top suspects should include those people they would want to bring in for questioning because they were likely to have committed the robbery. Participants were told that their list of top suspects might include all four robbers, if the available evidence didn't help them narrow down the possibilities, and that other times their list may contain some subset of the four robbers because the available evidence helped narrow down the possibilities. The pieces of evidence were presented on the screen in both word and picture format. As an example, for the first piece of evidence in Crime Scene A participants read, "First, you learn that the robber drove away in a sports car". A picture of a sports car was displayed on the screen. Appendix C presents an example of what participants viewed during the testing phase. To "list" a suspect, participants pressed the first letter of that suspect's name. When participants were finished listing suspects, they pressed the spacebar. This cycle of receiving a piece of evidence and then generating robbers continued until participants viewed four pieces of evidence for that crime scene. Then participants were informed that a new crime had occurred, and the participants were again needed to help determine who committed the crime. Participants again received four pieces of evidence and generated robbers after receiving each piece of evidence for this second crime scene. Participants generated hypotheses for four different robberies. However, only two crime scenes will be discussed in this paper. The order that crime scenes were presented was counterbalanced.

Table 2 presents the two orders in which evidence was presented. In Table 2, the two columns show the order that participants in the two cue order groups saw the cues. Participants in cue order group 2 received cues in the opposite order of participants in cue order group 1. This manipulation provided a comparison of generation when the same cues were presented in different orders. For instance, for Crime Scene A, participants in cue order group 1 saw the piece of evidence: "sports car" at cue position 1, "jewelry store" at cue position 2, "low value" at cue position 3, and "north" at cue position 4. Participants in cue order group 2 saw exactly the same cues after the fourth and final cue was presented, but they saw the cues in the opposite order: "north" at cue position 1, "low value" at cue position 2, "jewelry store" at cue position 3, and "sports car" at cue position 4. Therefore any differences in generation after all cues were presented could be attributed to order effects. Figure 3 presents the objective probabilities of each possible cause (robber) after each cue was presented in each of the two cue orders for each crime scene (Figure 3 top panel presents Crime Scene A; bottom panel presents Crime Scene B). In Crime Scene A, participants viewed the most diagnostic cue ("north", which is diagnostic of

Table 2

		Cue Order Group 1	**Cue Order Group 2**
Crime Scene A			
	Cue 1	Sports Car	North
	Cue 2	Jewelry Store	Low Value
	Cue 3	Low Value	Jewelry Store
	Cue 4	North	Sports Car
Crime Scene B			
	Cue 1	West	Van
	Cue 2	Low Value	Jewelry Store
	Cue 3	Jewelry Store	Low Value
	Cue 4	Van	West

Table 2 presents the cues of the two Crime Scenes. Participants in cue order group 2 saw all cues in the opposite order of participants in cue order group 1. Participants saw one location cue, one value cue, one car cue, and one job cue in each crime scene. In crime scenes A and B participants saw the most diagnostic information (location) either first or last in the sequence.

Figure 3

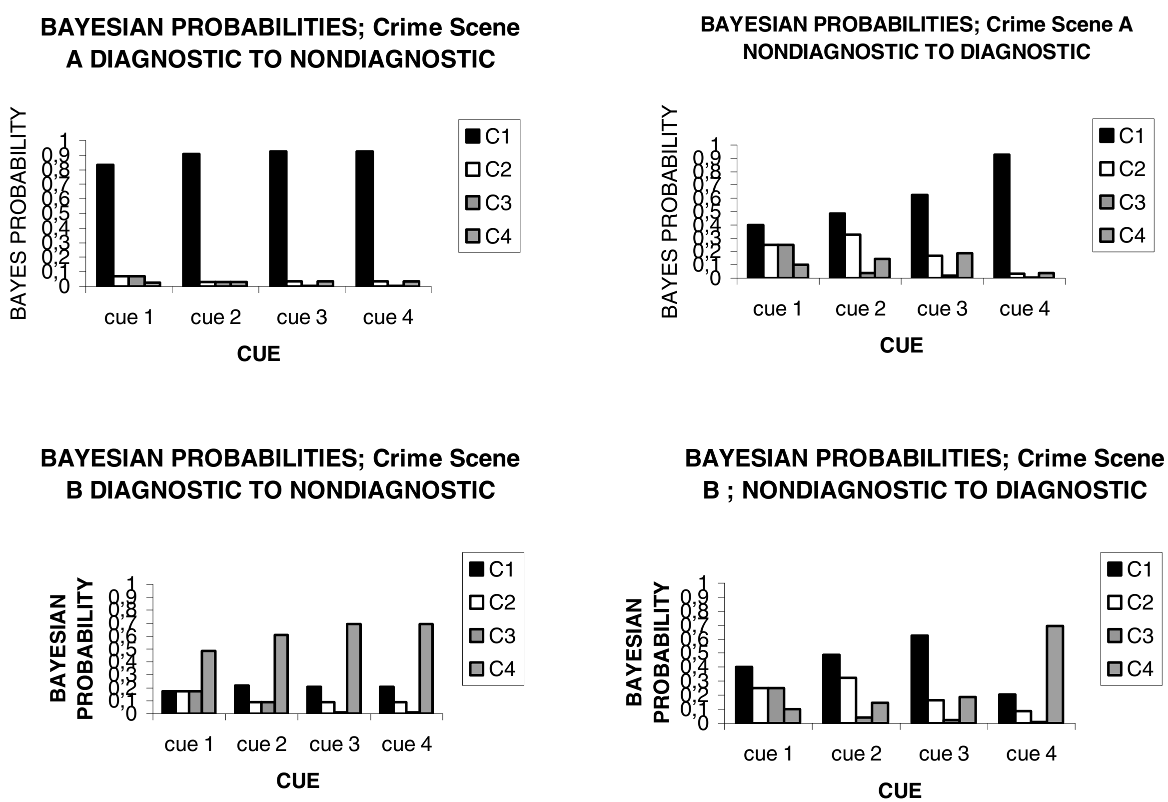


Figure 3 presents the objective probabilities for each possible cause after each cue was presented for the two cue order conditions. The top panel presents objective probabilities for Crime Scene A, and the bottom panel presents objective probabilities for Crime Scene B. In both crime scenes, one cue order presented a diagnostic cue first (left side) and the other cue order presented a diagnostic cue last (right side). The evidence in Crime Scene A pointed to H1 (the high base rate possibility) and the evidence in Crime Scene B pointed to H4 (the low base rate possibility). C1=possible cause 1; C2= possible cause 2; C3= possible cause 3; C4= possible cause 4.

possibility 1) either first (cue order group 2) or last (cue order group 1). Participants viewed non-diagnostic information ("sports car" which did not differentiate between robbers) either first (cue order group 1) or last (cue order group 2). In this crime scene, the most likely cause was possibility 1, the high base rate possibility, because it was much more likely than all other robbers. In Crime Scene B, the first and last cues were either highly diagnostic ("West" was diagnostic of possibility 4) or non-diagnostic ("van" did not differentiate between robbers) depending on cue order group. In Crime Scene B, the most likely cause was possibility 4, the low base rate possibility, because it was more likely than all other robbers.

Several factors were manipulated in this experiment. First, cue position was manipulated. Cues were presented at position 1, 2, 3, or 4. Second, in the testing phase the order of the cues was manipulated between subjects. Third, in the testing phase Crime Scenes were manipulated within subjects. Two crime scenes existed (see Table 2). A crime scene represented a single robbery, and the four pieces of evidence at that robbery (i.e. the location, type of getaway car, value of goods stolen, and type of job). Fourth, in the testing phase of the experiment, the cue presentation format was manipulated between subjects. Cues were presented either singularly or conjunctively. In the singular cue presentation condition, only the newest cue was presented on each screen. Thus, for Crime Scene A, cue order group 1, when the third cue was presented, participants in the singular condition read "Third, you learn that the robber drove away in a sports car," and participants saw a picture of a sports car. In the conjunctive cue presentation condition, each time a new cue was presented, the previous cues were also presented to remind participants what they had already learned about the crime scene. Thus, when the third cue was presented, participants in the conjunctive condition read, “Third, you learn that the robber drove away in a sports car. Thus, you now know that the robber committed the crime in South Crimeville, the robbery occurred in a bank, and the robber drove away in a sports car,” and participants saw pictures of south, the bank, and the sports car. Fifth, as can be seen in the right-most column of Table 1, in the training phase the base rate of the possible causes (i.e. robbers) was manipulated. Possibility 1 committed 80 robberies, possibilities 2 and 3 each committed 50 robberies, and possibility 4 committed 20 robberies.

Results

Data Analysis

Two dependent variables were analyzed: the number of hypotheses generated after each cue was presented and co-occurrence values. The number of hypotheses generated was simply the total number of hypotheses that participants listed after each cue was presented. Co-occurrence values are used to represent meaning in Burgess’s (Burgess & Lund, 1997) Hyperspace Analogue to Language (HAL) model of semantic knowledge. In the current experiments, co-occurrence values represent how closely a hypothesis “co-occurred” with a cue that was presented (i.e. the position at which a hypothesis was generated after a cue). Higher values represented that a hypothesis occurred more closely in time with the cue (i.e. sooner after the cue was presented). More specifically, the co-occurrence value ‘4’ was given to a hypothesis generated in the first position after the cue was presented; the co-occurrence value ‘3’ was given to a hypothesis generated in the second position after the cue was presented; the co-occurrence value ‘2’ was given to a hypothesis generated in the third position after the cue was presented and the co-occurrence value ‘1’ was given to a hypothesis generated in the fourth position after a cue was presented. If a hypothesis was not generated at all after a cue was presented, it was given the value 0 because it did not co-occur with the cue at all.

Analyses treated co-occurrence values (0, 1, 2, 3, or 4) as an interval dependent variable to be modeled as a function of independent variables (cue presentation format (conjunctive vs. singular); cue order group (original or reversed); cue position (position 1, position 2, position 3, or position 4); and hypothesis (hypothesis 1, 2, 3, or 4) using the General Linear Model (GLM).

As a preview, several findings are of interest in Experiment 1. First, hypothesis generation was more influenced by recent cues than by initial cues. Second, hypothesis generation was affected by the base rates of the possibilities. Third, participants tended to use more cues when all cues were present than when participants had to recall cues from memory. Fourth, participants tended to generate more hypotheses after an initial cue was presented and later generated fewer and fewer hypotheses after other cues were provided. Finally, participants tended to generate more hypotheses after the first cue when it was non-diagnostic than when the first cue was diagnostic.

Cue Order Effects

The first question of interest was whether generation differed when diagnostic information was presented first versus last in Crime Scenes A and B. If one weighted all of the cues equally, generation after all four cues had been presented should not differ for participants who received the most diagnostic information first (as cue 1) versus last (as cue 4). In contrast, if participants tended to use recent cues more than earlier cues, generation after cue 4 should be more affected by the diagnostic cue when it was presented last (as cue 4) than when it was presented first (as cue 1). If primacy occurred, participants would be most affected by diagnostic information when it was presented first versus last. For this comparison, only the singular cue presentation condition was examined, because in the conjunctive condition all cues were on the screen and thus participants did not need to recall previous cues. Thus, I examined co-occurrence values for the four hypotheses after cue 4 was presented, and compared cue order (whether the most diagnostic information was presented as cue 4 or as cue 1) in the single cue presentation condition.

Figure 4 presents mean co-occurrence values for the two cue order groups (diagnostic information first vs. last) after cue 4 for Crime Scenes A and B. The figure shows that in both crime scenes, higher co-occurrence values were found for the correct[2] hypothesis than for incorrect hypotheses. In both crime scenes, analyses found a main effect of hypothesis (Crime Scene A[3]: $F(3, 60)=11.22$, $p<0.0001$; Crime Scene B[4]: $F(3, 60)=6.50$, $p=0.0007$). Participants generated the most likely hypothesis more often and earlier than all other hypotheses.

Figure 4 shows a trend of higher co-occurrence values for the correct hypothesis when diagnostic information was presented last than when it was

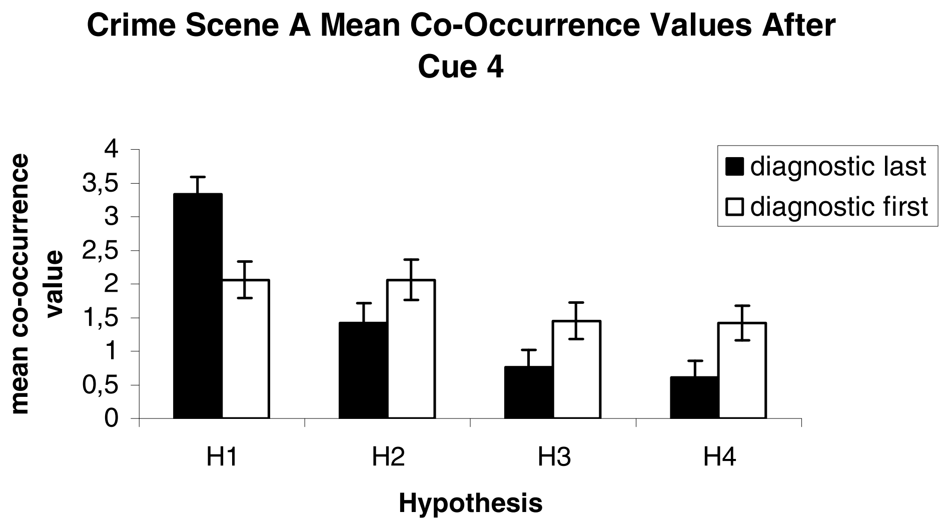
---

[2] "Correct hypothesis" refers to the hypothesis with the highest objective likelihood.

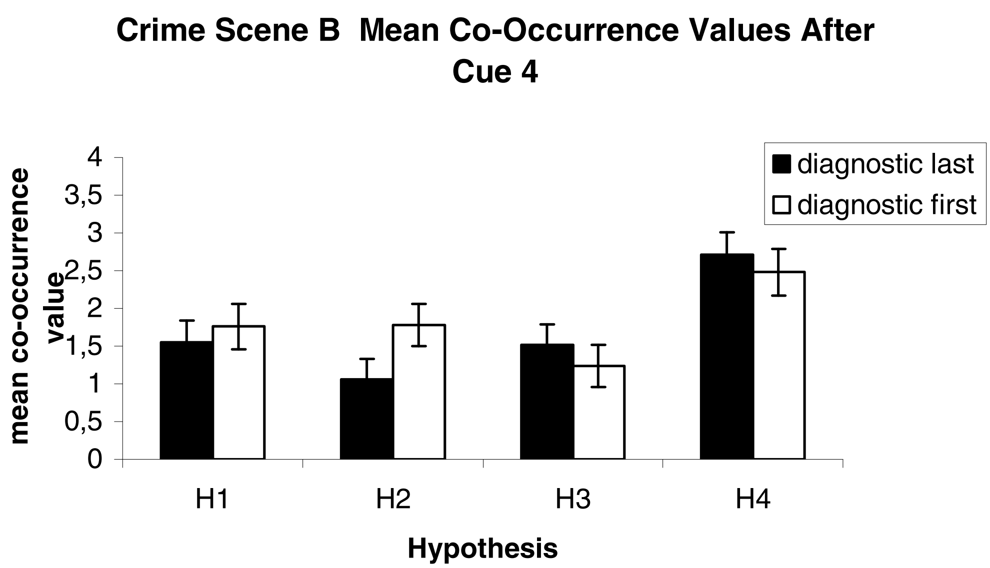
[3] Co-occurrence values for hypothesis 1 (the "correct" hypothesis) were higher than for hypotheses 3 and 4, but not 2 (H1 vs. H2: $t(63)= 1.80$, $p=0.076$; H1 vs. H3: $t(63)= 3.56$, $p=0.0007$; H1 vs. H4: $t(63)= 5.35$, $p<0.0001$)

[4] Co-occurrence values for hypothesis 4 (the "correct" hypothesis) were higher than for the other three hypotheses (H1 vs. H4: $t(63)= -2.47$, $p=0.016$; H2 vs. H4: $t(63)= -3.55$, $p=0.0007$; H3 vs. H4: $t(63)= -4.19$, $p<0.0001$)

Figure 4



A



B

Figure 4 displays the effect of cue order on hypothesis generation and compares when diagnostic information came first versus last. Panel A presents Crime Scene A results, and Panel B presents Crime Scene B results. The correct hypotheses are H1 for Crime Scene A and H4 for Crime Scene B.

presented first. Further, participants tended to generate incorrect hypotheses more when diagnostic information was presented first than when it was presented last. A significant interaction between cue order group and hypothesis was found for Crime Scene A, $F(3, 60)=4.20$, $p=0.009$. Participants who received the diagnostic cue last (as cue 4) had higher mean co-occurrence values for hypothesis 1 (the “correct” hypothesis) than did participants who received the diagnostic cue first (as cue 1), $t(62)=3.36$, $p=0.001$. On the other hand, participants who received the diagnostic cue last had lower mean co-occurrence values for hypothesis 4 (an “incorrect” hypothesis) than did participants who received the diagnostic cue first, $t(62)= -2.27$, $p=0.027$. These results suggest that participants who received the diagnostic information as cue 1 tended to weight it less than did participants who received diagnostic information as cue 4. However, no significant interaction between cue order group and hypothesis was found for Crime Scene B.

A second way I examined cue order effects was to compare generation after cue 4 when all cues were presented on the screen (conjunctive cue presentation condition) with generation after cue 4 when only the most recent cue was on the screen (singular cue presentation condition) when the most diagnostic information was presented first. It was hypothesized that if participants tended to weight cues differently based on their serial position, then differences should be found between the two presentation conditions. In the conjunctive presentation condition, forgetting or underweighting of early cues should not occur since all cues were available during generation. In contrast, in the singular presentation condition, participants could forget or underweight previous cues that were no longer on the computer screen during generation. If participants underweighted early cues in the singular presentation condition, then participants in the conjunctive condition should be more affected by that diagnostic information than participants in the singular condition when diagnostic information was presented early in the sequence.

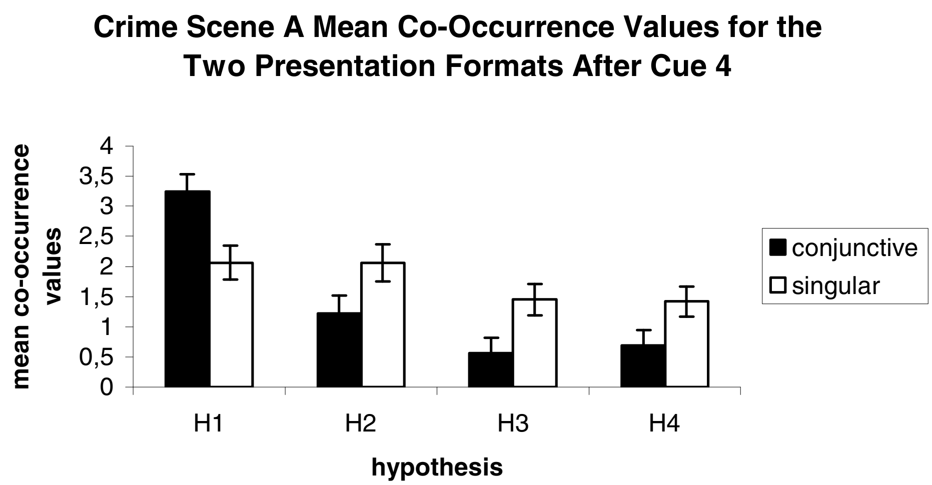
Figure 5 presents mean co-occurrence values for the four hypotheses after cue four for Crime Scenes A (top panel) and B (bottom panel). The figure compares the conjunctive and singular cue presentation conditions when diagnostic information was presented as cue 1. In Crime Scene A, the figure shows that co-occurrence values for the correct hypothesis was higher in the conjunctive condition than the singular condition. Further, mean co-occurrence values for incorrect hypotheses were lower in the conjunctive condition than in the singular condition. Analyses found that in one of two cases, there was a significant interaction between hypothesis and cue presentation condition (Crime Scene A: $F(3, 59)=3.95$, $p=0.012$). In Crime Scene A, participants in the conjunctive cue presentation condition had higher mean co-occurrence values for the correct hypothesis than did participants in the singular cue presentation condition. On the other hand, participants in the conjunctive cue presentation condition had lower mean co-occurrence values for incorrect hypotheses than did participants in the singular cue presentation condition[5]. These results suggest that

---

[5] Crime Scene A: There was a significant difference between the conjunctive and singular conditions for hypothesis one, $t(62)=2.98$, $p=0.004$; three, $t(62)= -2.44$, $p=0.018$; and four, $t(62)= -2.08$, $p=0.041$. Crime Scene B: There was a significant difference between the conjunctive and singular conditions for hypothesis one, two, and three (H1: $t(62)= -2.78$, $p=0.007$; H2: $t(62)=3.03$, $p=0.004$; H3: $t(62)= -2.17$, $p=0.034$). For hypothesis 2 (the correct hypothesis), participants in the conjunctive cue presentation condition (M= 3.44, SE=0.24) had higher mean co-occurrence values than did participants in the singular cue presentation condition (M=2.26, SE=0.31). For hypothesis 1 (an incorrect hypothesis), participants in the conjunctive cue presentation condition had lower mean co-occurrence values (M=0.94, SE=0.25) than did participants in the singular cue presentation condition (M=2.06, SE=0.32). Similarly for hypothesis 3 (an incorrect hypothesis), participants in the conjunctive condition again had

Figure 5

A



B

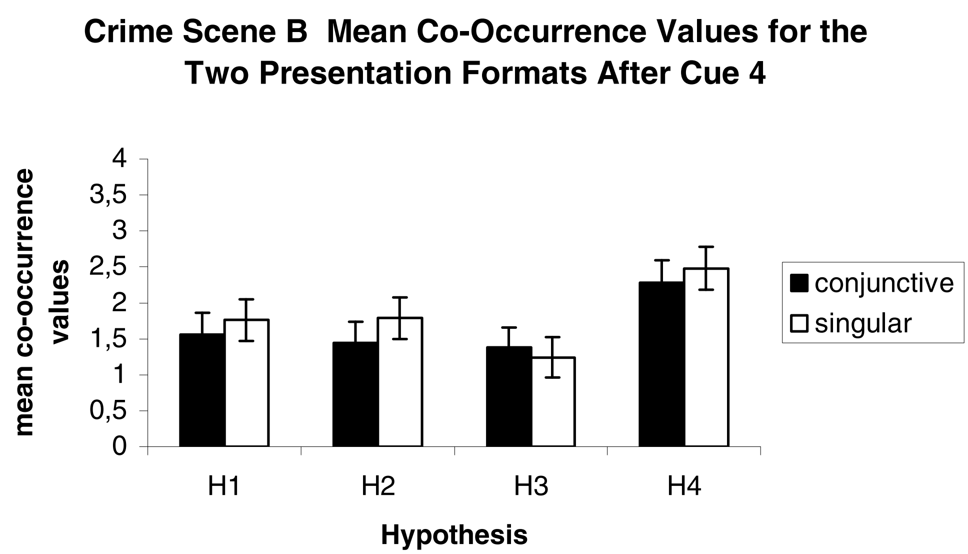


Figure 5 displays the effect of Presentation Format on hypothesis generation and compares the conjunctive and singular presentation conditions. Panel A presents Crime Scene A results; and Panel B presents Crime Scene B results. The correct hypotheses are: H1 for Crime Scene A and H4 for Crime Scene B.

---

lower co-occurrence values (M=0.81, SE=.24) than participants in the singular condition (M=1.61, SE=.28).

participants who received the diagnostic information as cue 1 tended to forget or underweight it when generating hypotheses after receiving the other 3 cues if the cues were presented singularly. As a result, they generated the "correct" hypothesis (hypothesis 1) less often and later in succession after cue 4 was presented than did participants who received the most diagnostic information as cue 4. They generated an "incorrect" hypothesis more often and earlier in succession after cue 4 was presented than did participants who received the most diagnostic information as cue 4.

Base Rate Effects

To test the effect of base rates on generation, I examined differences in generation after participants were given a non-diagnostic cue as their first piece of information. Then, the only information available to aid generation was the base rates of the hypotheses themselves. Thus, if participants were sensitive to differences in base rates, one would predict that high base rate items would be generated more often and in closer succession after the non-diagnostic cue was presented. Figure 6 presents mean co-occurrence values for each hypothesis when the first piece of evidence was non-diagnostic. One can see that co-occurrence values were higher for high and middle base rate hypotheses (H1, H2, and H3) than for the low base rate hypothesis (H4). Analyses found a main effect of hypothesis for Crime Scene A (Crime Scene A[6]: $F(3, 62)=5.25$, $p=0.002$; Crime Scene B: $F(3, 60)=2.57$, $p=0.063$). There was some sensitivity to base rates during hypothesis generation. Participants

[6]Contrasts revealed that hypotheses 1, 2, and 3 had higher co-occurrence values than did hypothesis 4. No other hypotheses differed reliably from each other. H1 vs. H4: $t(64)=3.61$, $p=0.0006$; H2 vs. H4: $t(64)=3.32$, $p=0.002$; H3 vs. H4: $t(64)=2.25$, $p=0.028$

Figure 6

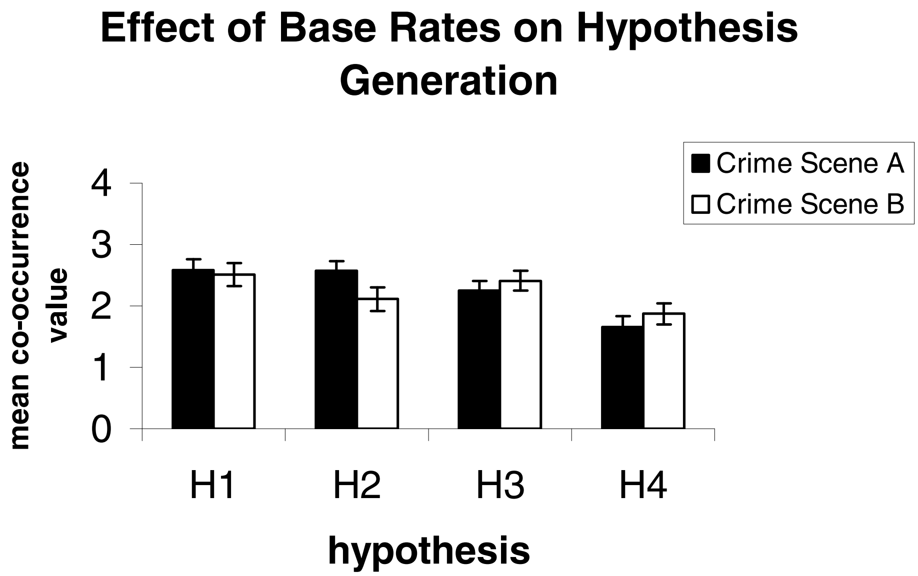


Figure 6 displays the effect of base rate on hypothesis generation. Hypothesis 1 (H1) had the highest base rate, followed by hypotheses 2 and 3 (H2 and H3), and then followed by hypothesis 4 (H4).

generated the high and middle base rate hypotheses more often and in closer relation to the non-diagnostic cue than the low base rate hypothesis. However, participants did not generate the high base rate item more often or in closer relation to the cue than the middle base rate items.

Number Generated

Figure 7 presents the number of hypotheses generated after each cue was presented for each Crime Scene. The top figures present data for Crime Scene A, and the bottom figures present data for Crime Scene B. The left figures present the diagnostic to non-diagnostic cue order and the right figures present the non-diagnostic to diagnostic cue order. Three general conclusions can be reached from examining how many hypotheses people generated after each cue was presented: First, participants tended to generate more hypotheses after an initial cue was presented and later generated fewer and fewer hypotheses after other cues were provided, and this was true in both cases (main effect of cue position: Crime Scene A[7]: F(3, 372)=55.68, p<0.0001; Crime Scene B[8]: F(3, 372)=30.10, p<0.0001). Second, participants tended to generate more hypotheses after the first cue when it was non-diagnostic than when the first cue was diagnostic (interaction between cue position and cue order group: Crime Scene A[9]: F(3, 372)=6.42, p=0.0003; Crime Scene B[10]: F(3, 372)=4.61,

---

[7] Participants generated more hypotheses: after cue 1 than after cue 3, t(127)=4.44, p<0.0001; after cue 1 than after cue 4, t(127)=9.02, p<0.0001; after cue 2 than after cue 3, t(127)=3.79, p=0.0002; after cue 2 than after cue 4, t(127)=10.18, p<0.0001; and after cue 3 than after cue 4, t(127)=8.42, p<0.0001.

[8] Participants generated more hypotheses after: cue 1 than after cue 2, t(127)=3.86, p=0.0002; after cue 1 than after cue 3, t(127)=4.72, p<0.0001; after cue 1 than after cue 4, t(127)=7.51, p<0.0001; after cue 2 than after cue 3, t(127)=2.25, p=0.026; after cue 2 than after cue 4, t(127)=5.65, p<0.0001; and after cue 3 than after cue 4, t(127)=3.93, p=0.0001.

[9] At cue position 1, participants who received a non-diagnostic cue generated more hypotheses (M=3.40, SE=.12) than participants who received a diagnostic cue (M=2.81, SE=.12), t(127)=3.49, p=.0007. There was no difference between the two groups for the other 3 cue positions.

Figure 7

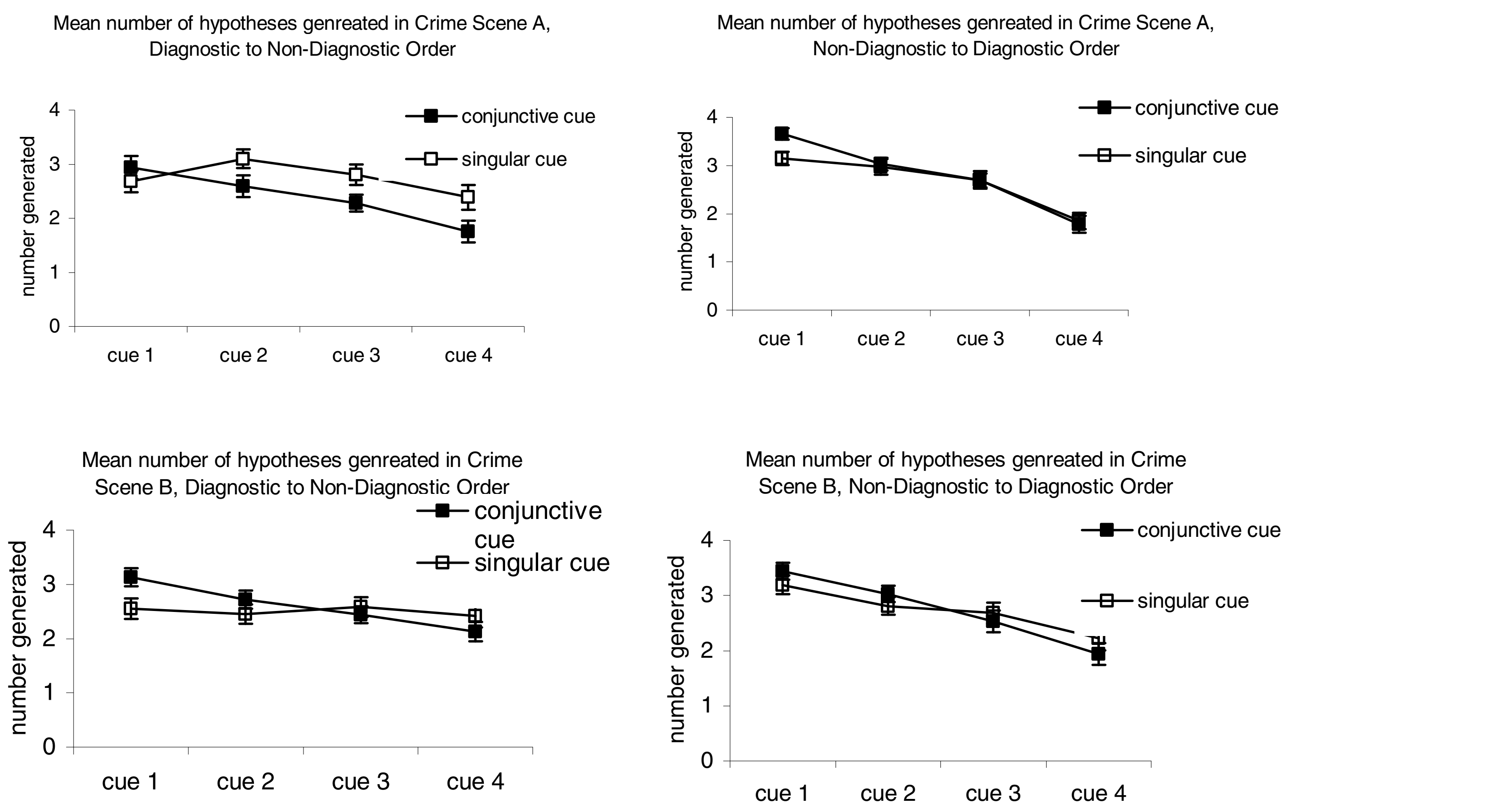


Figure 7 presents the number of hypotheses generated after each cue was presented for the two presentation format conditions. The top panel presents the number generated for Crime Scene A, and the bottom panel presents the number generated for Crime Scene B. In both crime scenes, one cue order presented the most diagnostic cue first (left side) and the other cue order presented the most diagnostic cue last (right side).

p=0.004). In both crime scenes, participants generated more hypotheses after cue 1 when cue 1 was non-diagnostic than when cue 1 was diagnostic, and in 1 of the 2 crime scenes this trend continued to cue positions 2 and 3. In both crime scenes, after cue 4 the two groups did not differ in the number of hypotheses they generated. Third, participants tended to narrow down the size of their hypothesis set after the 4th cue more in the conjunctive cue presentation format than in the singular cue presentation format, and this was true in both cases (interaction between cue position and presentation format: Crime Scene A[11]: F(3, 372)=6.00, p=0.0005; Crime Scene B[12]: F(3, 372)=5.76, p=0.0007). For both cases there was an unexpected difference between participants in the two presentation formats at cue position 1, surprising because at this point there were no differences in the manipulation between these groups.

For both cases, there was no main effect of presentation format, and there were no interactions between presentation format and cue order group, or between cue position, presentation format, and cue order group.

Discussion

In Experiment 1, evidence indicated that different cue presentation orders led participants to generate different hypothesis sets. For Crime Scene A, participants who received the most diagnostic cue at position 1 tended to forget or underweight it

---

[11] Participants generated more hypotheses in the conjunctive condition (M=3.30, se=.12) than in the singular condition (M=2.91, SE=.12), t=2.23, p=0.027, after cue 1, but not after any of the other 3 cues. However, this interaction was unexpected because after cue 1, there were no differences in the manipulations between conjunctive and singular groups (both groups received the same cue and only had that cue on the screen in front of them). Because after cue 1 there was no manipulated difference between the two groups, there was no reason to expect differences between groups at this point.

[12] After receiving cue 1, participants in the conjunctive condition (M=3.28, se=.12) generated more hypotheses than participants in the singular condition (M=2.87, se=.12), t(127)=2.40, p=0.018, and there were no difference between format for cues 2, 3, or 4.

when generating hypotheses after receiving the other 3 cues. As a result, they generated the “correct” hypothesis (hypothesis 1) less often and later in succession after cue 4 than participants who received the most diagnostic information at position 4. This result indicates that recency obtained, in that diagnostic information had stronger effects on final generation when it occurred later in the sequence than when it occurred earlier in the sequence. However, this result was not replicated for Crime Scene B, allowing no strong conclusions to be drawn. In Crime Scene A, participants in the conjunctive presentation condition had different generated sets than participants in the singular presentation condition when diagnostic information was presented early in the sequence and generation was examined after all four cues were presented. Participants in the conjunctive presentation condition generated the “correct” hypothesis more often and earlier than did participants in the singular condition. Also, participants in the conjunctive condition generated “incorrect” hypotheses less often and later in succession than did participants in the singular cue presentation condition. Again, these results indicate that recency effects obtained, in that participants weighted diagnostic information less when it was presented early in the sequence and later was not presented on the screen than when it was later presented on the screen as a reminder.

Experiment 1 also found that participants are sensitive to base rates when generating hypotheses. When completely non-diagnostic evidence was presented at position 1, participants generated high and middle base rate hypotheses earlier and more often than low base rate hypotheses, as indicated by having higher co-occurrence values. Interestingly, when people are asked to judge the likelihood of a hypothesis whose base rates are presented in a word problem, rather than learned through experience, judgments tend not to be sensitive to base rates in their judgments, a finding called “base-rate neglect”. However, as discussed, hypothesis generation does seem to take base rate information into account. Thus, perhaps previous tasks that do not allow for generation from experience underestimate people’s ability to incorporate base rate information into their judgments in real-world tasks when they themselves generate the hypotheses. Beyth-Marom and Arkes (1983) and Christensen and Beach (1983) suggested that people may give probability estimates in accordance with Bayes’s theorem by estimating conditional probabilities directly from their memory. Judgments using memory strength as input work well when the memory strength is correlated with actual frequencies. Wallsten (1981) found sensitivity to base rates in experienced physicians but not in medical students. This finding may suggest that physicians with more experience have learned the base rates over time and thus are sensitive to them in generation and judgment, whereas students who have only learned the base rates from book knowledge are insensitive to base rates in hypothesis generation and judgment.

In terms of the number of hypotheses generated, participants generated more hypotheses after an initial cue and generated fewer hypotheses after other pieces of information were presented. Participants seem to begin with a larger set of hypotheses and then rule out hypotheses as they receive more information. Further experiments will determine if this finding generalizes to other tasks and other domains. Perhaps the particular detective task used here induces a mental-set that encourages participants to reduce the number of hypotheses they consider as more evidence is received, even if new evidence does not rule out previously-considered hypotheses. Or, it could be the case that participants lose motivation to generate hypotheses as they view more information. The trend of a decreasing number of hypotheses generated was qualified by interactions with cue order and with presentation format. Participants initially receiving a non-diagnostic cue generated more hypotheses than participants initially receiving a diagnostic cue. This result suggests that participants recognized that non-diagnostic cues are consistent with more possible explanations than are diagnostic cues. Note, though, that participants who received a non-diagnostic cue at position 4 did not generate more hypotheses than participants who received a diagnostic cue at position 4. Thus, participants do not always generate more hypotheses after receiving a non-diagnostic cue, but rather do so only if no other diagnostic information has yet been presented. Participants tend to narrow down the size of their hypothesis set after the fourth cue more so in the conjunctive cue presentation format than in the singular cue presentation format. In the singular presentation-format, participants may forget or fail to use previous cues that provided information ruling out some hypotheses. In the conjunctive format, all cues were in front of participants when generating hypotheses leading participants to use more cues than participants in the singular format.

Chapter 3: Experiment 2

Whereas Experiment 1 examined hypothesis generation, Experiment 2 examined probability judgments when cues were presented sequentially. Most theories of probability judgment assume that people compare a focal hypothesis with at least one alternative hypothesis (Tversky & Koehler, 1994; Dougherty, Gettys & Ogden, 1999; Windschitl & Wells, 1998; Sprenger & Dougherty, 2006). In most cases, the assessment of competing hypotheses necessitates that the competing hypotheses be generated from long-term memory (Dougherty, Gettys, & Thomas, 1997; Dougherty & Hunter, 2003a; Gettys & Fisher, 1979). For example, a physician considering the likelihood that a patient has pneumonia presumably generates relevant competing alternatives to the pneumonia hypothesis prior to rendering a diagnosis (Elstein, Shulman, & Sprafka, 1978; Weber, Böckenbolten, & Hilton, 1993).

One can conceptualize the impact of hypothesis generation on probability judgment within the support theory framework. Tversky and Koehler (1994) proposed that judgments of probability are made by comparing the support for a focal hypothesis (A) with the support for a set of alternative hypotheses (B):

$$P(A,B) = \frac{s(A)}{s(A) + s(B)} \tag{2}$$

where s(A) and s(B) represent the support for A and B respectively, and P(A,B) is the probability of hypothesis A versus hypothesis B occurring. Research suggests that people "unpack" (i.e. generate) the hypotheses before evaluating their support (Dougherty & Hunter, 2003a, b; Dougherty & Sprenger, 2006; Sprenger &

Dougherty, 2006). Consider the case in which one is asked to judge the likelihood that the University of North Carolina will win the Atlantic Coast Conference (ACC) basketball tournament. In support theory terms: P(UNC win, UNC not win) = s(UNC)/ [s(UNC) + s(not UNC)]. People do not simply evaluate support for the packed alternative hypothesis, "not UNC", but rather unpack that hypothesis into other teams in the ACC, such as the University of Maryland and Duke. Judgments have been shown to decrease as the number of alternatives considered increases (Dougherty & Hunter, 2003; Sprenger & Dougherty, 2006). Further, Dougherty and Sprenger (2006) found that under certain conditions, participants consider irrelevant hypotheses, and consequently the accuracy of their judgments decreases. Thus, it appears that judgment accuracy is directly influenced by which and how many hypotheses people consider. Since hypothesis generation underlies people's probability estimation, it is of interest to examine how probability judgments are affected by the order of data presentation.

Revision of opinion research (Rapaport & Wallsten, 1972; Slovic & Lichtenstein, 1971; Fischhoff & Beyth-Marom, 1983) examined people's subjective probability judgments as information was presented sequentially. In a typical revision of opinion paradigm, participants were told that bookbag A contained 70% red poker chips and 30% blue poker chips, and bookbag B contained 70% blue poker chips and 30% red poker chips. Participants were then presented with poker chips picked out of a bookbag, and after viewing each poker chip participants estimated the probability that the bag in question was bookbag A. Peterson and DuCharme (1967) found a primacy effect such that information presented early in a sequence influenced judgments more than information presented later in a sequence. However, most revision of opinion research gave participants only two or three possible hypotheses, so hypothesis generation from long-term memory was not a necessary component of the judgment process.

Sequential effects in probability judgments have been examined in the belief updating literature. Belief updating research examines how beliefs about hypotheses or propositions are updated as new information is presented. In a review of the literature, Hogarth and Einhorn (1992) found that task variables, such as complexity, length, and response mode, affect the type of order effect that obtains. For instance, when tasks are simple (each piece of evidence needs little processing for its comprehension and the task is familiar), short (fewer than 12 pieces of evidence), and responses are required after each new piece of information, people tend to show recency effects. When tasks are simple, short, and responses are required only after viewing all evidence, people tend to show primacy effects. In longer tasks, people tend to show primacy. As is the case in revision of opinion research, most belief updating tasks give participants a small number of hypotheses to compare when updating beliefs, so hypothesis generation from long-term memory may not be a necessary component of the judgment process.

The goal of Experiment 2 was to examine how probability judgments are affected by the order of data presentation. It was hypothesized that probability judgments would show effects similar to those found with hypothesis generation in Experiment 1, because hypothesis generation determines which hypotheses are compared in probability judgment. Experiment 1 found that participants' generation was more affected by diagnostic information when it was presented later in the sequence than when it was presented earlier in the sequence. Therefore it was hypothesized that judgments, being based on the set of hypotheses generated, would similarly be more affected by diagnostic information when it was presented later in the sequence. Thus, it was hypothesized that after the final piece of evidence was presented, participants in cue order groups receiving the most diagnostic information last would give higher probability judgments for the correct hypothesis than participants receiving the most diagnostic information first. Further, it was hypothesized that participants receiving the most diagnostic information last would give lower probability judgments to incorrect hypotheses than participants receiving the most diagnostic information first.

Methods

Participants

University of Maryland undergraduate students (n=128) participated in Experiment 2 for course extra credit. Participants were run individually in single sessions lasting approximately 40 minutes.

Materials

Materials were the same as those used in Experiment 1. The instructions used in the experiment are presented in Appendix D.

Design and Procedure

The design and procedure of Experiment 2 were the same as Experiment 1 except for the following two changes. First, cue presentation format was not manipulated; all participants viewed a single cue on each screen, as in the singular presentation condition of Experiment 1. Second, participants judged the likelihood of hypotheses rather than generating hypotheses. More specifically, after receiving each piece of evidence participants judged the likelihood that a given robber committed that robbery. Participants judged the same robber after each cue of a given crime scene. Participants judged a different robber for each crime scene. The robber judged was counterbalanced across participants and crime scenes. Participants were instructed that a robbery had occurred and that they would judge the likelihood that a particular robber (as compared to other possible robbers) committed that robbery based on pieces of evidence learned at the scene of the robbery. Participants were instructed that a judgment of 0 meant that there was NO CHANCE that the suspect committed the crime and that a judgment of 100 meant that it was ABSOLUTELY CERTAIN that the suspect committed the crime. A judgment of 50 meant that the outcome had the same chance as a coin flip landing on heads rather than tails. Participants were told they could use any number between 0 and 100. After a piece of evidence was presented participants were asked, “Out of all possible robbers, what is the chance that the Black Bandit (Silver Swindler, Purple Pirate, Red Rogue) committed this crime?”

Results

Two general findings from Experiment 2 are of interest. First, judgments were more influenced by recent cues than by initial cues. Second, base rates did not influence probability judgments.

Cue Order Effects

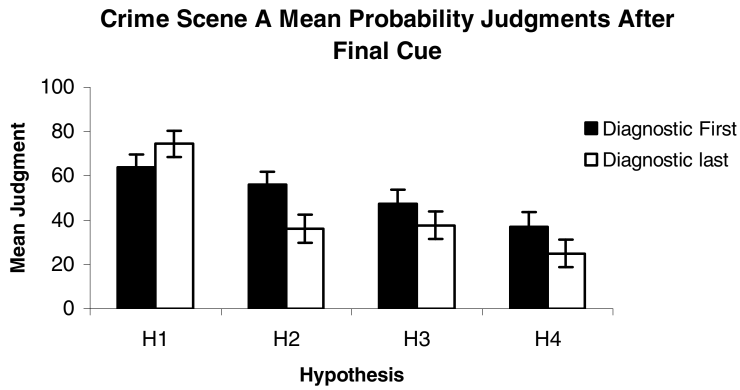
Figure 8 presents mean probability judgments for the two cue order groups (diagnostic information first vs. last) after cue 4 for Crime Scenes A (top panel) and B (bottom panel). The figure shows that in both crime scenes, higher probability judgments were made for the correct hypothesis than for incorrect hypotheses. In both crime scenes, analyses found a main effect of hypothesis (Crime Scene A[13]: $F(3, 120)=13.85$, $p<0.0001$; Crime Scene B[14]: $F(3, 120)=7.71$, $p<0.0001$).

Figure 8 shows a trend that judgments of the "correct" hypothesis tended to be higher when diagnostic information was presented last than when it was presented first. Further, there was a trend that judgments of "incorrect" hypotheses tended to be lower when diagnostic information was presented last than when it was presented first. Analyses found a marginally significant interaction between cue order group and hypothesis for Crime Scene A, $F(3,120)=2.46$, $p=0.066$, and a significant interaction between cue order group and hypothesis for Crime Scene B, $F(3, 120)=3.04$, $p=0.032$. Participants who received diagnostic information as cue 1 tended to weight it less than participants who received diagnostic information last. As a result, they gave lower judgments of the "correct" hypothesis than participants who received the most diagnostic information as cue 4, and gave higher judgments to incorrect hypotheses than when the diagnostic information was presented at cue position 4. These results suggest that recency, more than primacy, affected participants' judgments.

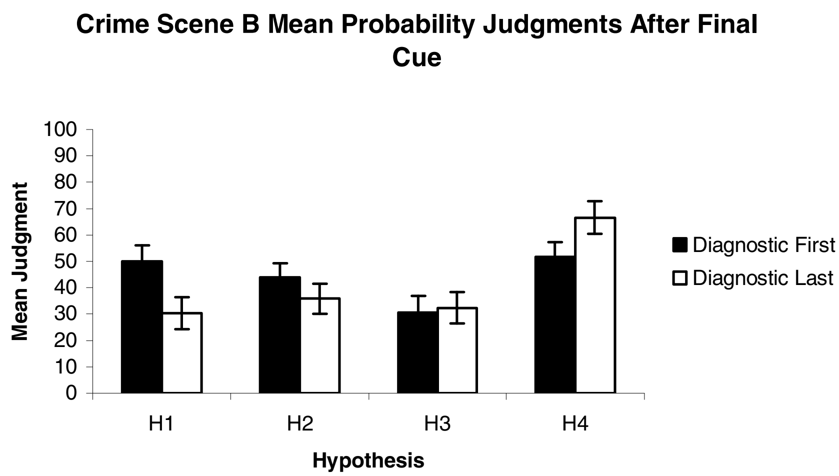
---

[13] Probability judgments for hypothesis 1 (the "correct" hypothesis) were higher than for the other three hypotheses (H1 vs. H2: $p=.0002$; H1 vs. H3: $p<.0001$; H1 vs. H4: $p<.0001$).
[14] Probability judgments for hypothesis 4 (the "correct" hypothesis) were higher than for the other three hypotheses (H4 vs. H1: $p=0.002$; H4 vs. H2: $p=0.001$; H4 vs. H3: $p<0.0001$)

Figure 8



A



B

Figure 8 presents the mean probability judgment of each hypothesis after all pieces of evidence have been presented. Panel A presents judgments for Crime Scene A and panel B presents judgments for Crime Scene B. H1 is the correct hypotheses for Crime Scene A and H4 is the correct hypothesis for Crime Scene B.

Base Rate Effects

Based on the generation results of Experiment 1, it was hypothesized that participants' probability judgments would be sensitive to hypothesis base rates. It was predicted that judgments of hypotheses with higher base rates would be higher than judgments of hypotheses with lower base rates when participants were initially presented with completely non-diagnostic information.

Base rate effects were examined by comparing mean judgments of each hypothesis after participants were given a non-diagnostic cue as their first piece of information. Then, the only information available to influence judgments was the base rates of the hypotheses themselves. Figure 9 presents mean judgments for each hypothesis in each crime scene when the first piece of evidence was non-diagnostic. Analyses found no main effect of hypothesis for either Crime Scene (Crime Scene A, group 1: $F(3, 61)=0.84$, $p>.10$; Crime Scene B, group 2: $F(3, 59)=0.08$, $p>0.10$). In contrast with hypothesis generation, participants' probability judgments were not sensitive to base rates. Participants gave similar judgments to the high, middle, and low base rate hypotheses.

<u>*Discussion*</u>

Experiment 2 extended the results of Experiment 1 by showing that probability judgment, like hypothesis generation, was sensitive to the order that data were presented. When diagnostic information was presented last as opposed to first in a sequence, participants gave higher probability judgments to correct hypotheses and

Figure 9

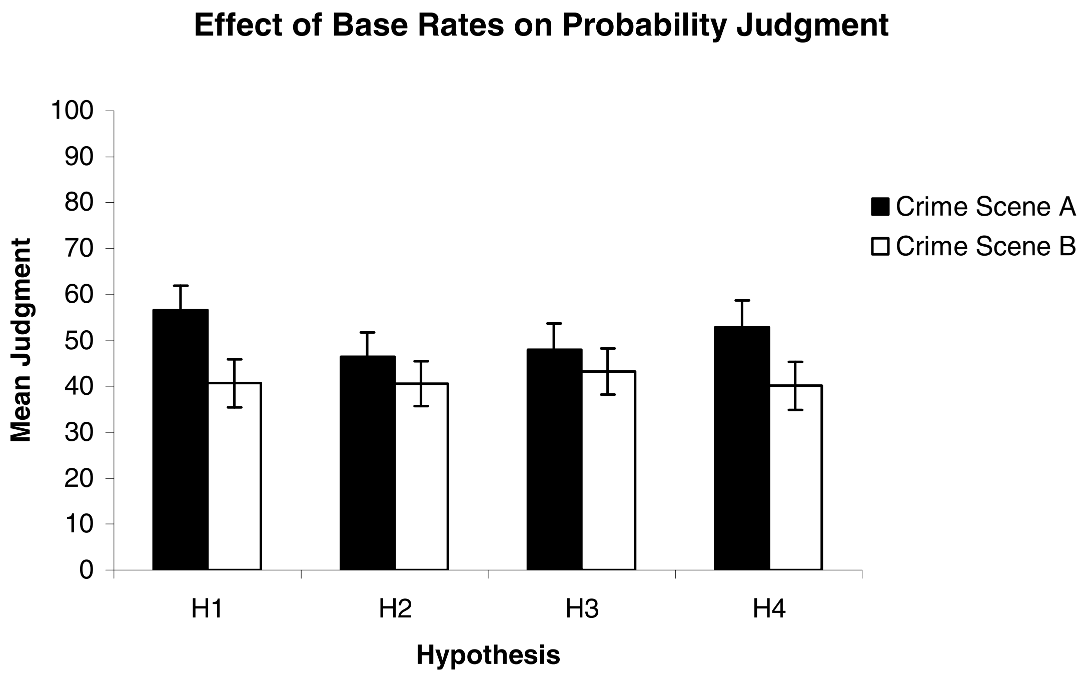


Figure 9 presents the effect of base rates on probability judgment. Mean judgments for each of the four hypotheses are shown after one piece of non-diagnostic information has been presented.

lower judgments to incorrect hypotheses. Although recency obtained for hypothesis generation and probability judgments in Experiments 1 and 2, the finding could be due to the way that responses were elicited. Participants generated hypotheses each time a new cue was presented, rather than only once at the end of the entire sequence of cues. In impression formation, primacy can be changed to recency by manipulations designed to equalize attention to all adjectives (Anderson, 1973). For instance, in one impression-formation experiment, merely asking participants pronounce words aloud produced recency rather than primacy effects (Hendrick & Constantini, 1970). In Experiments 1 and 2, forcing participants to respond after each cue was presented may have led participants to attend each cue more than they otherwise would. Perhaps when participants are asked to generate hypotheses and make judgments only after all cues are presented, primacy would obtain. Hogarth and Einhorn (1992) reviewed the belief updating literatures and found that in 16 out of 16 studies in which simple, short sequences of data were used with a step-by-step (SBS) response mode, participants displayed recency. In the SBS response mode participants responded after viewing each datum. When simple, short sequences of data were used with end-of-sequence (EOS) response modes, primacy was found in 19 out of 27 cases. In the EOS response mode participants responded only after viewing all data. Experiment 3 compared the EOS and SBS response modes, to examine whether primacy obtains in EOS conditions and recency in SBS conditions.

One limitation of Experiments 1 and 2 was that only 4 possible causes (robbers) were possible. In real-world environments in which people generate hypotheses, many possible causes and pieces of data exist. Within the medicine domain Gordon (1970) estimated the number of diseases to be approximately 6,000 and the number of symptoms (data) to be approximately 20,000. It could be argued that because only 4 causes were possible in Experiments 1 and 2, participants did not rely as heavily on long-term memory search as they would when more causes were possible. Rather, in Experiments 1 and 2 participants possibly maintained all 4 hypotheses and all observed data in working memory. Then, participants could simply output a hypothesis when it received enough support to meet a decision threshold without searching for the hypothesis in LTM. Thus Experiment 3 examined hypothesis generation and probability judgment when 8 possible causes and 4 possible data existed to see if the pattern of results was similar to that found in Experiments 1 and 2. Further, Experiment 3 compared SBS and EOS response modes to examine whether response mode affects the type of order effects that obtain.

Chapter 4: Experiment 3

Experiment 3 examined how cue order affects hypothesis generation, and it extended the first two experiments in two ways. First, Experiment 3 examined how participants were affected by cue order when eight causes were possible, rather than only four. Second, Experiment 3 examined whether different response modes resulted in different order effects. As in Experiments 1 and 2, one cue order presented the most diagnostic piece of information early in the cue sequence and the least diagnostic information late in the sequence, and the other cue order presented the least diagnostic information early in the cue sequence and the most diagnostic information late in the sequence. This manipulation shows whether diagnostic information is weighted more heavily when it occurs early or late in the sequence. In a second cue order, a diagnostic cue was presented at two positions in the sequence. An early cue pointed to one cluster of possible causes and a later cue pointed to a different cluster of possible causes. If primacy obtained, participants would generate mostly hypotheses from the cluster pointed to by the early cue. If recency obtained, participants would generate mostly hypotheses pointed to by the most recent cue. Another possibility was that participants would weight all of the cues equally and simply increase the number of hypotheses they considered as they received more cues indicative of other possible clusters. This would also be an interesting result, because it contrasts the findings of Experiment 1 where participants tended to generate fewer hypotheses as they learned more information.

It was previously mentioned that Hogarth and Einhorn (1992) reviewed the belief updating literature and reported that in different kinds of tasks, different order effects obtain. For instance, when people are presented with short sequences of data in simple tasks (meaning that each datum is a single word rather than a description), the order effects that obtain depend on response mode. When tasks require end-of-sequence (EoS) responses, primacy effects obtain. When tasks require step-by-step (SbS) responses, recency effects obtain. To examine whether different kinds of tasks lead to different order effects for hypothesis generation, response mode was varied in Experiment 3. Participants generated hypotheses either in a SbS or an EoS response mode. This experiment had simple data (single word and picture cues), and was short (six cues). It was predicted that if hypothesis generation was affected by similar factors as is belief updating, based on Hogarth and Einhorn's (1992) review primacy would obtain in the EoS response condition and recency would obtain in the SbS response condition.

Methods

Participants

University of Maryland undergraduate students (N=128) participated in this experiment for course extra credit.

Materials

The materials for this experiment were similar to those used in the previous experiments, except that there were eight robbers rather than four. The stimuli, and exact instructions for the entire experiment, are presented in Appendix E. Table 3 presents the frequencies with which each piece of data was associated with each

Table 3

		Data 1: Type of Job	Data 2: location of Job	Data 3: type of getaway vehicle	Data 4: type of cover	
		(bank, jewelry)	**(north, south)**	**(helicopter, sports car)**	**(smoke, mask)**	**Base rate**
Cluster A	**C1**	24,6	24,6	15,15	15,15	30
	C2	8,2	8,2	5,5	5,5	10
Cluster B	**C3**	6, 24	24,6	15,15	15,15	30
	C4	2, 8	8,2	5,5	5,5	10
Cluster C	**C5**	6, 24	6, 24	15,15	15,15	30
	C6	2, 8	2, 8	5,5	5,5	10
Cluster D	**C7**	6, 24	6, 24	15,15	15,15	30
	C8	2, 8	2, 8	5,5	5,5	10
						Total: 160

Table 3 presents an initial frequency distribution for 8 possible causes and 4 data for experiment 3. Clusters represent groups of possible causes that are similar to each other. The labels C1-C8 represent the 8 possible causes. The two numbers in each cell represent the frequency that data value A and B occur respectively for a given possible cause.

possible cause in the training phase of the task. In Table 3, one can see that there were 8 total possible causes and 4 types of data, and each datum had two levels. Further, there were four clusters of possible causes. The possible causes in each cluster were highly similar to each other. Possible causes in clusters A and B were similar to each other and clusters C and D were similar to each other. Data 1 separated possibilities in cluster A from possibilities in clusters B, C, or D. Data 2 separated possibilities in clusters A and B from those in clusters C and D. Data 3 and 4 were non-diagnostic.

Two crime scenes tested the effect of different cue orders on hypothesis generation (see Table 4).

Crime Scene A. Crime Scene A consisted of four pieces of evidence: Data 2 value B, Data 3 value A, Data 4 value A, and data 1 value A. Participants in cue order group 1 first observed Data 2, value B (diagnostic of clusters C and D) then observed two non-diagnostic cues (data 3 and 4) and finally observed data 1, value A (diagnostic of cluster A). Participants in cue order group 2 saw the same cues in the opposite order. In other words, participants in cue order group 1 observed cues pointing first toward possible causes in clusters C and D and later toward possible causes in cluster A, and participants in cue order group 2 observed cues pointing first toward possible causes in cluster A and later toward possible causes in clusters C and D.

Crime Scene B. Crime scene B consisted of four pieces of evidence: data 1, value A, data 2 value A, data 3, value A, and data 4, value A. Participants in cue order 1 observed data 1, value A first (diagnostic of cluster A) and data 2, value A second (diagnostic of clusters A and B). Then participants observed data 3, value A (a non-

Table 4

		Cue Order Group 1	Cue Order Group 2
Crime Scene A			
	Cue 1	Data 2, value B	Data 1, value A
	Cue 2	Data 3, value A	Data 4, value A
	Cue 3	Data 4, value A	Data 3, value A
	Cue 4	Data 1, value A	Data 2, value B
Crime Scene B			
	Cue 1	Data 1, value A	Data 4, value A
	Cue 2	Data 2, value A	Data 3, value A
	Cue 3	Data 3, value A	Data 2, value A
	Cue 4	Data 4, value A	Data 1, value A

Table 4 presents the cues of the three Crime Scenes. Participants in cue order group 2 saw all cues in the opposite order of participants in cue order group 1. In the first crime scene participants saw the a piece of diagnostic information indicative of one cluster of possible causes and later saw a piece of diagnostic information indicative of a second, different cluster of possible causes. In the third crime scene, in cue order group 1 participants saw non-diagnostic cue first and diagnostic cues last. Cue order group 2 saw the reverse order.

diagnostic cue), and then data 4, value A (a non-diagnostic cue).Thus, participants in cue order group 1 observed cues in a diagnostic to non-diagnostic order, and participants in cue order group 2 observed cues in a non-diagnostic to diagnostic order.

Figure 10 presents the objective probabilities of each possible cause after each cue was presented for the two crime scenes. The top panel presents the values for Crime Scene A. Note that after cue 4, the objective probabilities of each possible cause was the same for both cue order conditions, since both groups saw the same pieces of evidence at that point. For the 5,7 to 1 order (see the top, left panel of Figure 10), the first cue pointed toward possibilities 5 and 7 and the last cue pointed toward possibility 1. For the 1 to 5,7 order (see the top, right panel of Figure 10), the first cue pointed toward possibility 1 and the last cue pointed toward possibilities 5 and 7. After all cues were presented, possibilities 1, 5, and 7 had equal objective probabilities. The bottom panel of Figure 10 presents the objective probabilities for Crime Scene B. Here, cues either went in a diagnostic to non-diagnostic order (left figure) or in a non-diagnostic to diagnostic order. The diagnostic information pointed toward possibility 1, which was the most likely possibility after viewing all data.

Design and Procedure

Several factors were manipulated in this experiment.

Order of cues. During the testing phase, participants viewed pieces of evidence and generated hypotheses that they thought accounted for the pieces of evidence observed. In Crime Scene A, diagnostic cues were presented early in the sequence and late in the sequence. The diagnostic cue presented early in the sequence

Figure 10

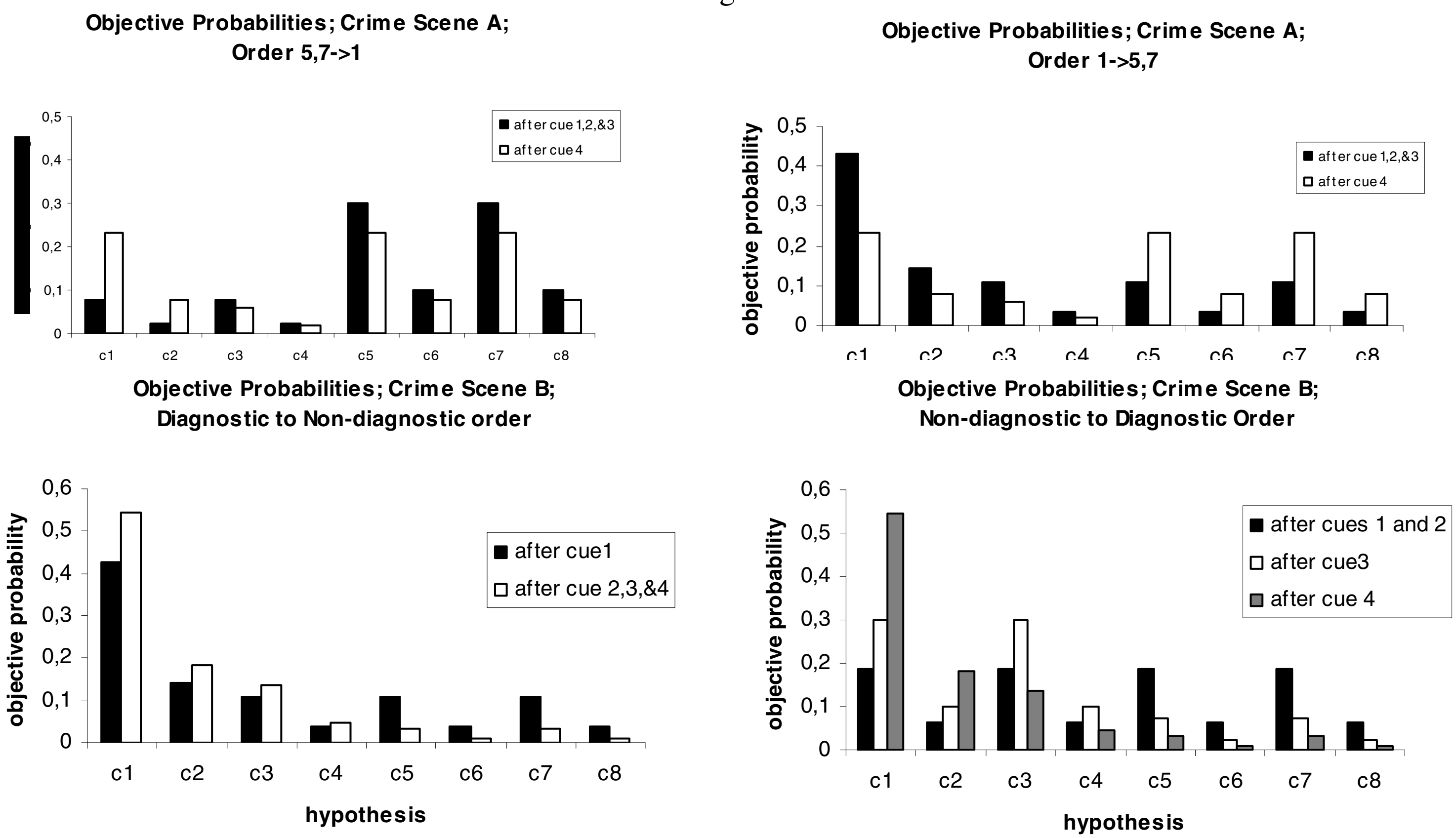


Figure 10 presents the objective probabilities for each possible cause after each cue was presented for the two cue order conditions. The top panel presents objective probabilities for Crime Scene A, and the bottom panel presents objective probabilities for Crime Scene B. In Crime Scene A, one cue order presented a cue supporting possibilities 5 and 7 first and a cue supporting possibility 1 last (left side) and in the other cue order a cue supporting possibility 1 was presented first and a cue supporting possibilities 5 and 7 was presented last (right side). In Crime Scene B, one cue order presented a diagnostic piece of evidence first and a non-diagnostic piece of evidence last (left side) and in the other cue order a non-diagnostic piece of evidence was presented first and a diagnostic piece of evidence was presented last. The evidence in Crime Scene A pointed to H1, H5, and H7 and the evidence in Crime Scene B pointed to H1.

pointed to a different cluster of possible causes than the diagnostic cue presented later in the sequence. In Crime Scene B, cues either went in a diagnostic to non-diagnostic order or vice versa. As in the previous experiments, the order of cues was varied such that participants in one cue order group saw cues in an opposite order than participants in the second cue order group.

Possible Causes. Eight causes (robbers) were possible.

Response mode. In Experiment 3 participants either generated hypotheses in a step by step (SbS) mode or in an end of sequence (EoS) mode. In the SbS mode, participants generated hypotheses after each piece of evidence was presented. In the EoS mode, participants generated hypotheses only after all evidence was presented.

As in Experiments 1 and 2, Experiment 3 consisted of two main phases, a learning phase and a generation phase. The learning phase was identical to that of Experiments 1 and 2, except that participants learned about eight robbers rather than four robbers, and each datum had two levels rather than 2, 3, or 4 levels. The testing phase of the experiment was also similar to that of Experiments 1 and 2, except that response mode was manipulated. Participants in the step-by-step response mode condition generated hypotheses each time a new cue was presented. Participants in the end-of-sequence response mode condition generated hypotheses only after all four cues were presented. Also, participants judged two hypotheses after all cues were presented. For Crime Scene A, participants evaluated hypotheses 1 and 7. For Crime Scene B, participants evaluated hypotheses 1 and 5. Participants were asked, "Based on the clues you have seen, how likely is it that the robber who committed this crime was: Rob?" A picture of the robber in question was displayed on the lower half of the screen. As in Experiment 2, participants responded by typing a number between 0 and 100 into a textbox.

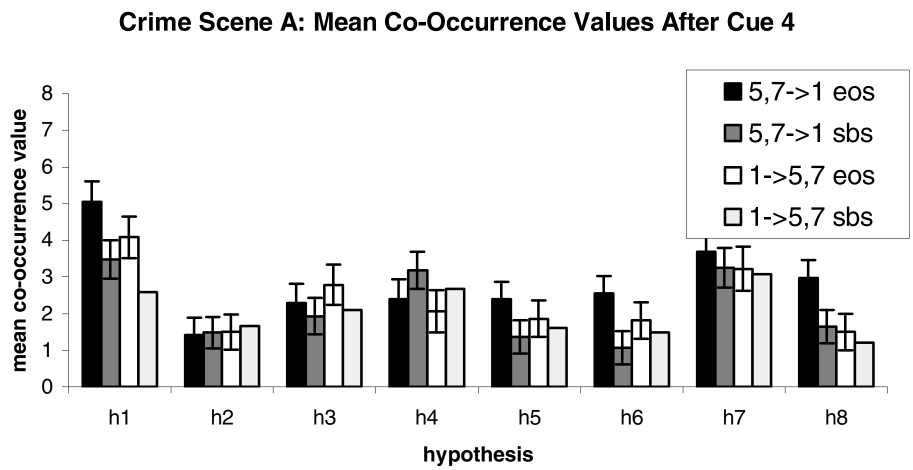
Results

Cue Order Effects

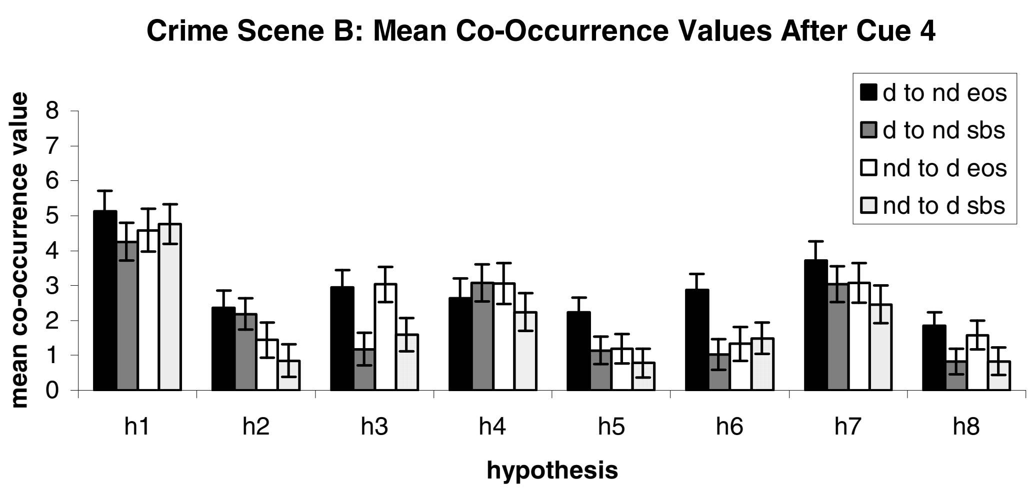
In Crime Scene A, I examined how generation differed when early information supported possible causes 5 and 7 and late information supported possible cause 1 versus when early information supported possible cause 1 and late information supported possible causes 5 and 7. In Crime Scene B, I examined how generation differed when diagnostic information was presented first versus when it was presented last. For both crime scenes, if participants weighted all of the cues equally, generation after cue 4 would not differ based on the order that cues were presented. If participants tended to use recent cues more than earlier cues, generation after cue 4 should be more affected by cue 4. If participants tended to use early cues more than recent cues, participants would be most affected by cue 1. Further, it was hypothesized that the type of order effect obtained would depend on the response mode; it was hypothesized that recency would obtain in the SBS condition, and primacy would obtain in the EOS response mode. Thus, mean co-occurrence values for the eight hypotheses were examined after cue 4 was presented, and the independent variables of cue order (whether the cue pointing toward possible causes 5 and 7 was presented as cue 4 or as cue 1) and response mode (SBS vs. EOS) were compared.

Figure 11, the top panel presents mean co-occurrence values for the two cue order groups after cue 4 for Crime Scene A, and the bottom panel presents the same

Figure 11



A



B

Figure 11 presents mean co-occurrence values after all pieces of evidence have been presented. Panel A presents Crime Scene A, and Panel B presents Crime Scene B. In Crime Scene A the correct hypotheses were H1, H5, and H7. In Crime Scene B the correct hypothesis was H1.

information for Crime Scene B. For Crime Scene A, Figure 11 shows that higher co-occurrence values were found for hypotheses 1 and 7 than for other hypotheses. For Crime Scene B, the figure shows that again, hypotheses 1 and 7 had higher co-occurrence values than other hypotheses. In both crime scenes, analyses found a main effect of hypothesis, (Crime Scene A[15]: F(7, 1085)=10.65, p;0.0001; Crime Scene B[16]: F(7, 1085)=21.72, p<0.0001). In Crime Scene A, it was hypothesized that hypotheses 1, 5, and 7 would have higher co-occurrence values than the other hypotheses, since these three hypotheses were most likely after all data were observed. However, only hypotheses 1 and 7 had higher co-occurrence values than others. For Crime Scene B, it was hypothesized that hypothesis 1 would have higher co-occurrence values than all other hypotheses, because it had the highest objective probability. However, hypothesis 7 also had relatively high values.

It was predicted that hypotheses would have higher co-occurrence values when the most diagnostic information for those hypotheses came first for the EOS condition, and last for the SBS condition. However, for both Crime Scenes, no three-way interaction between cue order group, response mode, and hypothesis were found, (Crime Scene A: F(7, 1085)<1; Crime Scene B: F(7, 1085)<1). Further, no

[15] For the post hoc comparisons, Bonferroni's adjustment was used. Thus, for each comparison alpha was 0.002. Hypothesis 1 differed significantly from all other hypotheses except hypothesis 7; H2(t(158)=6.13, p<0.0001); H3(t(158)=4.04, p<0.0001); H4(t(158)=3.11, p=0.002); H5(t(158)=5.48, p<0.0001); H6(t(158)=5.60, p<0.0001); H8(t(158)=5.29, p<0.0001). Hypothesis 7 differed significantly from all other hypotheses except hypotheses 1, 3 and 4: H2 (t(158)=-4.85, p<0.0001); H5 (t(158)=-4.51, p<0.0001); H6 (t(158)=-4.39, p<0.0001); H8 (t(158)=4.16, p<0.0001). Further, hypothesis 4 differed significantly from hypothesis 6 (t(158)=2.50, p=0.013).

[16] For the post hoc comparisons, Bonferroni's adjustment was used. Thus, for each comparison alpha was 0.002. Hypothesis 1 differed significantly from all other hypotheses; H2(t(158)=7.34, p<0.0001); H3(t(158)=6.56, p<0.0001); H4(t(158)=4.66, p<0.0001); H5(t(158)=9.38, p<0.0001); H6(t(158)=7.62, p<0.0001); H7(t(158)=3.85, p=0.0002); H8(t(158)=9.80, p<0.0001). Hypothesis 7 differed significantly from all other hypotheses except hypotheses 3 and 4; H2(t(158)=-3.61, p=0.0004); H5(t(158)=-5.13, p<0.0001); H6(t(158)=-4.08, p<0.0001); H8(t(158)=5.57, p<0.0001). Further, hypothesis 4 differed significantly from hypothesis 5 (t(158)=4.14, p<0.0001) and hypothesis 8 (t(158)=4.51, p<0.0001).

interactions were found between cue order group and hypothesis, (Crime Scene A: F(7, 1085)<1; Crime Scene B: F(7, 1085)<1), or between response mode and hypothesis, (Crime Scene A: F(7, 1085)=1.82, p>0.05: Crime Scene B: F(7, 1085)<1).

Base Rate Effects

To test the effect of base rates on generation, differences in generation were examined after participants were given a non-diagnostic cue as their first piece of information. Then, the only information available to aid generation was the base rates of the possible causes themselves. Thus, if participants were sensitive to differences in base rates, high base rate items would be generated more often and in closer succession after the cue was presented. In Crime Scene B, the SBS response mode, and the ND to D cue order, participants received a non-diagnostic piece of information as cue one. Figure 12 presents mean co-occurrence values for each hypothesis when the first piece of evidence was non-diagnostic. One can see that co-occurrence values were higher for some high base rate hypotheses (H1 and H7) but not for other high base rate hypotheses (H3 and H5). Further, one low base rate hypothesis (H4) received a high mean co-occurrence value. Analyses found a main effect of hypothesis[17] F(7, 280)=7.24, p<0.0001. However, the effect was not straight forward. It was not the case that all high base rate hypotheses had the highest co-

[17] For the post hoc comparisons, Bonferroni's adjustment was used. Thus, for each comparison alpha was set at 0.002. Hypothesis 1 differed significantly from hypothesis 8 (t(40)=3.76, p=0.0005). Hypothesis 4 differed significantly from hypothesis 5 (t(40)=3.73, p=0.0006) and hypothesis 8 (t(40)=4.98, p<0.0001). Hypothesis 7 differed significantly from all other hypotheses except hypothesis 4; H2 (t(40)=-3.77, p=0.0005); H3 (t(40)=-3.63, p=0.0008); H5 (t(40)=-5.15, p<0.0001); H6(t(40)=-3.70, p=0.0007); H8(t(40)=5.87, p<0.0001).

Figure 12

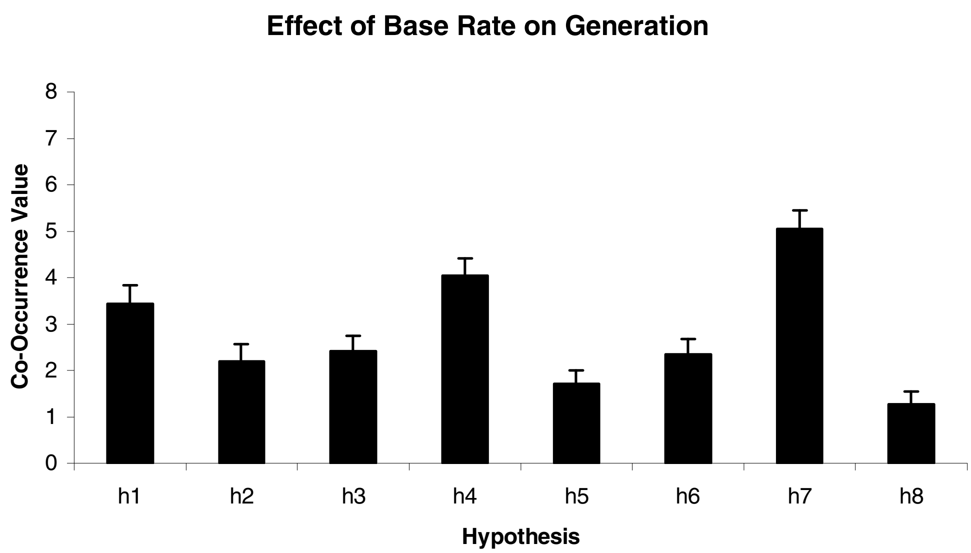


Figure 12 presents the effect of base rate on hypothesis generation. Hypotheses 1, 3, 5, and 7 were the high base rate hypotheses.

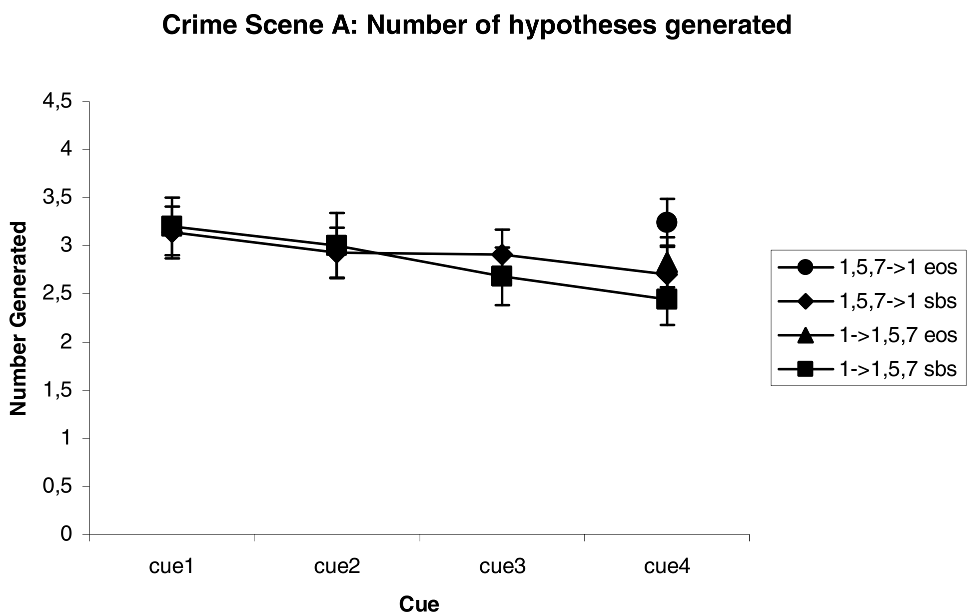
occurrence values, and not the case that all low base rate hypotheses had the lowest co-occurrence values.

Number Generated

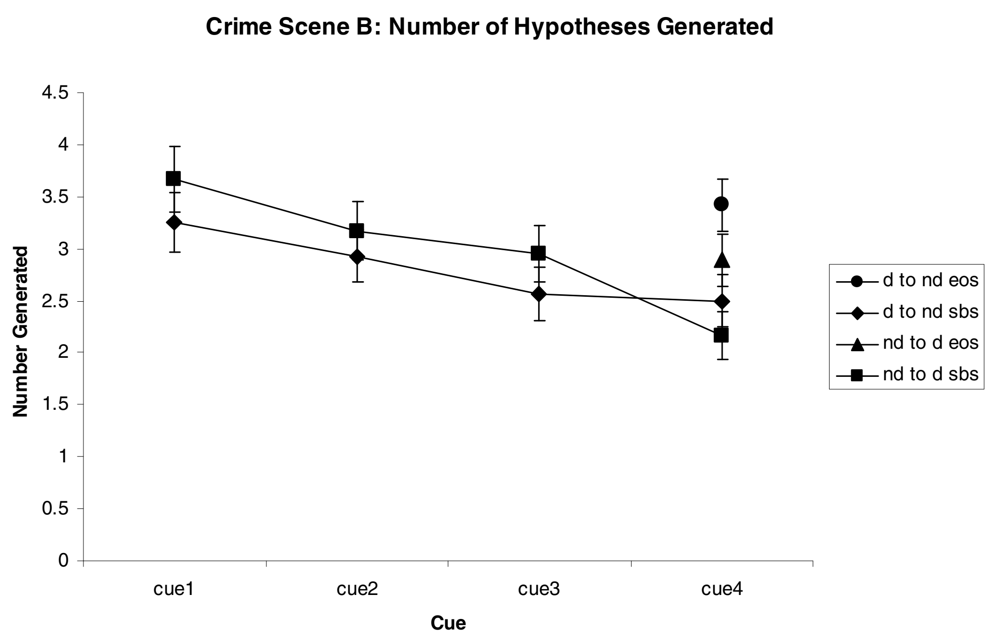
Figure 13 presents the mean number of hypotheses generated after each cue was presented for each Crime Scene. The top figure presents Crime Scene A and the bottom figure presents Crime Scene B. Several general conclusions can be reached from examining how many hypotheses people generated after each cue was presented: First, although there were eight possible causes, participants generated on average between two and four hypotheses. This is consistent with previous literature that has found that participants generate around four hypotheses on average (Weber et al., 1993; Joseph & Patel, 1990; Barrows et al., 1982; Elstein et al., 1978; Mehle, 1982). Second, in the step-by-step condition, participants generated more hypotheses after an initial cue was presented and generated fewer and fewer hypotheses after other cues were provided, and this was true in both cases (main effect of cue position: Crime Scene A[18]: $F(3, 249)=4.72$, $p=0.003$; Crime Scene B[19]: $F(3, 249)=17.98$, $p<0.0001$). Third, as in Experiment 1, participants tended to generate more hypotheses after the first cue when it was non-diagnostic than when the first cue was diagnostic (in the Step-by-Step condition in Crime Scene B, interaction between cue position and cue order group: $F(3, 249)=2.39$, $p=0.069$). Although this interaction

[18] Participants generated more hypotheses: after cue 1 than after cue 3, $t(127)=4.44$, $p<0.0001$; after cue 1 than after cue 4, $t(127)=9.02$, $p<0.0001$; after cue 2 than after cue 3, $t(127)=3.79$, $p=0.0002$; after cue 2 than after cue 4, $t(127)=10.18$, $p<0.0001$; and after cue 3 than after cue 4, $t(127)=8.42$, $p<0.0001$.
[19] Participants generated more hypotheses after: cue 1 than after cue 2, $t(127)=3.86$, $p=0.0002$; after cue 1 than after cue 3, $t(127)=4.72$, $p<0.0001$; after cue 1 than after cue 4, $t(127)=7.51$, $p<0.0001$; after cue 2 than after cue 3, $t(127)=2.25$, $p=0.026$; after cue 2 than after cue 4, $t(127)=5.65$, $p<0.0001$; and after cue 3 than after cue 4, $t(127)=3.93$, $p=0.0001$.

Figure 13



A



B

This figure presents the mean number of hypotheses generated after each cue was presented for Crime Scene A (Panel A) and Crime Scene B (Panel B).

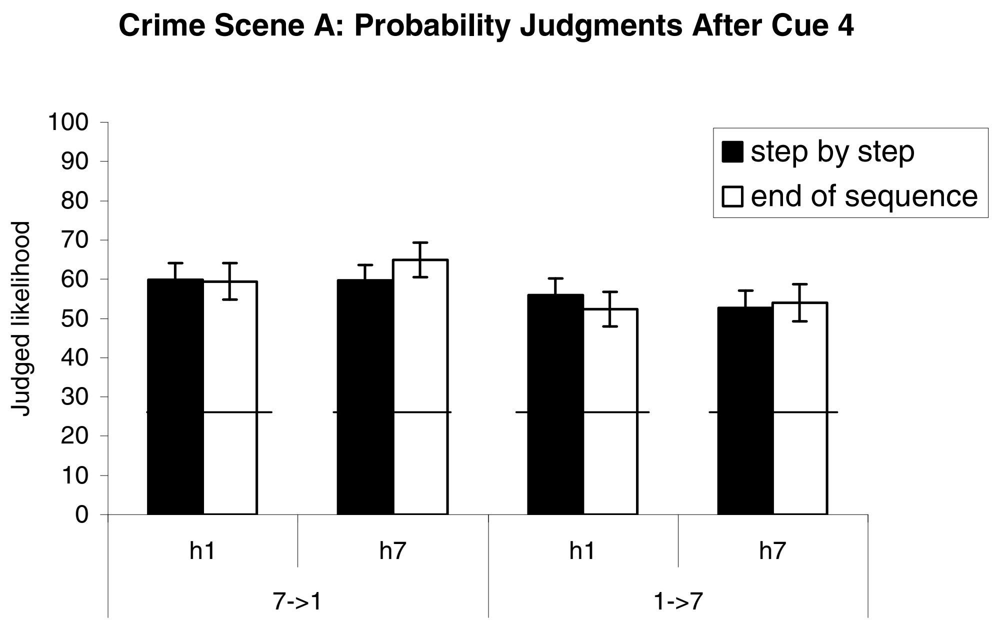
was non-significant, the trend was the same as the trend found in Experiment 1. Further, there were fewer participants in this analysis than in Experiment 1 due to only half of all participants being in the SBS condition.

I examined the effect of response mode and cue order on the number of hypotheses generated *after* cue 4. There was a tendency for participants in the end-of-sequence condition to generate more hypotheses than participants in the step-by-step condition. There was a significant main effect of response mode for Crime Scene B, $F(1, 155)=11.07$, $p=0.001$, and a marginally significant main effect of response mode for Crime Scene A, $F(1, 155)=2.91$, $p=0.090$. In both cases, participants generated more hypotheses in the EOS condition than in the SBS condition. Perhaps participants in the SBS condition more actively weeded out hypotheses after viewing each cue, since they had to focus on each cue to generate hypotheses. For both crime scenes there was no main effect of cue order, and no interaction between cue order and response mode.

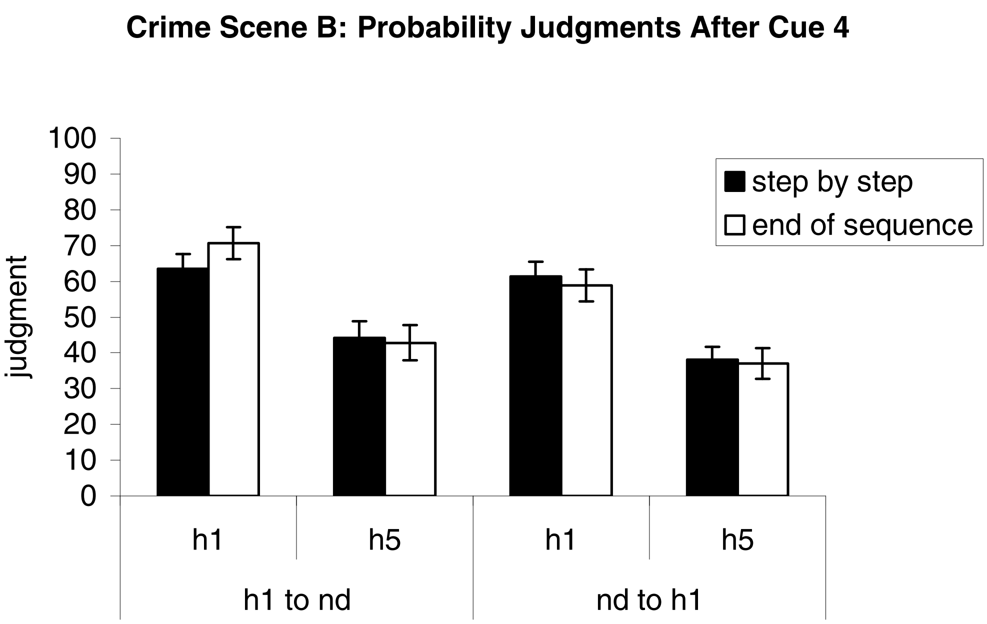
Judgments

As with generation, it was predicted that judgments would be affected more by recent information in the SBS condition and more by early information in EOS condition. Figure 14 presents mean probability judgments for each hypothesis as a function of cue order and response mode. For both crime scenes no interactions were found between hypothesis, cue order, and response mode (Crime Scene A: $F(1, 153)<1$; Crime Scene B: $F(1, 154)<1$). Further, no interactions were found between hypothesis and cue order (Crime Scene A: $F(1, 153)<1$; Crime Scene B: $F(1, 154)<1$) or between hypothesis and response mode (Crime Scene A: $F(1, 153)=1.06$, $p>0.05$;

Figure 14



A



B

Figure 14 presents participants' mean probability judgments for each hypothesis after all cues were presented for Crime Scene A (Panel A) and Crime Scene B (Panel B).

Crime Scene B: $F(1, 154)<1$). In both crime scenes, there was a significant main effect of cue order, (Crime scene A: $F(1, 153)=4.09$, $p=0.045$; Crime Scene B: $F(1, 154)=3.98$, $p=0.048$). Participants tended to give higher judgments in the step-by-step response mode condition than in the end-of-sequence response mode. For Crime Scene B, participants were sensitive to the differences in objective probabilities for hypotheses 1 and 5, as indicated by a significant main effect of hypothesis, $F(1, 154)=86.11$, $p<0.0001$.

Discussion

Experiment 3 examined how hypothesis generation and probability judgment were affected by the presentation order of data when eight causes were possible (rather than only four in Experiments 1 and 2). Further, Experiment 3 examined whether response mode affects the type of order effects that obtain on response mode. Unlike Experiment 1, Experiment 3 found no evidence of order effects on hypothesis generation. Further, Experiment 3 found no indication that response mode affects the type of order effect that obtains. Why were no order-effects found for this experiment? Perhaps people were unable to learn the cue diagnosticities. Experiment 3 was more complex than the previous two experiments, in that participants had to learn about eight robbers rather than just four. In order to maintain attention throughout the learning phase, participants viewed only 160 total crimes. To achieve this goal, the high base rate hypotheses were shown only 30 times, and the low base rate hypotheses were shown 10 times each. In contrast, in Experiment 1 the high base rate hypothesis was shown 80 times, and the low base rate hypothesis was shown 20 times. It is possible that seeing a hypothesis only 30 times is not enough to learn which cues are predictive of that hypothesis. Support for this explanation comes from the finding that in Experiment 1 all significant order effects occurred with Crime Scene A, when cues pointed toward the high base rate hypotheses. In Crime Scene B, cues pointed toward the low base rate hypothesis and no significant order effects obtained. Thus it seems that when base rates were lower, participants could not learn the relationships between those low base rate hypotheses and data as well and thus their generation was then not affected by the presentation order of cues.

As in Experiment 1, participants tended to reduce the number of hypotheses they generated as they received more and more information. This was true even when the most diagnostic pieces of information were presented early in the sequence. Perhaps participants assumed that additional information should rule out hypotheses, even when the new information was non-diagnostic. Note, though, that participants narrowed their hypothesis set more when later information was diagnostic than when later information was non-diagnostic. A second possibility is that participants' motivation to generate hypotheses decreases as the number of times they are asked to generate increases. An interesting finding was that after cue 4, participants in the EOS condition generated more hypotheses than did participants in the SBS condition. When participants were asked to generate after viewing each new piece of evidence, they more actively narrowed their hypothesis sets than when participants only responded after viewing all pieces of evidence.

Chapter 5: Experiment 4

Experiment 3 found no indication of order effects and no differences in order effects as a function of response mode. However, it was unclear whether the null results obtained because order effects do not exist under the conditions present in Experiment 3 or because order effects could not be detected due to methodological flaws. Perhaps the order effects found in Experiment 1 were due to the small number of causes possible. Then, the null findings of Experiment 3 could reflect a true lack of order effects when many causes are possible. Alternatively, the learning phase in Experiment 3 may not have been long enough for participants to learn relationships between cues and possible causes. Perhaps without enough knowledge of these relationships order effects do not obtain because participants' generation is impaired. Thus, a fourth experiment was conducted to again examine order effects on hypothesis generation when many causes were possible. In Experiment 4, rather than having participants learn the relationships between possible causes and data in the laboratory, participants relied on previously stored knowledge about relationships between possible causes and data to generate hypotheses. Then, rather than ensuring that all participants received equal learning a priori, participants' domain knowledge was measured a posteriori. Again, response mode was manipulated to examine whether the type of order effect that obtained depended on the way participants were asked to generate hypotheses.

In Experiment 4, participants observed a series of words found in a description of an undergraduate course offered by the University of Maryland psychology department. Participants were asked to list all of the courses that they thought contained the observed words in their course descriptions. As in Experiment 3, two types of generation task were shown to participants. In the first, participants received non-diagnostic cues early in the sequence and more diagnostic cues later in the sequence (or vice versa). As an example, the word "psychology" was considered a non-diagnostic cue because most psychology courses have that word in their descriptions. On the other hand, "abnormal" was considered a diagnostic cue in that only clinical and counseling courses tended to have that word in their course descriptions. One group of participants received the cue "psychology" early in the sequence and the cue "abnormal" late in the sequence. The other group of participants received the cue "abnormal" early in the sequence and the cue "psychology" late in the sequence. In a second generation task, participants received cues consistent with one cluster of possible causes early in the sequence and cues consistent with a different cluster of possible causes later in the sequence. One group of participants received cues pointing toward social psychology courses early in the sequence and cues pointing toward cognitive neuroscience courses late in the sequence. The other group of participants received cues pointing toward cognitive neuroscience courses early in the sequence and cues pointing toward social psychology courses late in the sequence. Participants' experience with the knowledge domain was assessed by examining which psychology courses they had taken. Then, it was possible to examine whether experience with the knowledge domain was related to order effects in hypothesis generation. Finally, at the end of Experiment 4 participants were given a surprise recall task in which they were asked to recall as many cues from the most recent generation task as they could remember. Fisher (1987) also used a surprise recall task to examine how data presented serially were differentially weighted in a hypothesis generation task. Fisher assumed that participants would best recall data that they had "used" most during generation. Similarly a recall task was included in Experiment 4 to gain an initial assessment of how cues were weighted in the generation task. In this experiment, many possible causes existed, as there were at least 66 psychology courses at the university where the participants attended.

It was hypothesized that as in Experiments 1 and 2, participants' hypothesis generation would be most affected by recent information in the Step-by-Step condition. Based on Hogarth and Einhorn's (1992) review of belief updating literature, it was hypothesized that participants' hypothesis generation would be most affected by early information in the End of Sequence response condition.

Methods

Participants

University of Maryland undergraduate students (N=179) participated in Experiment 4 for course extra credit.

Materials

The materials for this experiment consisted of words describing a psychology course. Words were selected in the following way. First, course descriptions for 64 psychology courses offered at the University of Maryland were submitted to Latent Semantic Analysis (LSA). LSA is a mathematical/statistical technique for extracting and representing the similarity of meaning of words and passages by analysis of large bodies of text. It uses singular value decomposition, a general form of factor analysis, to condense a very large matrix of word-by-context data into a much smaller, but still large-typically 100-500 dimensional-representation (Deerwester, Dumais, Furnas, Landauer & Harshman, 1990). The resulting similarity matrix was then submitted to cluster analysis, to determine how courses clustered together. Then, topics for each cluster of courses were derived by submitting course descriptions within each cluster of courses to Steyvers, Griffiths, & Dennis's (2006) Topics model. Topics models are based upon the idea that documents are mixtures of topics, where a topic is a probability distribution over words. A topic model is a generative model for documents: it specifies a simple probabilistic procedure by which documents can be generated. To make a new document, one chooses a distribution over topics. Then, for each word in that document, one chooses a topic at random according to this distribution, and draws a word from that topic. Standard statistical techniques can be used to invert this process, inferring the set of topics that were responsible for generating a collection of documents. Stimulus-words for experiment 4 were topic-words selected based on their diagnosticities (the degree to which they occurred only in one cluster of courses vs. in many clusters of courses).

Eighteen words were used as cues for two generation tasks. For the first generation task, the words used were: theory, psychology, behavior, assessment, abnormal, diagnosis, disorder, drugs, and treatment. The words "theory", "psychology", and "behavior" were non-diagnostic, in that many course descriptions contained those words. The other words were diagnostic of the cluster of clinical and counseling types of courses (such as "Abnormal Psychology" and "Introduction to Clinical Psychology"). For the second generation task, the cue words were:

relationship, negotiation, communication, helping, research, neural, memory, thinking, and brain. The words “relationship”, “negotiation”, “communication”, and “helping” were diagnostic of the cluster of social psychology courses (such as “Introduction to Social Psychology”, “Communication and Persuasion”, and “Interpersonal Relationships”). The words “neural”, “memory”, “thinking”, and “brain” were diagnostic of the cluster of cognitive and neuroscience courses (such as “Introduction to Memory and Cognition”, “Biological Bases of Behavior”, and “Developmental Biopsychology”). The exact instructions for the entire experiment are presented in Appendix F.

Design and Procedure

Two factors were manipulated in this experiment.

Order of cues. As in the previous experiments, the order of cues was manipulated such that participants in one cue order group saw cues in an opposite order than participants in the second cue order group. In Generation Task 1, cues were ordered either from diagnostic to non-diagnostic (D to ND) or vice versa (ND to D). In Generation Task 2, in one ordering diagnostic cues early in the sequence pointed toward social psychology courses and cues late in the sequence pointed toward cognitive neuroscience courses (S to CN ordering), and in the second ordering the cues were reversed (CN to S ordering). All participants completed both generation tasks, and the order of task was counterbalanced. Cue order was manipulated between participants.

Response mode. Participants either generated hypotheses in a step by step (SbS) mode or in an end of sequence (EoS) mode. In the SbS mode, participants generated hypotheses after each piece of evidence was presented. In the EoS mode, participants generated hypotheses only after all evidence was presented. Response mode was manipulated between participants.

The entire experiment consisted of four main tasks. First, participants completed two generation tasks. Second, after each generation task, participants judged the likelihood of two courses. Third, participants recalled the cue words from the second generation task. Finally, participants indicated which courses they had taken.

Hypothesis Generation Tasks. All participants were instructed to guess which undergraduate psychology course(s) were described by words presented on the computer screen. Each word was displayed individually for three seconds on the computer screen. Participants were told that the words they saw were selected from undergraduate psychology course descriptions. Further, participants were told that they would be presented with a sequence of these course description words, one after another. Participants were asked to consider all of the words they saw when coming up with their list of possible courses. In the end of sequence condition, participants were instructed that *after seeing all of the words*, they would be asked to list the courses they thought were best described by the words they saw. Participants in the step-by-step condition were instructed that *after seeing each word*, they would be asked to list the courses that they thought were best described by the words they had seen. Participants were asked to consider all observed words when coming up with possible courses.

Judgment Task. After completing each generation task, participants made likelihood estimates for two hypotheses. One hypothesis was more consistent with words presented early in the sequence and the other hypothesis was more consistent with words presented later in the sequence. In the first generation task, the two hypotheses judged were: "Introduction to Psychology" and "Abnormal Psychology". In the second generation task, the two hypotheses judged were: "Introduction to Memory and Cognition" and "Basic Helping Skills". Participants were asked, "Based on the words you have seen, how likely is it that the course being described is: Introduction to Memory and Cognition (Basic Helping Skills, Introduction to Psychology, Abnormal Psychology)?" Participants responded by clicking on a 10 point verbal scale ranging from "impossible" to "certain".

Recall Task. After completing both generation and judgment tasks, participants were given a surprise recall task. This task was intended to provide some preliminary indication of which cues participants attended most during the generation task. Participants were asked, "Recall the nine words from the last course description. Type all of the words that you can remember in the space below." Participants typed their responses into a blank textbox.

Experience measure. The experience measure provided information about participants' experience with psychology courses at the University of Maryland. Participants viewed a series of 64 psychology courses offered at the University of Maryland, and for each course participants indicated whether they: had completed the course, were currently enrolled in the course, had started but dropped out of the course, or had never taken the course. Participants were instructed, "We are interested in which psychology courses you have taken. In the following series, you will see the names of psychology courses. For each, please indicate "yes" if you have completed the course (or an equivalent course); "no" if you have not taken the course; "currently enrolled" if you are currently taking the course; and "did not finish" if you started but did not finish the course." Participants also listed how many credits they had completed prior to the current semester, and participants listed their major(s) of study.

Results

Data Analysis

Courses generated fell into one of 10 clusters: cognitive, neuropsychology, developmental, clinical/counseling, social, industrial/organizational, introduction to psychology, research, individual differences, and other. Appendix G lists all of the courses within each cluster. Participants were divided into three experience groups based on the number of psychology courses they had taken. Participants who had not yet completed any courses, but were currently enrolled in one course were classified as having low experience (n=60). Participants who had taken between 1 and 4 courses were classified as having medium experience (n=60). Participants who had taken between 5 and 16 courses were classified as having high experience (n=59).

Cue Order Effects

The first question of interest was whether hypothesis generation differed when diagnostic information was presented first versus last in Generation Task 1. If participants weighted all of the cues equally, generation after all cues were presented would not differ for participants who received the most diagnostic information early in the sequence versus those participants who received the most diagnostic information late in the sequence. In contrast, if participants tended to use recent cues more than early cues, generation after the final cue would be more affected by the diagnostic cue when it was presented late in the sequence than when it was presented early in the sequence. If participants tended to use early cues more than recent ones, they would be most affected by diagnostic information when it was presented early in the sequence. Thus, I examined the number of hypotheses generated in each cluster after cue 9 (the final cue) was presented, and compared cue order (whether the most diagnostic information was presented early or late in the sequence), response mode (whether participants generated hypotheses after each cue or only after all cues were presented), and experience (the number of psychology courses participants had taken). Figure 15 presents the mean number of courses generated in each cluster for the two cue order groups (diagnostic information first vs. last) and the two response mode groups (EOS or SBS) after cue 9 for Generation Task 1. The figure shows that participants generated mostly hypotheses from the clinical/counseling cluster and the "other" cluster. Analyses found a main effect of hypothesis cluster[20], $F(9, 1485)=112.94$,

[20] A Bonferroni adjustment was used,. Therefore, alpha for each comparison was set to 0.001. Participants generated more clinical/counseling than: cognitive courses, $t(177)=-11.80$, $p<0.0001$; bio/neuropsychology courses, $t(177)=-11.91$, $p<0.0001$; developmental courses, $t(177)=-11.32$, $p<0.0001$; social courses, $t(177)=10.44$, $p<0.0001$; industrial/organizational courses, $t(177)=13.65$, $p<0.0001$; introduction courses, $t(177)=6.47$, $p<0.0001$; research courses, $t(177)=12.58$, $p<0.0001$; and individual differences courses, $t(177)=13.00$, $p<0.0001$. Participants generated fewer industrial/organizational than: cognitive courses, $t(177)=3.92$, $p=0.0001$; bio/neuropsychology courses, $t(177)=3.66$, $p=0.0003$; developmental courses, $t(177)=3.51$, $p=0.0006$; social courses, $t(177)=5.09$, $p<0.0001$; "other" courses, $t(177)=-12.49$, $p<0.0001$; individual difference courses, $t(177)=-5.15$, $p<0.0001$. Participants generated more introduction than: cognitive courses, $t(177)=-6.96$, $p<0.0001$; bio/neuropsychology courses, $t(177)=-6.12$, $p<0.0001$; developmental courses, $t(177)=-6.48$, $p<0.0001$; social courses, $t(177)=-5.23$, $p<0.0001$; industrial/organizational courses, $t(177)=-10.76$, $p<0.0001$; research courses, $t(177)=9.06$, $p<0.0001$; and individual differences courses, $t(177)=5.70$, $p<0.0001$. Participants generated more "other" courses than; cognitive courses, $t(177)=-12.30$,

Figure 15

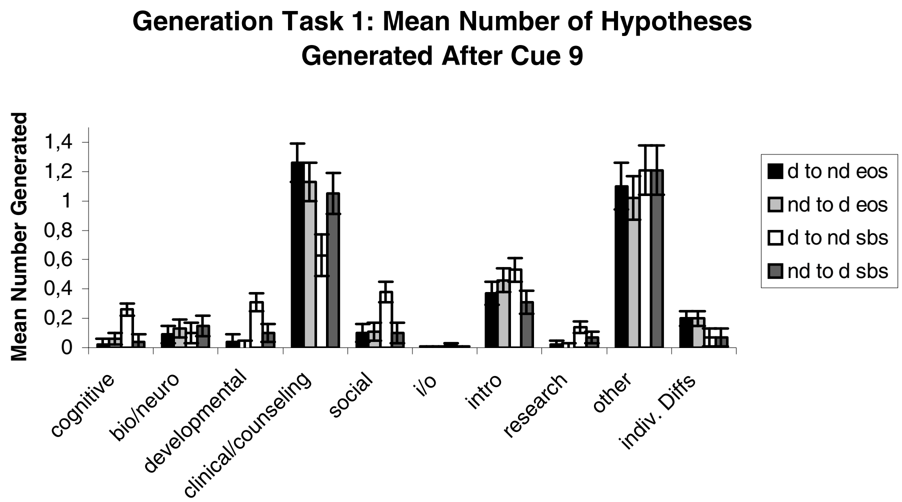


Figure 15 presents the mean number of hypotheses generated from each cluster after the final cue was presented as a function of response type and cue order for Generation Task 1.

$p<0.0001$; bio/neuropsychology courses, $t(177)=-13.27$, $p<0.0001$; developmental courses, $t(177)=-12.58$, $p<0.0001$; social courses, $t(177)=-11.24$, $p<0.0001$; introduction courses, $t(177)=-6.54$, $p<0.0001$; research courses, $t(177)=-12.00$, $p<0.0001$; and individual differences courses, $t(177)=10.23$, $p<0.0001$. Participants generated more social than research courses, $t(177)=3.25$, $p=0.001$.

$p<0.0001$. Moreover, the figure shows a trend that in the SBS condition, participants generated more hypotheses from the "correct" clinical/counseling cluster when diagnostic information was presented late in the sequence rather than early in the sequence. Further, participants in the SBS condition who received diagnostic information first tended to generate more hypotheses from "incorrect clusters" than did participants in the SBS condition who received diagnostic information last. Participants in the EOS condition generated the same number of hypotheses from the correct and incorrect clusters irrespective of cue order. Analyses found a significant interaction between hypothesis cluster, cue order, and response mode, $F(9, 1485)=2.65$, $p=.005$. In general, there were no differences between cue order conditions for the EOS group, but the SBS group tended to generate more hypotheses from the correct (clinical/counseling) cluster and fewer hypotheses from the incorrect clusters when the diagnostic information was presented late in the sequence[21].These results suggest that order effects obtained only in the SBS condition. In the SBS condition, participants were most affected by recent cues, replicating the findings of Experiment 1.

---

[21] For the cognitive cluster, there was no difference in the number generated in the EOS condition, but in the SBS condition participants generated more cognitive courses when the diagnostic information came early in the sequence (M=0.26, SE=0.04) than late in the sequence (M=0.04, SE=0.04), $t(88)=-3.39$, $p=0.0009$. For the developmental cluster, there was no difference in the number generated in the EOS condition, but in the SBS condition participants generated more developmental courses when the diagnostic information came early in the sequence (M=0.31, SE=0.06) than when it came late in the sequence (M=0.10, SE=0.06), $t(88)=-2.63$, $p=0.009$. For the clinical/counseling cluster, there was no difference in the number generated in the EOS condition, but in the SBS condition participants generated more clinical/counseling courses when the diagnostic information came late in the sequence (M=1.05, SE=0.14) than early in the sequence (M=0.63, SE=0.14), $t(88)=2.09$, $p=0.038$. For the social cluster, there was no difference in the number generated in the EOS condition, but in the SBS condition participants generated more social courses when the diagnostic information came early in the sequence (M=0.38, SE=0.07) than late in the sequence (M=0.10, SE=0.07), $t(88)=-2.96$, $p=0.004$. There were no differences in the bio/neuropsychology cluster, the industrial/organizational cluster, the introduction to psychology cluster, the research cluster, the "other" cluster, or the individual differences cluster.

The second question of interest was whether generation differed when early diagnostic information pointed to one cluster of courses and late diagnostic information pointed to a different cluster of courses in Generation Task 2. If one weighted all of the cues equally, generation after all cues were presented should not differ for participants who received the same information in opposite orders. Cues pointed toward the social cluster of courses and toward the cognitive neuroscience cluster of courses. Figure 16 presents the mean number of courses generated in each cluster for the two cue order groups (cognitive neuroscience to social (CN to S) vs. social to cognitive neuroscience (S to CN)) and the two response mode groups (EOS or SBS) after cue 9 for Generation Task 2. The figure shows that participants generated mostly hypotheses from the cognitive, bio/neuropsychology, social, clinical/counseling, introductory psychology, and "other" hypothesis clusters. Analyses found a main effect of hypothesis cluster[22], $F(9, 1485)=88.31$, $p<0.0001$. Moreover, the figure shows that in the SBS condition, participants generated more hypotheses from the cluster pointed to by the most recent cues. Participants in the

---

[22] A Bonferroni adjustment was used,. Therefore, alpha for each comparison was set to 0.001. Participants generated more cognitive courses than: developmental courses, t(177)=6.27.80, p<0.0001; industrial/organizational courses, t(177)=9.64, p<0.0001; research courses, t(177)=8.99, p<0.0001; and individual differences courses, t(177)=8.85, p<0.0001. Participants generated more bio/neuropsychology courses than: developmental courses, t(177)=6.24, p<0.0001; industrial/organizational courses, t(177)=8.62, p<0.0001; research courses, t(177)=8.23, p<0.0001; and individual differences courses, t(177)=8.92, p<0.0001. Participants generated more clinical/counseling courses than: developmental courses, t(177)=-4.00, p<0.0001; industrial/organizational courses, t(177)=-6.17, p<0.0001; research courses, t(177)=5.96, p<0.0001; and individual differences courses, t(177)=6.90, p<0.0001. Participants generated more "other" courses than; cognitive courses, t(177)=-11.48, p<0.0001; bio/neuropsychology courses, t(177)=-11.43, p<0.0001; developmental courses, t(177)=-13.96, p<0.0001; clinical/counseling courses, t(177)=-12.87, p<0.0001; social courses, t(177)=-7.56, p<0.0001, industrial/organizational courses, t(177)=-14.34, p<0.0001; introductory psychology courses, t(177)=-9.14, p<0.0001; research courses, t(177)=-14.12, p<0.0001 and individual differences courses, t(177)=14.35, p<0.0001. Participants generated more social courses than: developmental courses, t(177)=-6.50, p<0.0001; industrial/organizational courses, t(177)=9.47, p<0.0001; research courses, t(177)=8.72, p<0.0001; and individual differences courses, t(177)=8.73, p<0.0001. Participants generated more introductory psychology courses than: developmental courses, t(177)=-6.04, p<0.0001; industrial/organizational courses, t(177)=-9.44, p<0.0001; research courses, t(177)=9.21, p<0.0001, and individual differences courses, t(177)=9.32, p<0.0001.

Figure 16

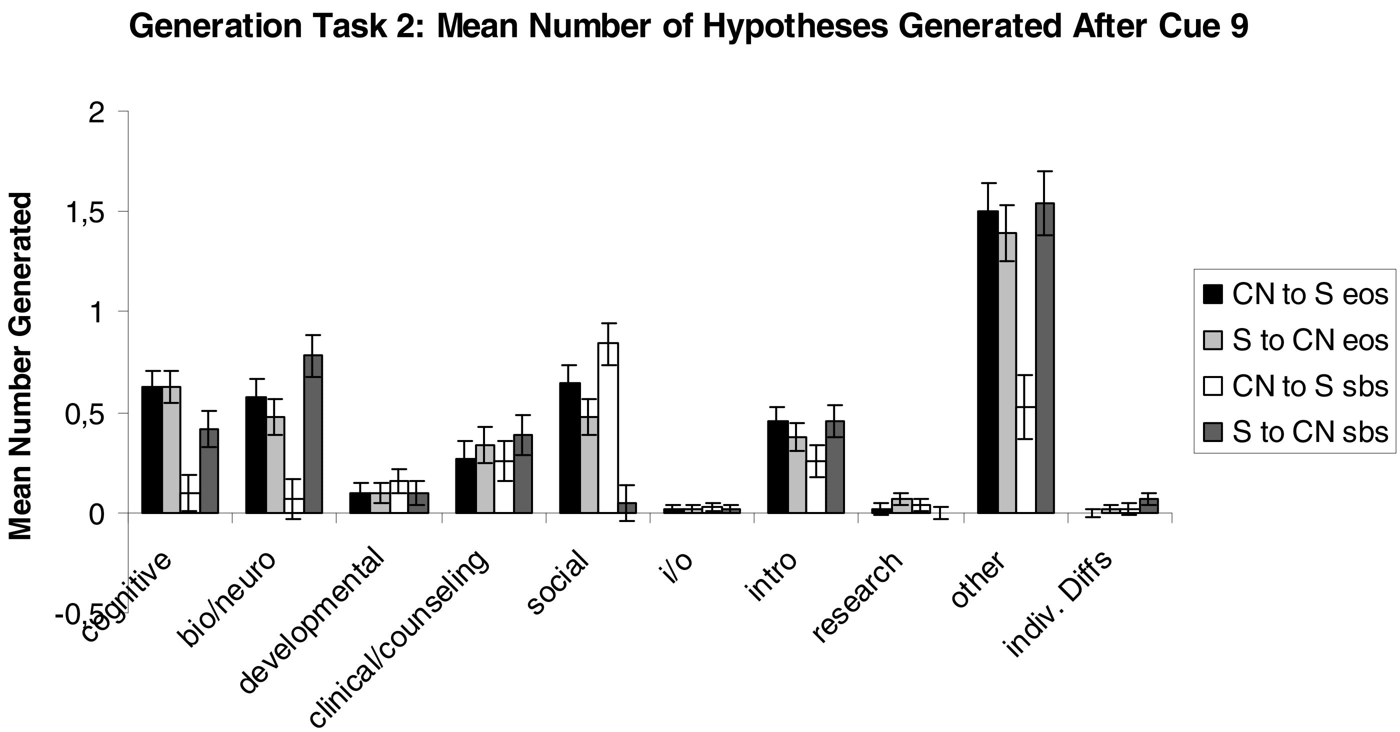


Figure 16 presents the mean number of hypotheses generated from each cluster after the final cue as a function of response type and cue order for Generation Task 2.

SBS condition generated more social courses when later cues pointed toward social psychology courses than when early cues pointed toward social psychology courses. Similarly, participants in the SBS condition generated more cognitive and bio/neuropsychology courses when later cues pointed toward cognitive neuroscience courses than when early cues pointed toward cognitive neuropsychology courses. On the other hand, participants in the EOS condition generated the same number of hypotheses from the correct and incorrect clusters irrespective of cue order. Analyses found a significant interaction between hypothesis cluster, cue order, and response mode, $F(9, 1485)=9.85$, $p<.0001$. In general, there were no differences between cue order conditions for the EOS group, but the SBS group tended to generate more hypotheses from the clusters that the most recent cues pointed toward and fewer hypotheses from the clusters that early cues pointed toward.[23] These results suggest that order effects obtained only in the SBS condition. In that condition, participants were most affected by recent cues.

Analyses also found a cluster by response mode interaction for both generation tasks (Generation Task 1: $F(9, 1485)=4.27$, $p<0.0001$; Generation Task 2:

[23] For the cognitive cluster, there was no difference in the number generated in the EOS condition, but in the SBS condition participants generated more cognitive courses when the cues pointing toward cognitive neuroscience came late in the sequence (M=0.42, SE=0.09) than early in the sequence (M=0.10, SE=0.09), $t(88)=2.45$, $p=0.016$. For the bio/neuropsychology cluster, there was no difference in the number generated in the EOS condition, but in the SBS condition participants generated more bio/neuropsychology courses when the cues pointing toward cognitive neuroscience came late in the sequence (M=0.78, SE=0.19) than when it came early in the sequence (M=0.07, SE=0.10), $t(88)=5.03$, $p<0.0001$. For the social cluster, there was no difference in the number generated in the EOS condition, but in the SBS condition participants generated more social courses when the cues pointing toward social came late in the sequence (M=0.84, SE=0.10) than early in the sequence (M=0.05, SE=0.09), $t(88)=-5.86$, $p<0.0001$. For the "other" cluster, there was no difference in the number generated in the EOS condition, but in the SBS condition participants generated more "other" courses when the cues pointing toward the cognitive/neuropsychology cluster came late in the sequence (M=1.54, SE=0.16) than early in the sequence (M=0.53, SE=0.16), $t(88)=4.49$, $p<0.0001$. There were no differences in the developmental cluster, the clinical/counseling cluster, the industrial/organizational cluster, the introduction to psychology cluster, the research cluster, or the individual differences cluster.

F(9, 1485)=4.33, p<0.0001). For generation task 1, participants in the SBS (M=0.02, SE=0.04) condition generated more hypotheses from the developmental cluster than did participants in the EOS condition[24] (M=0.20, SE=0.04), t(177)=-3.28, p=0.001. For Generation Task 2, participants in the EOS condition (M=0.63, SE=0.06) generated more hypotheses from the cognitive cluster than did participants in the SBS condition (M=0.26, SE=0.06), t(177)=4.25, p<0.0001.

In Generation Task 2, participants in the EOS condition generated more hypotheses than did participants from the SBS condition[25], F(1, 165)=8.62, p=0.004. This finding is consistent with Experiment 3, that participants tended to narrow down their hypothesis set more when in the SBS condition than when in the EOS condition.

Analyses found a significant interaction between cue order and response mode for Generation Task 2, F(1, 165)=7.35, p=0.007. Analyses also found a significant interaction between cluster and cue order for Generation Task 2[26], F(9, 1485)=9.67, p<0.0001. Both of these interactions were qualified by the three-way interaction between cue order, response mode, and cluster described above.

---

[24] Bonferroni adjustments were used so that for the 10 post hoc comparisons, alpha was set to .005 for each.

[25] Participants in the EOS condition generated more hypotheses than participants in the SBS condition for the cognitive cluster, t(177)=4.25, p<0.0001; and the "other" cluster, t(177)=2.73, p=0.007.

[26] For the bio/neuropsychology cluster, participants who received the cues pointing toward the bio/neuropsychology cluster late in the sequence (M=0.63, SE=0.07) generated more hypotheses than participants receiving theses cues early in the sequence (M=0.33, SE=0.07), t(177)=-3.19, p=0.002. In contrast, for the social cluster, participants who received the cues pointing toward the bio/neuropsychology cluster late in the sequence (M=0.26, SE=0.06) generated fewer hypotheses than participants receiving these cues early in the sequence (M=0.74, SE=0.06), t(177)=5.28, p<0.0001. Finally, for "other" cluster, participants who received the cues pointing toward the bio/neuropsychology cluster late in the sequence (M=1.46, SE=0.11) generated more hypotheses than participants receiving theses cues early in the sequence (M=1.02, SE=0.11), t(177)=-2.97, p=0.003.

Experience

Participants with more experience tended to generate more hypotheses than participants with less experience. In both generation tasks, a main effect of experience obtained (Generation Task 1[27]: F(2, 165)=21.29, p<.0001; Generation Task 2[28]: F(2,

[27] For the cognitive cluster, participants with high experience (M=.17, SE=.04) generated more hypotheses than did participants with low experience (M=.03, SE=.04), t(117)=-2.48, p=0.014. For the bio/neuropsychology cluster, participants with high experience (M=0.29, SE=0.06) generated more hypotheses than did participants with medium experience (M=0.03, SE=0.06), t(118)=-3.20, p=0.002; and participants with high experience generated more hypotheses than participants with low experience (M=0.04, SE=0.06), t(117)=-3.15, p=0.002. For the developmental cluster, participants with medium experience (M=0.22, SE=0.05) generated more hypotheses than participants with low experience (M=0.00, SE=0.05), t(117)=3.31, p=0.001.For the clinical/counseling cluster, participants with high experience (M=1.62, SE=0.12) generated more hypotheses than participants with medium experience (M=0.89), t(118)=-4.31, p<0.0001 or low experience (M=0.55, SE=0.12), t(117)=-6.34, p<0.0001, and participants with medium experience generated more hypotheses than participants with low experience, t(117)=2.06, p=0.040. For the social cluster, participants with high experience (M=0.23, SE=0.06) generated more hypotheses than participants with low experience (M=0.04, SE=0.05), t(117)=-2.41, p=0.017, and participants with medium experience (M=0.26, SE=0.05) generated more hypotheses than participants with low experience, t(117)=2.91, p=0.004. For the introduction to psychology cluster, participants with low experience (M=0.54, SE=0.07) generated more hypotheses than did participants with high experience (M=0.29, SE=0.07), t(117)=2.58, p=0.011. For the "other" cluster, participants with high experience (M=1.92, SE=.14) generated more hypotheses than did participants with medium experience (M=1.10, SE=.14), t(118)=-4.09, p<0.0001 and participants with low experience (M=0.39, SE=0.14), t(117)=-7.68, p<0.0001. Further, participants with medium experience generated more hypotheses than participants with low experience, t(117)=3.65, p=0.0003. There were no differences in experience for the industrial/organizational cluster, the research cluster, or the individual differences cluster.

[28] For the cognitive cluster, participants with high experience (M=.55, SE=.08) generated more hypotheses than did participants with low experience (M=.28, SE=.07), t(117)=-2.51, p=0.013. Participants with medium experience (M=0.51, SE=0.07) generated more hypotheses than did participants with low experience, t(117)=2.19, p=0.030. For the bio/neuropsychology cluster, participants with high experience (M=0.71, SE=0.09) generated more hypotheses than did participants with medium experience (M=0.42, SE=0.08), t(118)=-2.47, p=0.015; and participants with high experience generated more hypotheses than participants with low experience (M=0.30, SE=0.08), t(117)=-3.55, p=0.0005. For the clinical/counseling cluster, participants with high experience (M=0.45, SE=0.08) generated more hypotheses than participants with low experience (M=0.21, SE=0.08), t(117)=-2.12, p=0.035. For the social cluster, participants with high experience (M=0.72, SE=0.08) generated more hypotheses than participants with low experience (M=0.21, SE=0.08), t(117)=-4.54, p<0.0001, and participants with medium experience (M=0.58, SE=0.085) generated more hypotheses than participants with low experience, t(117)=3.40, p=0.0009. For the introduction to psychology cluster, participants with low experience (M=0.52, SE=0.06) generated more hypotheses than did participants with high experience (M=0.31, SE=0.07), t(117)=2.22, p=0.028 or with medium experience (M=0.33, SE=0.06), t(117)=-2.04, p=0.043. For the "other" cluster, participants with high experience (M=1.71, SE=.13) generated more hypotheses than did participants with medium experience (M=1.26, SE=.13), t(118)=-2.42, p=0.017 or participants with low experience (M=0.75, SE=0.13), t(117)=-5.13, p<0.0001. Further, participants with medium experience generated more hypotheses than participants with low experience, t(117)=2.79, p=0.006. There were no differences in experience for the developmental cluster, the industrial/organizational cluster, the research cluster, or the individual differences cluster.

165)=15.85, p<.0001). Further, participants with more experience tended to generate more hypotheses from clusters other than introduction to psychology than did participants with less experience. Both experiments found significant hypothesis cluster by Experience interactions (Generation Task 1: F(18, 1485)=17.48, p<.0001; Generation Task 2: F(18, 1485)=6.16, p<.0001).

For Generation Task 2, a significant interaction between cluster, experience, and cue order was found, F(18, 1485)=2.22, p=0.002. The experience effect, although significant alone, did not obtain equally in all cue orders and for all clusters.

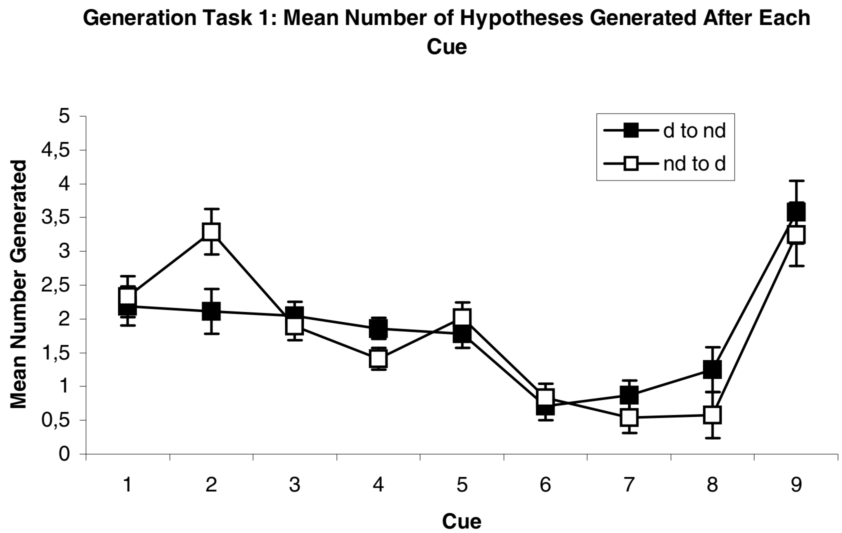
Number Generated

Experiment 1 found that participants tended to generate fewer and fewer hypotheses as they received more and more information, especially when each new cue was more diagnostic than previous ones. It was hypothesized that similar effects would obtain in Experiment 4. Figure 17 presents the mean number of hypotheses generated after each cue was presented in the SBS condition as a function of experience and cue order. Indeed, in both generation tasks, a main effect of cue position obtained (Generation Task 1[29]: F(8, 592)=28.02, p<0.0001; Generation Task

[29] Bonferroni adjustments were used fort he 36 post hoc comparisons. Alpha for each comparison was set to .001. Cue position 1 had a higher mean number generated than: cue position 6, t(80)=5.92, p<0.0001; cue position 7: t(80)=5.72, p<0.0001; and cue position 8: t(80)=4.15, p<0.0001. Cue position 2 had a higher mean number generated than: cue position 4, t(80)=3.72, p=0.0004; cue position 6: t(80)=7.19, p<0.0001; cue position 7: t(80)=6.58, p<0.0001; and cue position 8: t(80)=5.09, p<0.0001. Cue position 3 had higher mean number generated than: Cue position 6, t(80)=6.39, p<0.0001; cue position 7: t(80)=6.06, p<0.0001; and cue position 8: t(80)=3.90, p=0.0002. Cue position 4 had a higher mean number generated than: cue position 6: t(80)=5.07, p<0.0001; cue position 7: t(80)=4.87, p<0.0001; . Cue position 5 had a higher mean number generated than: cue position 6: t(80)=5.88, p<0.0001; cue position 7: t(80)=5.40, p<0.0001; cue position 8: t(80)=3.45, p=0.0009. Cue position 9 had higher mean number generated than: cue position 3: t(80)=-4.55, p<0.0001; cue position 4: t(80)=-5.03, p<0.0001; cue position 5: t(80)=-4.69, p<0.0001; cue position 6: t(80)=-7.40, p<0.0001; cue position 7: t(80)=-8.18, p<0.0001; cue position 8: t(80)=-7.69, p<0.0001.

Figure 17

A



B

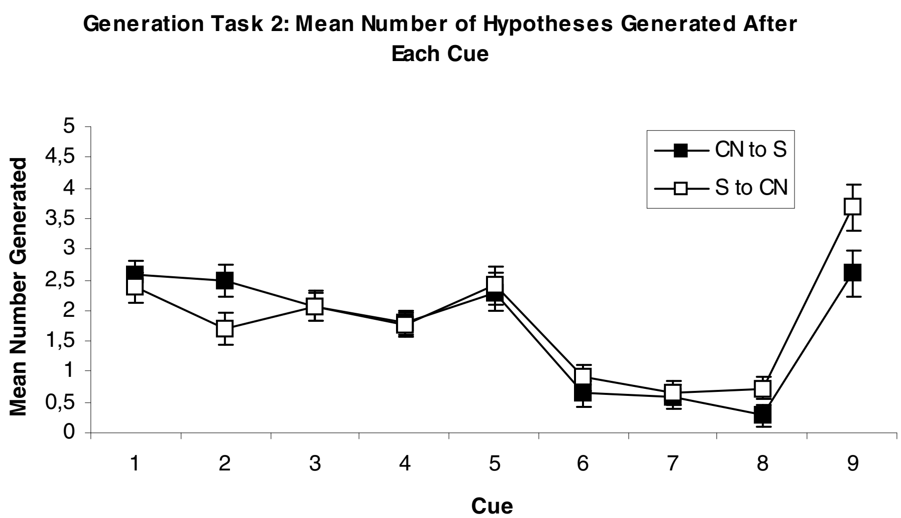


Figure 17 presents the mean number of hypotheses generated after each cue as a function of experience and cue order for participants in the Step-By-Step condition. Panel A presents data for the first generation task and Panel B presents data for the second generation task.

2[30]: F(8, 592)=35.46, p<0.0001). Participants generated fewer hypotheses late in the sequence than early in the sequence. The exception was that participants again generated more hypotheses after the final cue.

In both generation tasks, there was an interaction between cue position and cue order group (Generation Task 1[31]: F(8, 592)=2.47, p=0.012; Generation Task 2[32]: F(8, 592)=2.60, p=0.009). In Generation Task 1, after early cues participants tended to generate more hypotheses in the non-diagnostic to diagnostic order, and after later cues tended to generate more hypotheses in the diagnostic to non-diagnostic cue order. In general, participants generated fewer courses after diagnostic cues were presented than after non-diagnostic cues were presented. In Generation Task 2, after early cues participants tended to generate more hypotheses in the cognitive neuroscience to social cue order than in the reverse cue order. After later cues,

---

[30] Bonferroni adjustments were used fort he 36 post hoc comparisons. Alpha for each comparison was set to .001. Cue position 1 had a higher mean number generated than: cue position 4, t(80)=4.42, p<0.0001; cue position 6: t(80)=8.98, p<0.0001; cue position 7: t(80)=9.45, p<0.0001; and cue position 8: t(80)=10.91, p<0.0001. Cue position 2 had a higher mean number generated than: cue position 6, t(80)=5.60, p<0.0001; cue position 7: t(80)=6.15, p<0.0001; and cue position 8: t(80)=6.47, p<0.0001. Cue position 3 had higher mean number generated than: Cue position 6, t(80)=6.07, p<0.0001; cue position 7: t(80)=6.98, p<0.0001; and cue position 8: t(80)=7.46, p<0.0001. Cue position 4 had a higher mean number generated than: cue position 6: t(80)=5.38, p<0.0001; cue position 7: t(80)=6.43, p<0.0001; and cue position 8: t(80)=7.68, p<0.0001. Cue position 5 had a higher mean number generated than: cue position 6: t(80)=6.33, p<0.0001; cue position 7: t(80)=7.58, p<0.0001; cue position 8: t(80)=7.58, p<0.0001. Cue position 9 had higher mean number generated than: cue position 3: t(80)=-3.41, p=0.001; cue position 4: t(80)=-4.99, p<0.0001; cue position 6: t(80)=-8.10, p<0.0001; cue position 7: t(80)=-8.88, p<0.0001; cue position 8: t(80)=-9.23, p<0.0001.

[31] After cue 2, participants generated more hypotheses in the ND to D cue order (M=3.29, SE=0.34) than in the D to ND order (M=2.11, SE=0.33), t(80)=-2.50, p=0.015. After cue 4, participants generated more hypotheses in the D to ND cue order (M=1.86, SE=0.16) than in the ND to D order (M=1.41, SE=0.16), t(80)=2.01, p=0.048.

[32] After cue 2, participants generated more hypotheses in the CN to S cue order (M=2.48, SE=0.25) than in the S to CN cue order (M=1.70, SE=0.25), t(80)=2.22, p=0.029. After cue 9, participants generated more hypotheses in the S to CN cue order (M=3.68, SE=0.38) than in the CN to S cue order (M=2.61, SE=0.38), t(80)=-1.99, p=0.050.

participants tended to generate more hypotheses in the social to cognitive neuroscience cue order than in the reverse order.

In both generation tasks, participants generated more hypotheses when they had higher levels of experience (Generation Task 1: $F(2, 74)=9.50$, $p=0.0002$; Generation Task 2: $F(2, 74)=11.43$, $p<0.0001$). Further, in both generation tasks this main effect was qualified by an interaction between cue position and experience (Generation Task 1[33]: $F(16, 592)=3.01$, $p<0.0001$; Generation Task 2[34]: $F(16, 592)=2.23$, $p=0.004$). Although in general participants with high experience generated more hypotheses than participants with medium experience, and participants with medium experience generated more hypotheses than participants with low experience, these effects were not equal across cue positions.

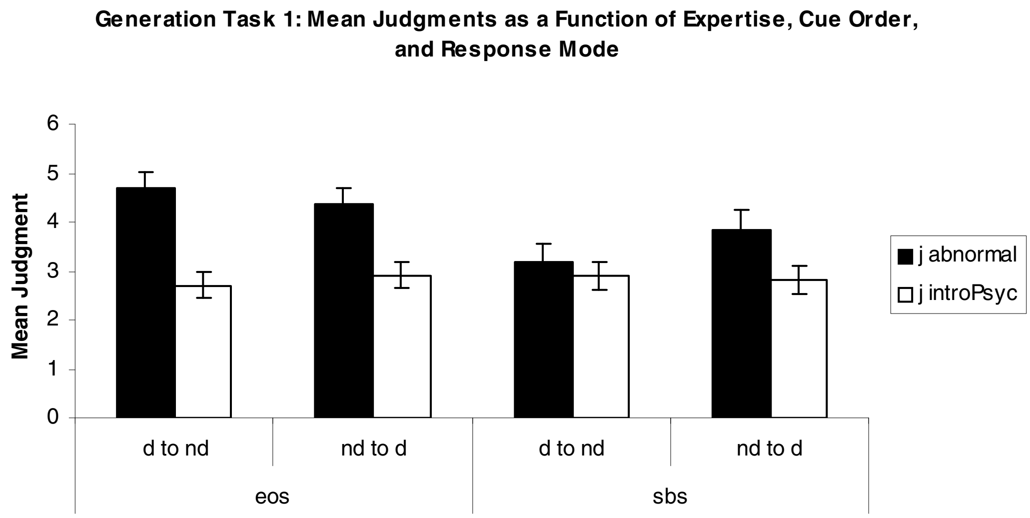
Judgments

Figure 18 presents mean probability judgments as a function of experience, cue order, and response mode. Panel A presents data for Generation Task 1, and panel B presents data for Generation Task 2. For Generation Task 1, it was predicted that participants would judge the Abnormal Psychology course as more likely than the Introduction to Psychology course, because the diagnostic cues pointed toward

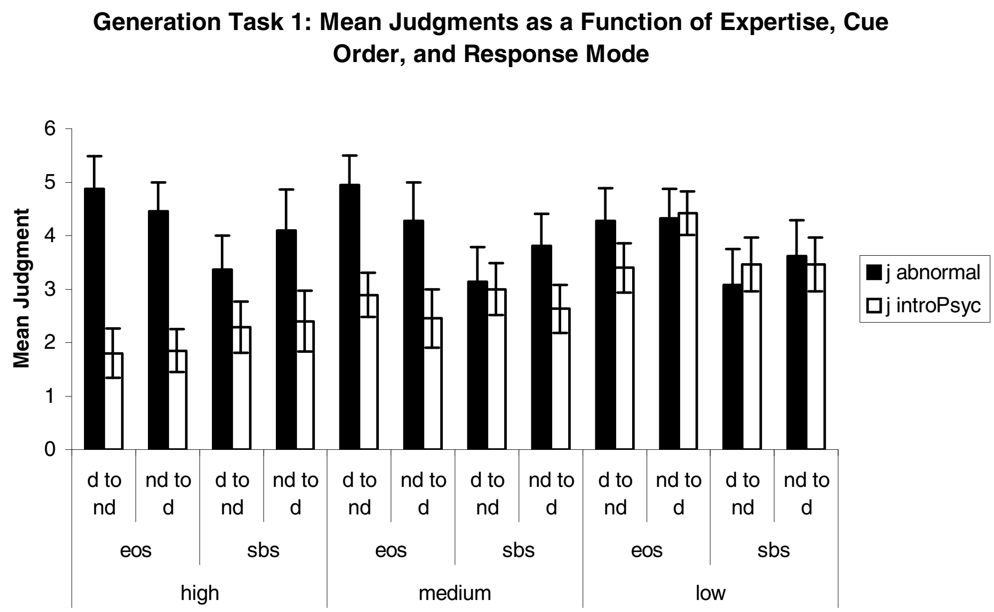
[33] Bonferroni adjustments were used fort he 27 post hoc comparisons. Alpha for each comparison was set to .002. After cue 1, participants with high experience generated more hypotheses than did participants with low experience, $t(48)=-3.60$, $p=0.0006$. After cue 2, participants with high experience generated more hypotheses than participants with medium experience, $t(54)=-3.25$, $p=0.0017$; and participants with low experience, $t(48)=-3.63$, $p=0.0005$. After cue 9, participants with high experience generated more hypotheses than participants with low experience, $t(48)=-4.00$, $p=0.0002$.

[34] Bonferroni adjustments were used fort he 27 post hoc comparisons. Alpha for each comparison was set to .002. After cue 1, participants with high experience generated more hypotheses than did participants with low experience, $t(48)=-3.75$, $p=0.0003$. After cue 4, participants with high experience generated more hypotheses than participants with low experience, $t(48)=-3.57$, $p=0.0006$. After cue 5, participants with high experience generated more hypotheses than participants with low experience, $t(48)=-3.59$, $p=0.0006$. After cue 9, participants with high experience generated more hypotheses than participants with low experience, $t(48)=-3.78$, $p=0.0003$.

Figure 18



A



B

Figure 18 presents mean probability judgments as a function of cue order, and response mode. Panel A presents data for the first generation task and Panel B presents data for the second generation task.

clinical and counseling-type courses. Indeed, analyses found a main effect of judged item, $F(1, 167)=25.31$, $p<0.0001$. Participants judged the Abnormal Psychology course more likely than the Introduction to Psychology course. It was also predicted that probability judgments would follow a pattern similar to hypothesis generation, such that differences between the items judged would be greatest when diagnostic information was presented late in the sequence for SBS conditions. No interaction between item judged, response mode, and cue order was found, however. Analyses found a main effect of response mode, $F(1, 167)=4.42$, $p=0.037$. Participants in the EOS condition made higher judgments than participants in the SBS condition. These main effects were qualified by an interaction between the item judged and response mode, $F(1, 167)=5.21$, $p=0.024$. When judging the abnormal psychology course, participants in the EOS condition ($M=4.51$, $SE=0.25$) gave higher judgments than participants in the SBS condition ($M=3.52$, $SE=0.27$), $t(177)=2.73$, $p=0.007$. However, no response mode differences were found for judgments of the introduction to psychology course. Finally, analyses found a significant interaction between the item judged and experience, $F(2, 167)=5.97$, $p=0.0031$. No experience effects were found when judging the abnormal psychology course, but when judging the introduction to psychology course, participants with low experience ($M=3.69$, $SE=0.23$) gave higher judgments than participants with high experience ($M=2.08$, $SE=0.24$), $t(117)=4.79$, $p<0.0001$; and participants with low experience gave higher judgments than participants with medium experience($M=2.74$, $SE=0.24$), $t(117)=-2.84$, $p=0.005$.

For the second generation task, analyses found a significant main effect of judged item, $F(1, 167)=63.97$, $p<0.0001$. Participants judged the Introduction to Memory and Cognition course as more likely than the Helping Skills course. Further, a significant experience effect obtained[35], $F(2, 167)=8.51$, $p=0.0003$. Participants with more experience tended to give lower probability judgments. Finally, analyses found a significant main effect of cue order, $F(1, 167)=6.45$, $p=0.012$. Participants in the Social to Cognitive Neuroscience order made higher judgments on average than participants in the reverse order condition.

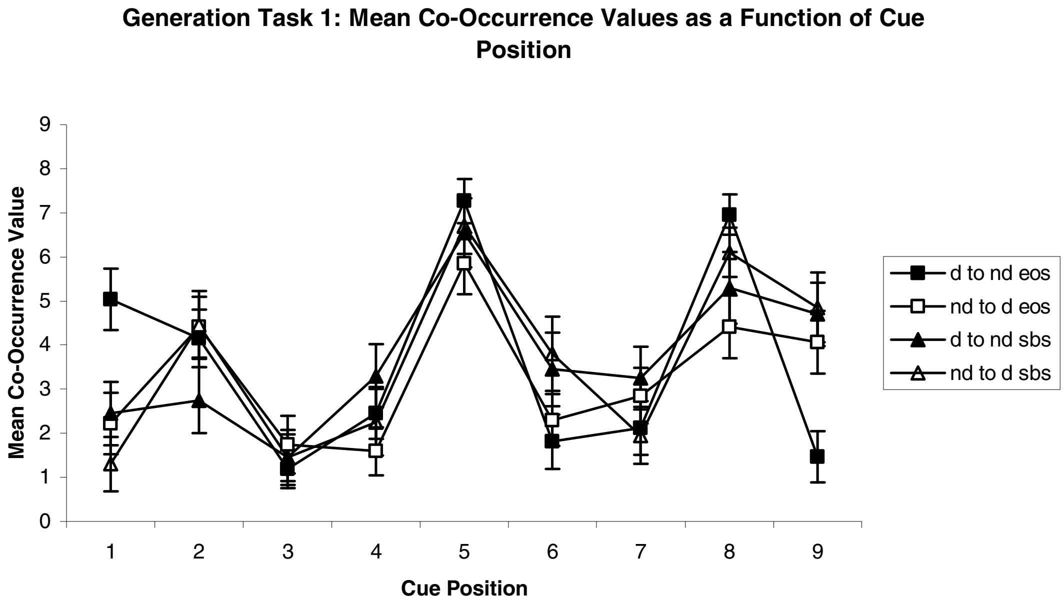
Recall

It was hypothesized that recall of cue words would provide insight about how participants weighed the nine cues during hypothesis generation, because participants were likely to best recall items that they attended most during observation of cues and hypothesis generation (Fisher, 1987). Mean recall co-occurrence values were analyzed as a function of cue position (the order that cues appeared), response type (EOS vs. SBS), and cue order (whether the most diagnostic information came first vs. last in Generation Task 1, and whether information pointed to social courses first and later to cognitive neuroscience courses or vice versa). Figure 19 presents mean recall co-occurrence values as a function of cue position, response mode, and cue order. The first main finding was that in both cases, participants tended to recall items from

[35] When judging the Helping Skills course, participants with high experience ($M=3.05$, $SE=0.23$) made significantly lower judgments than participants with low experience ($M=3.80$, $SE=0.23$), $t(117)=2.31$, $p=0.022$. When judging the Memory and Cognition Course, participants with high experience ($M=4.35$, $SE=0.28$) made lower judgments than participants with medium experience ($M=5.74$, $SE=0.27$), $t(118)=3.62$, $p=0.0004$; and participants with high experience made lower judgments than participants with low experience ($M=5.28$, $SE=0.27$), $t(117)=2.41$, $p=0.017$.

Figure 19

A



B

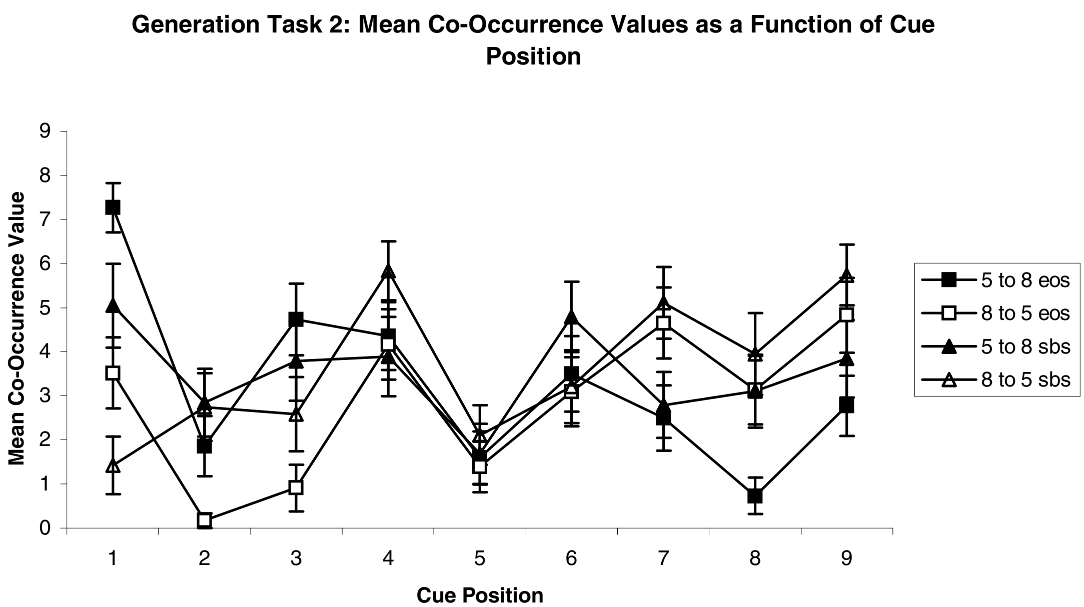


Figure 19 presents mean co-occurrence values for recall as a function of cue position, cue order, and response type. Panel A presents data for Generation Task 1 and Panel B presents data for Generation Task 2.

some cue positions better than from other cue positions, (Generation Task 1[36]: F(8, 648)=23.38, p<0.0001; Generation Task 2[37]: F(8, 568)=7.51, p<0.0001). For Generation Task 1, participants tended to best remember words from cue position 5 and 8, and to a lesser degree positions 2 and 9. Cue position 5 always contained the word "abnormal" which was a highly diagnostic word in the sequence. Thus, participants tended to best remember one of the diagnostic words. Further, one early cue (position 2) and two late cues (positions 8 and 9) were best remembered, indicating some primacy and recency obtained.

For Generation Task 2, participants tended to recall all cues equally well except for positions 2 and 5. Those two cue positions were recalled less well than the others. The second main finding with recall was a significant cue position by cue order interaction (Generation Task 1[38]: F(8, 648)=2.46, p=0.013; Generation Task

---

[36] Bonferroni adjustments were used fort he 36 post hoc comparisons. Alpha for each comparison was set to .001. Cue position 2 had a higher mean co-occurrence value than: cue position 3, t(93)=5.47, p<0.0001; and cue position 4: t(93)=3.36, p=0.001. Cue position 5 had a higher mean co-occurrence value than: cue position 1, t(93)=-7.82, p<0.0001; cue position 2: t(93)=-5.31, p<0.0001; cue position 3: t(93)=-11.55, p<0.0001; cue position 4: t(93)=-8.73, p<0.0001; cue position 5: t(93)=8.37, p<0.0001; cue position 7: t(93)=8.62, p<0.0001; and cue position 9: t(93)=6.01, p<0.0001. Cue position 8 had higher mean co-occurrence value than: Cue position 1, t(93)=-5.86, p<0.0001; cue position 4: t(93)=-6.93, p<0.0001; cue position 6: t(93)=-5.27, p<0.0001; cue position 7: t(93)=-6.84, p<0.0001; and cue position 9: t(93)=3.82, p=0.0002. Cue position 9 had a higher mean co-occurrence value than: cue position 3: t(93)=-4.39, p<0.0001

[37] Bonferroni adjustments were used fort he 36 post hoc comparisons. Alpha for each comparison was set to .001. Cue position 1 had a higher mean co-occurrence value than: cue position 2, t(83)=4.91, p<0.0001 and cue position 5, t(83)=5.24, p<0.0001. Cue position 4 had a higher mean co-occurrence value than cue position 2: t(83)=-5.56, p<0.0001; and cue position 5: t(83)=5.19, p<0.0001. Cue position 6 had a higher mean co-occurrence value than cue position 2: t(83)=-3.47, p=0.0008; and cue position 5, t(83)=-3.82, p=0.0003. Cue position 7 had a higher mean co-occurrence value than cue position 2: t(83)=-3.57, p=0.0006; and cue position 5, t(83)=-4.05, p=0.0001. Cue position 9 had a higher mean co-occurrence value than cue position 2: t(83)=-4.41, p<0.0001; and cue position 5: t(83)=-4.42, p<0.0001.

[38] Participants recalled cue position 1 more in the diagnostic to non-diagnostic cue order (M=3.68, SE=0.52) than in the non-diagnostic to diagnostic cue order (M=1.57, SE=0.51), t(93)=2.91, p=0.005.

2[39]: F(8, 568)=6.17, p<0.0001). In Generation Task 1, participants better remembered cue position 1 when that position contained a diagnostic word than when that position contained a non-diagnostic word. In Generation Task 2, participants tended to recall cue positions 1 and 3 best in the cognitive neuropsychology to social cue order and cue positions 7 and 9 best in the opposite order. These cue positions always contained the words “memory” and “brain”, suggesting that no matter which cue order participants saw, they tended to best recall the cue words “memory” and “brain”.

It was predicted that participants would recall mostly recent cues in the SBS response mode condition, and both early and late cues in the EOS condition. However, no significant interaction between cue position, cue order, and response mode was found. This contrasts with hypothesis generation findings, where the type of order effect that obtained depended on response mode.

For the first generation task, there was a significant interaction between cue position and response mode, F(8, 648)=2.82, p=0.005. Participants recalled cue position 1 better in the EOS (M=3.48, SE=0.48) than in the SBS condition (M=1.77, SE=0.55), t(93)=2.37, p=0.020. Participants recalled cue position 6 better in the SBS (M=3.84, SE=0.52) than in the EOS condition (M=2.09, SE=0.46), t(93)=-2.52, p=0.014. Finally, participants recalled cue position 9 better in the SBS (M=4.82, SE=0.52) than in the EOS condition (M=2.79, SE=0.46), t(93)=-2.89, p=0.005. For the first generation task, there was a significant 4-way interaction between cue

[39] Participants recalled cue position 1 more in the CN to S order (M=6.14, SE=0.55) than in the S to CN order (M=2.58, SE=0.58), t(83)=4.48, p<0.0001. Participants recalled cue position 3 more in the CN to S order (M=4.34, SE=0.57) than in the S to CN order (M=1.81, SE=0.60), t(83)=3.06, p=0.003. Participants recalled cue position 7 less in the CN to S order (M=2.74, SE=0.58) than in the S to CN order (M=5.22, SE=0.61), t(83)=-2.97, p=0.004. Participants recalled cue position 9 less in the CN to S order (M=3.49, SE=0.56) than in the S to CN order (M=5.40, SE=0.60), t(83)=-2.32, p=0.023.

position, experience, cue order, and response mode, F(16, 648)=2.02, p=0.010. In Generation Task 2, there was a significant main effect of response mode, F(1, 71)=6.34, p=0.014. Participants in the SBS condition tended to recall more cue words than did participants in the EOS condition. Perhaps generating after observing each cue led participants to attend each cue more, and later this led to increased recall in the SBS condition. In Generation Task 2, a significant interaction between cue position, experience, and cue order obtained, F(16, 568)=1.73, p=0.037. For cue position 1, participants with high experience tended to have higher mean co-occurrence values in the CN to S condition than in the S to CN condition, t(83)=3.86, p=0.0002. Participants with low experience tended to have higher mean co-occurrence values in the CN to S condition than in the S to CN condition, t(83)=-4.08, p=0.0001.

Discussion

Several general conclusions can be reached from Experiment 4. First, for hypothesis generation, the type of order effect that obtains depends on the response mode. In the end of sequence response mode, no order effects obtained, whereas in the step-by-step response mode, recency effects obtained. Although many more hypotheses were possible in Experiment 4 than in Experiment 1, similar results obtained in both experiments. Participants in the SBS condition were most influenced by recent cues in both cases. It is interesting that no order effects obtained in the EOS condition, whereas primacy effects tend to obtain in EOS belief-updating tasks (Hogarth & Einhorn, 1992). This study demonstrated that after viewing an entire sequence of cues participants are equally biased by early and late cues. It is possible that no order effects obtained because participants considered all cues equally, or because they weighed early and late cues more than middle cues. In real-world cases in which people must generate hypotheses, the true “response mode” is often a mixture between SBS and EOS modes. People observe a sequence of several cues, then form an initial set of hypotheses. Then people observe new cues, either passively due to chance encounters with new information or actively through information search processes. People then generate hypotheses again. Thus, future research could examine what types of order effects obtain when participants respond in a cross between SBS and EOS response modes.

Second, after viewing all cues, participants in the EOS condition tended to generate more hypotheses than participants in the SBS condition. This finding replicates the finding in Experiment 3 that participants generated more hypotheses in the EOS condition than in the SBS condition after all cues were presented. Two possible explanations could account for this finding. First, it is possible that participants in the SBS condition lose motivation to generate hypotheses when they have already generated hypotheses eight times prior to the final generation. Second, perhaps requiring participants to attend and respond after each cue leads participants to rule out more hypotheses as they view more cues. Then, after all cues are presented participants in the SBS condition have narrowed their hypothesis set down more so than participants in the EOS condition.

Third, participants with more experience generated more courses on average. Further, participants with less experience tended to generate more hypotheses from the “introduction to psychology” cluster. These findings are not unexpected, in that participants with low experience only know several psychology courses and participants with high experience know many psychology courses. When comparing high and low experience participants, high experience participants have a larger pool of courses from which to generate.

Fourth, as in Experiments 1 and 3, participants in Experiment 4 tended to generate fewer hypotheses as they viewed more cues. Then, when viewing the final cue, participants again generated more hypotheses. This effect could be due either to a natural process whereby participants attempt to rule out more and more hypotheses as they learn more and more information, or could be due to motivation decreasing as participants view more and more information. Participants generated fewer courses after diagnostic cues were presented than after non-diagnostic cues were presented. In Generation Task 1, after early cues participants generated more hypotheses in the non-diagnostic to diagnostic order, and after later cues participants generated more hypotheses in the diagnostic to non-diagnostic cue order. This finding shows that participants were sensitive to the diagnosticity of cues. By definition, non-diagnostic cues are consistent with more hypotheses than diagnostic cues, and therefore participants generate more hypotheses after non-diagnostic cues than after diagnostic ones.

Interestingly, no order effects obtained for participants' probability judgments. It was expected that probability judgments would follow patterns similar to the generation results, in which case participants would be most affected by recent information in the SBS condition and would be equally influenced by all information in the EOS condition. This lack of order effects contrasts with the findings of

Experiment 2, in which participants' judgments were most affected by recent information in an SBS format. In Experiment 1 participants responded on a numerical scale whereas in Experiment 4 participants responded on a verbal scale. Perhaps different order effects obtain depending on the type of scale participants are asked to respond on. The lack of order effects in probability judgments also contrasts with belief updating studies in which participants show primacy in EOS conditions and recency in SBS conditions (Hogarth & Einhorn, 1992). Unlike belief updating studies, participants in Experiment 4 had to generate alternatives to the focal hypothesis themselves; the alternative(s) were not simply presented to them. Future research should further examine when order effects obtain in probability judgments that require hypothesis generation from long-term memory.

In Experiment 4, the recall task provided some preliminary indication of the cues that participants attended most during the generation and judgment tasks. Several interesting findings obtained. First, participants tended to recall some items better than others, but the pattern of recall differed from those found in typical memory experiments. In typical memory experiments, when participants learn lists of words and recall the words immediately, participants tend to best recall items presented early and late in the sequence (Page & Norris, 1998; Rundus, 1971). In other words, a U-shaped function is found. In Experiment 4, Generation Task 1, however, participants recalled cues from the beginning, middle, and end of the sequence better than other cues. In other words, a W-Shaped function obtained. This recall function was similar to the U-shaped function, except for position 5. That cue position always contained the word "abnormal," a diagnostic word. In Generation

Task 2, participants tended to recall all items equally well except for items in positions 2 and 5. These findings suggest that in real-world applications, memory recall functions may differ from those typically found in memory research. A second interesting finding from the recall task was that participants tended to recall diagnostic cues better than less diagnostic cues. In Generation Task 1, participants better remembered cue position 1 when that position contained a diagnostic word than when that position contained a non-diagnostic word. In Generation Task 2, participants tended to recall cue positions 1 and 3 best in the cognitive neuropsychology to social cue order and cue positions 7 and 9 best in the opposite order. These cue positions always contained the words “memory” and “brain”, suggesting that no matter what cue order participants saw, they tended to best recall these two cue words. Participants in Generation Task 2 judged the memory and cognition course as more likely than the helping skills course, perhaps suggesting that participants deemed these two cues (memory and brain) as the most diagnostic. Fisher (1987) also found that participants tended to recall diagnostic cues better than other cues.

In contrast with the hypothesis generation findings, no interaction between cue position, cue order, and response mode was found for recall. This is surprising because if generation is based on the cues maintained in working memory, one would hypothesize that when participants were most affected by certain cues in generation tasks, they would also tend to best recall those cues. Thus, the recall and hypothesis generation tasks may have tapped into different processes. It is possible that participants tended to best recall the words, “abnormal”, “memory”, and “brain”, because recall was affected by the items judged in the judgment task. Participants recalled cue words *after* performing the judgment tasks, and two of the courses that participants judged were “Introduction to Memory and Cognition” and “Abnormal Psychology”. It is possible that the judgments of these items affected participants’ recall more so than the hypothesis generation task itself. Future research should examine cue recall functions directly after participants generate hypotheses to eliminate this confound.

It is interesting to note the diversity of courses that participants generated. Although UMD offered approximately 64 psychology courses, participants generated 137 different courses. Participants generated many new, non-existent courses. Some new courses combined several cue words, such as “Brain and Memory Psychology”. Other times, participants generated courses that were not from the psychology domain, such as “American Studies” and “Anatomy”. Also interesting, some participants generated “any psychology course”, a catchall hypothesis, rather than generating all individual courses they could think of.

Chapter 6: General Discussion

The purpose of this research was to examine how decision makers generate and evaluate hypotheses when data are presented sequentially. Several general findings from the experiments are of importance. First, hypothesis generation was most influenced by recent cues when participants responded after viewing each cue. Of all participants in the singular cue presentation condition in Experiment 1, participants who received the most diagnostic piece of information last weighted it more heavily than participants who received the most diagnostic piece of information first. In the Step-by-Step condition of Experiment 4, participants generated hypotheses consistent with the most recent cues, rather than cues presented early in the sequence. Second, the type of order effect that obtains depends on response mode. Experiments 1 and 4 found recency effects in the SBS condition and Experiment 4 found no order effects in the EOS condition. One possibility is that participants in the EOS condition weighted all cues equally in their generation. A second possibility is that participants in the EOS conditions weighted early and late cues more than middle cues. Third, people tend to narrow their sets of hypotheses as they learn new information.

Note that the order effects found in these studies differ from those found in belief updating studies. In belief updating studies, primacy usually obtains in EOS response formats. Anderson (1973) argued that primacy obtains in belief updating tasks because less attention is given to each cue as people view more cues. The exception is that when attention is directed to each cue, recency usually obtains because participants attend each new cue more than previous ones. However, primacy did not obtain in the EOS condition in Experiment 4. It seems that the way that people attend cues differs when people generate hypotheses versus when people update beliefs.

In Experiments 1, 3, and 4, participants tended to generate fewer hypotheses as more evidence was presented, and this was true even when the new evidence was non-diagnostic. Further, this finding obtained both in the detective-robbery paradigm (Experiments 1 and 3) as well as in the psychology courses paradigm (Experiment 4). This contrasts with Weber et al.'s (1993) finding that as more and more information was presented, doctors tended to generate the same number of hypotheses on average, but that their hypothesis sets tended to become more homogenous. In the present experiments, participants attempted to find the correct hypothesis, and thus eliminated alternative hypotheses even when new data was completely non-diagnostic.

After all cues were presented, participants in the EOS condition generated more hypotheses than participants in the SBS condition, and this finding occurred in Experiments 3 and 4. Two possible explanations of this finding exist. First, generating hypotheses after viewing each cue induces participants to weed out more hypotheses than when generating hypotheses only after viewing all cues. Second, the SBS condition causes people to lose motivation to generate many hypotheses because they are asked to generate multiple times. In Experiment 4, participants in the SBS condition were asked to generate hypotheses nine different times after viewing each of the nine cues. Repeated generation might lead participants to put forth less effort to come up with new hypotheses.

Theoretical Implications

One of the main goals of this paper was to begin providing an empirical basis for the development and testing of the Sequential HyGene model. I will now discuss several alternative versions of the HyGene model and how the data constrain each.

The first possible version of the Sequential HyGene model was specified in the introduction. In this version, instead of all data simultaneously probing long-term memory to initiate hypothesis generation, a *subset* of data is activated based on the cue activation function specified in equation 1. Activated data are integrated into a single composite probe, where the degree to which each cue contributes to the composite probe is weighted by the cue's activation in working memory.

The data from the current set of experiments constrain this model in several ways. First, the experiments suggest that when participants generate hypotheses after observing each new piece of information, recent information is weighted most. Thus, the cue memory function in which PI is high and RI is low is supported. These parameter settings produce the recency curves in figure 2. Second, although it is yet unclear exactly which cue memory parameters should be used when participants generate hypotheses after viewing all cues, the current experiments do constrain the model. In the current experiments, no order effects obtained in EOS conditions. These results rule out the cue memory function in which retroactive interference is high and proactive interference is low. These parameter settings produce the primacy curves in Figure 2. The results also rule out the cue memory function in which proactive interference is high and retroactive interference is low. These parameter settings produce the recency curves in Figure 2. Thus, when participants generate hypotheses after viewing all cues, one of two possible versions of the cue memory function is supported. First, it is possible that the primacy and recency curve will best account for participants' hypothesis generation in EOS tasks. A second possibility is that a function in which all cues are equally weighted will best account for participants' hypothesis generation in EOS tasks.

A second possible version of the Sequential HyGene model is informed by Gettys and Fisher's (1979; Fisher, Gettys, Manning, Mehle, & Baca, 1983; Fisher, 1987) model of hypothesis generation. Gettys and Fisher proposed that due to STM constraints, only a subset of data are used in retrieving hypotheses. Then, once a hypothesis is retrieved, before it is included in WM, it is compared with *all* data to verify that it is consistent with all data. The original HyGene model was consistent with the assumption that after a hypothesis is retrieved, it is compared with data to check for consistency before the hypothesis is included in WM. HyGene assumed that after a hypothesis was retrieved, each data portion of the hypothesis was compared with each observed data. If any data were dissimilar, the hypothesis was not included in WM. However, the original HyGene model assumed that all data were simultaneously used to retrieve hypotheses. The Sequential HyGene model assumes that only active data are used to probe LTM and for the consistency checking process. Thus, two competing assumptions about consistency checking are possible. In one version of the Sequential HyGene model, it would be assumed that all observed data are compared with generated hypotheses to check for consistency. This version of the model is consistent with Gettys and Fisher's theory. In the second version of the Sequential HyGene model, it would be assumed that only activated data are compared with generated hypotheses to check for consistency. This version of the model is consistent with Dougherty et al.'s theory.

The data from the current experiments are inconsistent with Gettys and Fisher's assumption that all data are used for consistency checking. Participants tended to weight diagnostic information less when it was presented early in the sequence than when it was presented late in the sequence. This finding suggests that participants forget some early data or that WM capacity is reached and some data are consequently "kicked out" to store new data and hypotheses. For instance, in experiment 1, Crime Scene A, when diagnostic information was presented early in the sequence, after viewing all data participants generated highly unlikely hypotheses (three hypotheses that each had objective probabilities less than .05). If all data were available for use in consistency checking, participants would rule out those hypotheses whose objective likelihood was low. Indeed, when diagnostic information was presented late in the sequence, participants did tend to rule out those hypotheses. These results suggest that not all data are available for consistency checking, but rather only the most recent data are available. Thus, my experiments constrain the Sequential HyGene model by indicating that only a subset of data are used in consistency checking, and in SBS cases those cues available are the most recent ones observed.

A second assumption of Gettys and Fisher's model that could be implemented in Sequential HyGene is that participants only attempt to generate new hypotheses when new data render the entire set of previously considered hypotheses implausible (below some threshold, k). They argued that participants engage in a "win-stay, loseswitch" strategy; so long as new data do not decrease participants' subjective plausibility estimates of their previous hypothesis set, participants simply maintain that previous hypothesis set (Fisher, 1987; Gettys & Fisher, 1979). If new data decrease participants' subjective plausibility estimates of their previous hypothesis set, participants engage hypothesis generation processes.

The assumption that participants only attempt to generate new hypotheses when new data render the entire set of previously considered hypotheses implausible (below some threshold, k), is challenged by the data from the current study. If participants simply maintained the same sets of hypotheses so long as new data did not indicate the previous set was unlikely, when the most diagnostic information was presented early in the sequence participants should maintain the exact same set of hypotheses as new data were presented. In contrast, in the current studies participants ruled out some hypotheses as they viewed additional information, even when the new information was non-diagnostic. When new information was non-diagnostic, participants' subjective evaluations of the plausibility of their current hypothesis sets are unlikely to change. This finding, that people narrow their hypothesis set even when new information is non-diagnostic, is inconsistent with Gettys and Fisher's argument that participants simply maintain the same hypothesis sets so long as new data do not render them less likely. The finding is instead consistent with the notion that as more data are learned, WM capacity is reached. Then, participants cannot maintain as many cues and data in WM, leading them to rule out or forget some previously considered hypotheses. In consequence, the number of hypotheses considered decreases as the number of cues observed increases.

An additional flaw in Gettys and Fisher's theory is that although they indicate that not all data are used to probe memory due to STM limitations, they do not specify which cues are used. Any model of sequential hypothesis generation will need to specify which data are active during hypothesis generation.

A third possible version of the sequential HyGene model is based on the buffer component of the SAM memory model (Raaijmakers & Shiffrin, 1981; Mensink & Raaijmakers, 1988). Note that this version of HyGene is consistent with the version in which cues are activated via the activation function (equation 1) and then are combined into a consolidated data probe. The version specified here simply details the psychological processes that result in the curves produced by the cue activation function. In this version of Sequential HyGene, it is assumed that working memory (WM) has a limited capacity; only a limited number of items can be retained, and rehearsed at one time. Each datum enters WM, and joins previous data in WM until WM-capacity is reached. Then, each new datum replaces one of the previous data in WM. Items in WM are transferred to LTM. Two types of information are stored in LTM: item information and context information. It is assumed that context information changes slightly with each unit increase in time. Thus, each datum stored would have different context associated with it, and the degree to which the context component of one datum is similar to the context component of other data depends on how far apart the data occurred temporally. How well item information is stored in LTM depends on how long it resides in WM. It is assumed that with each increment in time, additional features would be stored in the data's item representation. Before hypothesis generation begins, data are retrieved from WM and LTM. All data residing in WM are assumed to be recalled (unless a distractor task occurs between learning and generation). Data in LTM are retrieved by probing LTM. LTM is first probed with the current context. Then, after an item is retrieved it is used in combination with context to probe memory to attempt to retrieve additional items.

One interesting aspect of this version of the model is that predictions about which cues are activated and used to probe LTM are based on both context strength as well as item strength. It could be the case that recency obtains in SBS due to context changes. In SBS conditions, the amount of time that passes between viewing the original cue to viewing the final cue is longer than in EOS conditions, because participants generate hypotheses after observing each cue rather than only after viewing all cues. Because context changes as time passes, the context after the final cue in SBS conditions would tend to be more dissimilar to original contexts stored with early data than would the context after the final cue in the EOS conditions. Thus, in SBS conditions, context after the final cue would tend to be similar only to the most recent items. In contrast, the context after the final cue in EOS conditions would be similar to most of the items. In the SBS case, this model would predict recency and less primacy. Early items would be less likely to be retrieved because those items are less similar to the current context. Recent items would tend to be retrieved because they still resided in the buffer, and because they are similar to the current context. In the EOS case, this model would predict both primacy and recency. Early items are more likely to be retrieved due to still being consistent with the current context (since less time has passed) as well as because those items were stored most strongly in

LTM due to being in buffer longer. Recent items would tend to be retrieved because they still resided in the buffer, and because they are similar to the current context.

How do my data relate to this third version of the model? If the implemented model actually makes the predictions I am hypothesizing, this version of the model would be supported by the data in the current set of experiments. Recency would obtain in SBS cases, and primacy and recency would obtain in EOS cases.

Finally, in comparison with the various versions of HyGene specified thus far, one could develop a sequential hypothesis generation model based on Hogarth and Einhorn's (1992) belief updating model. Hogarth and Einhorn's model was developed to account for how a person's belief in a singular hypothesis is updated as new information is learned. Hogarth and Einhorn proposed that people's belief in a given hypothesis is updated by a general, sequential anchoring and adjustment process in which current opinions are adjusted by the impact of succeeding pieces of information as specified in equation 3:

$$S_k = S_{k-1} + w_k \left[s(x_k) - R \right] \qquad (3)$$

where S_k represents the degree of belief in some hypothesis after k pieces of information; S_{k-1} represents the prior opinion; $s(x_k)$ represents the subjective evaluation of the kth piece of evidence, R represents the reference point against which the kth piece of evidence is evaluated, and w_k represents the adjustment weight for the kth piece of evidence. Hogarth and Einhorn accounted for primacy in EOS tasks and recency in SBS tasks using this model. The authors did not attempt to explain how people generate to-be-updated hypotheses. However, one could extend Hogarth and Einhorn's model by assuming that participants update all possible hypotheses in

LTM using equation 2, and those hypotheses whose belief strength is above some criterion are entered into WM.

The current experiments do not support this model. First, the model predicts primacy in EOS cases, and the current experiments suggest no order effects obtain for hypothesis generation in EOS response modes. Second, the model places heavy demands on memory. It is unclear how all hypotheses could be simultaneously updated, unless the process is assumed to occur automatically. Third, the model does not make predictions about how the number of hypotheses generated changes as new data are observed. The model would simply assume that the number of hypotheses in WM at any given time would equal WM capacity.

Future Experiments

Several future experiments are necessary to further develop and constrain the Sequential HyGene model. First, although the current experiments clearly indicate that the recency curves should be used for SBS cases, it is unclear which type of curve should be used for EOS cases. A future experiment will further examine order effects in EOS cases by positioning the most diagnostic information in the middle of the sequence in one condition, at the beginning of the sequence in a second condition, and at the end of the sequence in the third condition. If primacy and recency obtain in EOS cases, participants will be more affected by diagnostic information when it is presented either early or late in the sequence than when presented in the middle of the sequence. If, on the other hand, participants tend to weight all information equally in EOS cases, then no differences should be found between the three conditions.

The current experiments ruled out primacy-only order effect explanations of sequential hypothesis generation. However, in the belief updating literature, primacy tends to obtain when many data have been observed, irrespective of the response format used. In a future experiment, EOS and SBS conditions will be compared when *many* data are possible (more than 12). Further, diagnostic information will be presented early, middle, or late in the sequence to determine what types of order effects obtain when many data are observed.

A third future experiment will examine one prediction of the SAM-inspired version of the HyGene model described previously. In that theory, it was argued that recency obtains in SBS cases and not in EOS cases because context changes more in the SBS case than in the EOS case due to more time passing. In this experiment, time would be equalized between EOS and SBS conditions by inserting a distractor task between each item in the EOS case. Then, like generation in the SBS condition, context will change significantly between observing cue 1 and the final cue. It is predicted that participants in this EOS condition with a distractor task would demonstrate recency.

Finally, the two versions of the model that are supported by the current study will be tested. First, simulations of Sequential HyGene using the cue activation function will be implemented to verify that they predict results comparable with those obtained in current study. We will examine predictions of this version of the model under several different assumptions: Cues do not take up space in WM; Cues do take up space in WM and cause hypotheses to be kicked out; Cues do take up space in WM, and the cues and hypotheses in WM tend to maintain each other. Second, the SAM-buffer version of the model will be implemented to learn exactly what predictions this version of the model makes.

Applications of Research

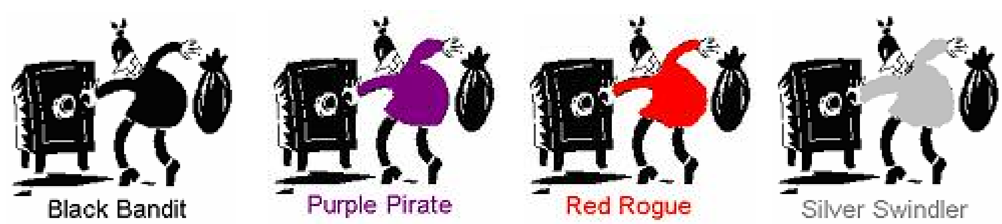
The current research has potential applications for decision-making tasks in which decision makers receive information sequentially. Further, the research has applications for both decision makers (such as clinicians) as well as those dependent on decision makers (such as patients). First, consider the finding that when the most diagnostic information is presented early or late in the sequence, no order effects obtain when participants generate hypotheses after viewing all evidence. For those dependent on decision makers, this finding suggests that when presenting information, people should attempt to position the most informative evidence early or late in the sequence, and should not to interrupt the decision maker by asking for intermediate hypotheses. Interrupting the decision maker may lead them to overweight recent information. Decision makers need to be aware that they may be biased by the order that information is presented, and perhaps should try to use counterfactual reasoning skills by asking themselves if they would generate the same hypotheses if the information was presented in various other possible orders.

The purpose of this research was to investigate how people generate hypotheses in response to data they directly observe in the environment as the data unfold over time. Understanding how people generate hypotheses is as important as understanding how people evaluate and choose among hypotheses. Most decisions begin with structuring of the decision. Structuring requires specifying: possible states of the world or hypotheses, possible action alternatives, and outcomes that will result if a particular action is chosen (Gettys & Fisher, 1979). Structuring the decision is a major determinant of the quality of the ultimate decision, and has been neglected in research. If a physician never *generates* the correct disease diagnosis, testing hypotheses optimally will still not lead to optimal results.

Appendix A

Instructions for Experiment 1:

Imagine you are a detective. Your specialty is catching thieves. You just moved to the city of Crimeville, and therefore are going to review case files in order to learn about robberies that have been committed in the past year. Knowing the history of the city's crimes will help you in the future when you need to predict who committed a robbery. All of the robberies in Crimeville have been committed by one of four different criminals pictured below:



Each of these thieves always acts alone.

You will learn 4 things about each robbery committed in last year:

1. the location:

2. the kind of job:

3. the value of items stolen:

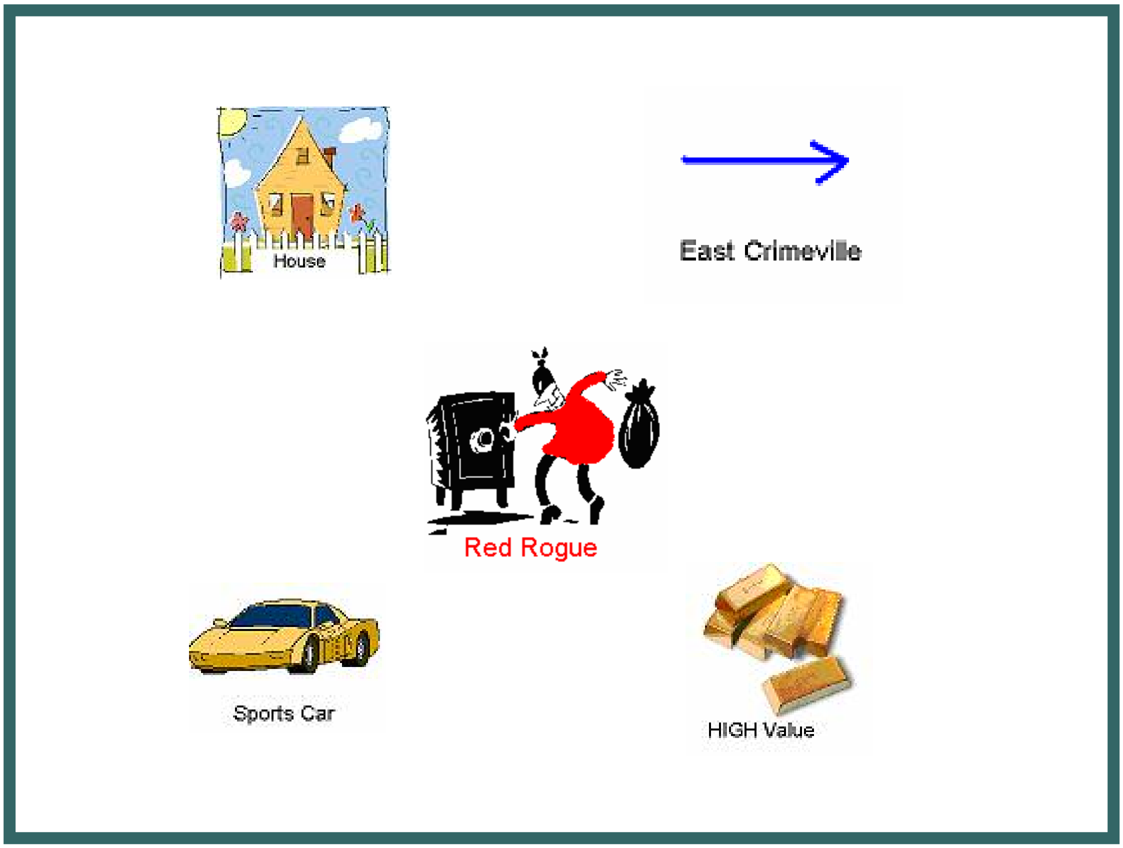
4. the kind of getaway car:

Each time you view a new record, you will see the name of the criminal who committed the robbery and details about that theft on the computer screen. ***To move to the next record, you must press the first letter of the name of the criminal.*** For instance, if the Black Bandit committed the crime, you must press 'B' to continue. Be sure to pay attention so that you will be able to predict who commits future robberies based on the details of the crime you are given.

Now participants view the 200 records. Then they read: You have finished studying the case files and are now ready to get to work. A robbery has occurred, and you have been called in to figure out who committed the robbery. As you learn details about the crime scene, use your knowledge of past crimes to generate a list of your top suspects. Your list of top suspects should include those people you would want to bring in for questioning, because you think they are quite likely to have committed the robbery. Sometimes your list of top suspects may include all of the 4 robbers you have learned about because the information you receive doesn't help you narrow down the possibilities. Other times, your list of top suspects may include a subset of the 4 robbers you have learned about because the information you receive helps you narrow down the possibilities. On the next screen you will be given a piece of evidence about the crime scene. Based on this piece of evidence, list the suspect(s) you are considering. To "list" a suspect, you must simply press the first letter of that suspect's name (i.e., press 'S' for the Silver Swindler; 'R' for the Red Rogue; 'B' for the Black Bandit; 'P' for the Purple Pirate). Continue typing the first letter of suspects until you have listed each suspect you are currently considering on your top-suspect list. When you have finished, press the spacebar. List only those suspect(s) that you are considering, whether that be 1, 2, 3, or all 4 of the robbers of Crimeville.

Before generating hypotheses for the second case, participants read: A new robbery has occurred! You are again called in to figure out who the robber was. You will be given pieces of evidence about this new crime scene, and then asked to list the suspect(s) you are considering by pressing the first letter of their name (i.e., press 'S' for the Silver Swindler; 'R' for the Red Rogue; 'B' for the Black Bandit; 'P' for the Purple Pirate). Again, list only those suspect(s) that you are considering, whether that be 1, 2, 3, or all 4 of the robbers of Crimeville.

Appendix B



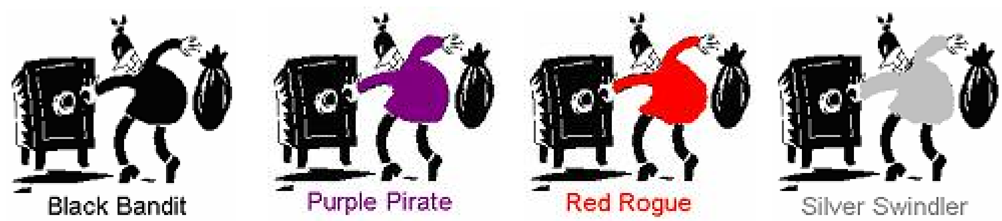
Appendix C

"First, you learn that the robber drove away in a SPORTS CAR. Please indicate each of the suspect(s) you are including in your suspect list by typing the first letter of the suspect's name. When you have finished indicating all of the suspect(s) on your list, press the spacebar."

Appendix D

Instructions for Experiment 2:

Imagine you are a detective. Your specialty is catching thieves. You just moved to the city of Crimeville, and therefore are going to review case files in order to learn about robberies that have been committed in the past year. Knowing the history of the city's crimes will help you in the future when you need to determine who committed a robbery. All of the robberies in Crimeville have been committed by one of four different criminals pictured below:



Each of these thieves always acts alone.

You will learn 4 things about each robbery committed last year:

5. the location:

6. the kind of job:

7. the value of items stolen:

8. the kind of getaway car:

Each time you view a new record, you will see the name of the criminal who committed the robbery and details about that theft on the computer screen. ***To move to the next record, you must press the first letter of the name of the criminal.*** For instance, if the Black Bandit committed the crime, you must press 'B' to continue. Be sure to pay attention so that you will be able to determine who commits robberies in the future based on the details of those crimes.

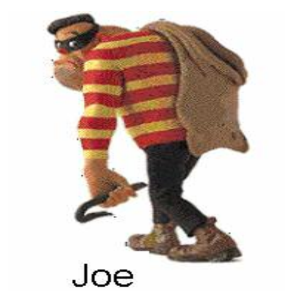
Now participants view the 200 records. Then they read : You have finished studying the case files and are now ready to get to work. A robbery has occurred, and you have been called in to help figure out who committed the robbery. You will now learn details about the crime scene. You will be asked to evaluate the chance that a particular robber (as compared to other possible robbers) committed this crime. Use your knowledge of past crimes to evaluate the chance that the suspect in question committed this crime. On the next screen you will be given a piece of evidence about the crime scene. Then you will be asked the chance that a particular robber committed the crime based on the evidence presented to you. Then you will be given other pieces of evidence. After each piece of evidence, you will be asked to estimate the chance that the robber committed the crime. When asked to judge the chance that a given robber committed this crime, please respond on a chance scale from 0 to 100. A judgment of 0 means that there is NO CHANCE that the suspect committed this crime. A judgment of 100 means that it is ABSOLUTELY CERTAIN that the suspect committed this crime. A response of 50 means that the outcome has the same chance as a coin flip landing on heads rather than tails. You can use ANY number between 0 and 100. Use the number pad at the right of your keyboard to make your responses.

Before judging hypotheses for the second case, participants read: A new robbery has occurred! You are again called in to figure out who the robber was. You will be given pieces of evidence about this new crime scene, and then asked the chance that a particular robber committed the crime based on the evidence presented to you.

Appendix E

Instructions for Experiment 3:

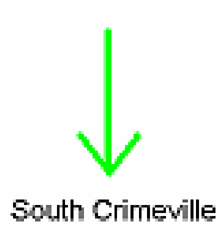
Imagine you are a detective. Your specialty is catching thieves. You just moved to the city of Crimeville, and are going to review case files in order to learn about robberies that have been committed in the past year. Knowing the history of the city's crimes will help you in the future when you need to predict who committed a robbery. All of the robberies in Crimeville were committed by the criminals pictured below:



Each of these thieves always acts alone.

You will learn 4 things about each robbery committed last year:

1. where the robbery took place:



2. what kind of business was robbed:

3. how the robber disguised himself:

4. how the robber got away from the robbery:

Each time you view a new record, you will see the name of the criminal who committed the robbery and details about that theft on the computer screen. ***To move to the next record, you must press the first letter of the name of the criminal.*** For instance, if the Billy committed the crime, you must press 'B' to continue. Be sure to pay attention so that you will be able to predict who commits future robberies based on the details of the crime you are given.

Now participants view the 160 records. Then they read : You have finished studying the case files and are now ready to get to work. A robbery has occurred, and you have been called in to figure out who committed the robbery. As you learn details about the crime scene, use your knowledge of past crimes to generate a list of your top suspects. Your list of top suspects should include those people you would want to bring in for questioning, because you think they are quite likely to have committed the robbery. Sometimes your list of top suspects may include many of the robbers you have learned about because the information you receive doesn't help you narrow down the possibilities. Other times, your list of top suspects may include few robbers you have learned about because the information you receive helps you narrow down the possibilities. On the next screens you will be given pieces of evidence about the crime scene, one at a time. Based on the evidence you see list all of the suspect that you think have potentially committed this crime.

Before generating hypotheses for the second case, participants read: A new robbery has occurred! You are again called in to figure out who the robber was. You will be given pieces of evidence about this new crime scene, and then asked to list the suspect(s) you think have potentially committed this robbery.

Appendix F

Instructions for Experiment 4, Step-by-Step condition: In this experiment, your job is to guess what undergraduate psychology course(s) are being described by words presented on the computer screen. The words you will see are listed in undergraduate psychology course descriptions. You will be presented with a sequence of these course description words, one after another. After seeing each word, you will be asked to list the courses you think are best described by the words you have seen. After seeing each new word, consider that word in addition to the other words you previously saw when coming up with your list of possible courses. Sometimes you may only be able to think of one course described by the words you have seen thus far, and other times you may be able to think of many courses described by the words you have seen thus far. That is fine—your job is to try to list all of the courses that you think are likely candidates given the words you have seen.

Instructions for Experiment 4, End-of-Sequence condition: In this experiment, your job is to guess what undergraduate psychology course(s) are being described by words presented on the computer screen. The words you will see are listed in undergraduate psychology course descriptions. You will be presented with a sequence of these course description words, one after another. After seeing all of the words, you will be asked to list the courses you think are best described by the words you have seen. Try to consider all words you have seen when coming up with your list of possible courses that are being described. Sometimes you may only be able to think of one course described by the words you have seen thus far, and other times you may be able to think of many courses described by the words you have seen thus far. That is fine—your job is to try to list all of the courses that you think are likely candidates given the words you have seen.

Appendix G

Cognitive

- Experimental psychology: cognitive processes
- Research methods: cognitive psychology
- Research in Cognitive Psychology
- Introduction to Memory and Cognition
- Brain and Memory Psychology
- Cognitive and Behavioral Psychology
- Decision Making
- Thinking and Problem Solving
- Cognitive Neural Psychology
- Seminar in Animal Intelligence
- Psychology of the Complex Mind
- Cognitive Functioning
- Psychology of Language
- Mental Psychology
- Communication and Memory

Bio/Neuropsychology

- Special Topics in Psychology: Introduction to Neuroscience Through Sleep
- Experimental Psychology of Sensory Processes
- Neurobiological Psychology
- Developmental Biopsychology
- Honors Seminar in Biopsychology of Aggression
- Biological Basis of Behavior
- Psychology of the Brain
- Biological Basis of Behavior Laboratory
- Neural Systems Behavior
- Perception
- Genetic Psychology

Developmental

- Advanced Psychology 1: Death and Dying
- Applied Developmental Psychology
- Research Methods in Developmental Psychology
- Life-Span Development
- Child Psychology
- Adult Psychology
- Adolescent Psychology
- Infant Psychology
- Psychology of Adulthood and Aging
- Psychology of Adolescents and Adulthood

Counseling/Clinical

Special Topics in Psychology: Abnormal Child Psychology
Introduction to Clinical Psychology
Introduction to Community Psychology
Family and Community Psychology
Psychology of Adjustment
Introduction to Counseling Psychology
Abnormal Psychology
Mental Health
Severe Mental Disorders: Etiology and Treatment
Psychiatry/Psychiatric Therapy
Assessment of Drug Therapy vs. Behavioral Therapy in Resolving Abnormal Behavior
Psychology of Drug Treatment
Theories of the Link Between Drugs and Abnormal Behavior
Substance Abuse and Drug Abuse
Special Topics in Psychology: Assessment and Treatment of Addictive Behaviors
Introduction to Behavioral Pharmacology
Community Interventions, Theory, and Research
Basic Helping Skills Lab
Basic Helping Skills
Community Interventions: Service Learning
Experiential Learning
Field Experience: Special Assignments for Honors Students
Psychotherapy Course
Clinical Behavioral Specialization Course
Domestic Violence
Humanistic Psychology
Psychoanalytic Theory

Social

Advanced Psychology II: Personality Development
Advanced Social Psychology
Personality Theories
Experimental Psychology: Social Processes
Social Psychology
Communication and Persuasion
Applied Psychology in Social Relationships
Intrapersonal Psychology
Psychology of Interpersonal Relationships
Intimate Relations
Family Studies

Industrial/Organizational	
	Psychology of Organizational Processes
	Survey of Industrial Organizational Psychology
	Psychology of Motivation Attitudes
	Field Research in Organizational Psychology
	Psychological Foundations of Personnel Selection
	Training
	Psychology of Leaders in Work Organizations
	Negotiation
	Cross Cultural Negotiation
Introduction to Psychology	
	Introduction to Psychology
	Psychology 101
	Modern Psychological Theories
	Psychology
Research Psychology	
	Research Methods
	Experimental Psychology
	Research Labs
	Statistical Methods in Psychology
	Principles of Psychological Testing
Other	
	Special Topics in Psychology: Major Transitions from Undergraduate to Professional
	Independent Study in Psychology
	Senior Seminar
	History of Psychology
	Special Research Problems in Psychology
	Higher Level 400 Courses
	Research in Memory, Cognition, and Counseling
	Special Topics: CyberPsychology
	Environmental Ecological Psychology
	Introduction to Evolutionary Theory
	Animal Behavior
	Cultural Context of Psychological Development
	Psychology of Human Sexuality
	Honors Thesis Proposal Preparation
	Ericsonian Course
	All Non-Application Psychology Courses
	Any Psychology Course
	Jungian Course
	Freudian Course
	Adoptive Psychology
	Behavioral Psychology
	Psychology 140

Applied Psychology / Practical Psychology
Medical Psychology
UNIV
Biology 105
Food and Nutrition
KNES
Psychology 105
Biology
Chemistry
Educational Psychology
American Studies
Social Work Courses
Social Work Graduate Courses
Sociology
Psychology 340
Paranormal Psychology
Social Cognition
Sleeping
Psychology 344
Psychology 300
Anatomy

Individual Differences

Psychology of Women
Cross-Cultural Psychology
Psychology of Individual Differences
Gender Psychology

References

Adelman, L., Bresnick, T., Black, P. K., Marvin, P. K., & Sak, S. G. (1996). Research with Patriot Air Defense officers: Examining information order effects. *Human Factors, 38,* 250-261.

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

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

Anderson, N. H., & Hubert, S. (1963). Effects of concomitant verbal recall on order effects in personality impression formation. *Journal of Verbal Learning and Verbal Behavior, 2,* 379-391.

Asch, S. E. (1946). Forming impressions of personality. *Journal of Abnormal Social Psychology, 41,* 258-290.

Barrows, H. S., Norman, G. R., Neufeld, V. R., & Feightner, J. W. (1982). The clinical reasoning of randomly selected physicians in general medicine practice. *Clinical and Investigative Medicine, 5,* 49-55.

Berbaum, K. S., et al. (1991). Time course of satisfaction of search. *Investigative Radiology, 26,* 640-648.

Beyth-Marom, R., & Arkes, H. R. (1983). Being accurate but not necessarily Bayesian: Comments on Christensen-Szalanski and Beach. *Organizational Behavior and Human Performance, 31,* 255-257.

Botti, M., & Reeve, R. (2003). Role of knowledge and ability in student nurses' clinical decision making. *Nursing and Health sciences, 5,* 39 – 49.

Burgess, C., & Lund, K. (1997). Modeling parsing constraints with high-dimensional context space. *Language and Cognitive Processes, 12,* 177-210.

Carroll, J. S., & Siegler, R. S. (1977). Strategies for the use of base-rate information. *Organizational Behavior and Human Performance,19,* 392-402.

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology, 4,* 55-81.

Christensen-Szalanski, J. J. J., & Beach, L. R. (1982). Experience and the base-rate fallacy. *Organizational Behavior and Human Performance,29,* 270-278.

Christensen-Szalanski, J. J. J., & Beach, L. R. (1983). Believing is not the same as testing: A reply to Beyth-Marom and Arkes. *Organizational Behavior and Human Performance, 31,* 258-261.

Christensen-Szalanski, J. J. J. & Bushyhead, J. B. (1981). Physicians' use of probabilistic information in a real clinical setting. *Journal of Experimental Psychology: Human Perception and Performance, 7,*928-935.

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

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

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing By Latent Semantic Analysis. *Journal of the American Society For Information Science, 41*, 391-407.

Dougherty, Harbison, Sprenger, & Thomas (2007). The sequential sampling model of HyGene. In preparation.

Dougherty, M. R. P. & Hunter, J. E. (2003a). Hypothesis generation, probability judgment, and individual differences in working memory capacity. *Acta Psychologica, 113,* 263 – 282.

Dougherty, M. R. P. & Hunter, J. E. (2003b). Probability judgment and subadditivity: The role of working memory capacity and constraining retrieval. *Memory & Cognition, 31,* 968 – 982.

Dougherty, M. R. P. (2001). Integration of the ecological and error models of overconfidence using a multiple-trace memory model. *Journal of Experimental Psychology: General, 130,* 579 – 599.

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

Dougherty, M. R. P., Gettys, C. F., Thomas, R. P. (1997). The role of mental simulation in judgments of likelihood. *Organizational Behavior & Human Decision Processes, 70,* 135-148.

Dougherty, M. R., & Sprenger, A. (2006). The influence of improper sets of information on judgment: How irrelevant information can bias judged probability. *Journal of Experimental Psychology, General, 135,* 262-281.

Elstein, A. S., Shulman, L. S., & Sprafka, S. A. (1978). *Medical problem solving: An analysis of clinical reasoning*. Cambridge: Harvard University Press.

Elstein, A. S., & Schwartz, A. (2006). Evidence base of clinical diagnosis: Clinical problem solving and diagnostic decision making: Selective review of the cognitive literature. *British Medical Journal, 324,* 729-732.

Ericsson, K. A., & Polson, P. G. (1988). An experimental analysis of the mechanisms of a memory skill. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14,* 305-316.

Farris, H. H., & Revlin, R. (1989). Sensible reasoning in two tasks: Rule discovery and hypothesis evaluation. *Memory & Cognition, 17,* 221-232.

Fischhoff, B. (1977). Perceived informativeness of facts. *Journal of Experimental Psychology: Human Perception and Performance, 3,* 349-358.

Fischhoff, B., & Beyth-Marom, R. (1983). Hypothesis evaluation from a Bayesian perspective. *Psychological Review, 90,* 239-260.

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

Gettys, C. F., Fisher, S. D., & Mehle, T. (1978). Hypothesis generation and plausibility assessment (Annual Rep. TR 15-10-78). University of Oklahoma, Decision Processes Laboratory. (Can be obtained from the Defense Documentation Center ADA060-787).

Gettys, C. F., & Fisher, S. D. (1979). Hypothesis plausibility and hypothesis generation. *Organizational Behavior and Human Decision Processes, 24,* 93-110.

Gettys, C. F., Pliske, R. M., Manning, C., & Casey, J. T. (1987). An evaluation of human act generation performance. *Organizational Behavior and Human Decision Processes, 39,* 23 – 51.

Gigerenzer, G., Hell, W., & Blank, H. (1988). Presentation and content: The use of base rates as a continuous variable. *Journal of Experimental Psychology: Human Perception and Performance, 14,* 513-525.

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

Gordon, B. L. (1970). Terminology and content of the medical record. *Computational Biomedical Research, 3,* 436-444.

Greene, R. L. (1986). A common basis for recency effects in immediate and delayed recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 12,* 413-418.

Hendrick, C., & Costantini, A. F. (1970). Effects of varying trait inconsistency and response requirements on the primacy effect in impression formation. *Journal of Personality and Social Psychology, 15,* 158-164.

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

Joseph, G. M., & Patel, V. L. (1990). Domain knowledge and hypothesis generation in diagnostic reasoning. *Medical Decision Making, 10,* 31-46.

Klahr, D., & Dunbar, K. (1988). Dual space search during scientific reasoning. *Cognitive Science, 12,* 1-48.

Koehler, D. J. (1994). Hypothesis generation and confidence in judgment. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 461-469.

Koppenaal, L. & Glanzer, M. (1990). An examination of the continuous distractor task and the 'long-term recency effect'. *Memory & Cognition, 18,* 183-195.

Libby, R. (1985). Availability and the generation of hypotheses in analytical review. *Journal of Accounting Research,* 23, 646-665.

Medin, D. L., & Edelson, S. M. (1988). Problem structure and the use of base-rate information from experience. *Journal of Experimental Psychology, General, 117*, 68-85.

Mehle, T. (1982). Hypothesis generation in an automobile malfunction inference task. *Acta Psychologica, 52,* 87-106.

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

Pei, B. K., & Tuttle, B. M. (1999). Part-set cueing effects in a diagnostic setting with professional auditors. *Journal of Behavioral Decision Making, 12,* 233-256.

Peterson, C. R., & DuCharme, W. M. (1967). A primacy effect in subjective probability revision. *Journal of Experimental Psychology, 73,* 61-65.

Poltrock, S. E., & MacLeod, C. M. (1977). Primacy and recency in the continuous distractor paradigm. *Journal of Experimental Psychology: Human Learning & Memory, 3,* 560-571.

Raaijmakers, J. G. W., & Shiffrin, R. M. (1981). Search of associative memory. *Psychological Review, 88,* 93-134.

Rapaport, A., & Wallsten, T. S. (1972). Individual decision behavior. *Annual Review of Psychology, 23,* 131-176.

Rundus, D. (1971). Analysis of rehearsal processes in free recall. *Journal of Experimental Psychology, 89,* 63-77.

Slovic, P., & Lichtenstein, S. (1971). Comparison of Bayesian and regression approaches to the study of information processing in judgment. *Organizational Behavior and Human Performance, 6,* 649-743.

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

Steyvers, M., Griffiths, T. L., & Dennis, S. (2006). Probabilistic inference in human semantic memory. *Trends in Cognitive Sciences, 10,* 327-334.

Thomas, R., Dougherty, M. R., & Sprenger, A. (2007). Diagnostic hypothesis generation and human judgment. Under Revision.

Tversky, A., & Koehler, D. J. (1994). Support theory: A nonextensional representation of subjective probability. *Psychological Review, 101,* 547-567.

Vermande, M. M., van den Bercken, J. H., De Bruyn, E. E. (1996). Effects of diagnostic classification systems on clinical hypothesis-generation. *Journal of Psychopathology & Behavioral Assessment, 18,* 49-70.

Wallsten, T. S. (1981). Physician and medical student bias in evaluating diagnostic information. *Medical Decision Making, 1,* 145-164.

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology, 12,* 129-140.

Weber, E. U., Böckenholt, U., Hilton, D. J., & Wallace (1993). Determinants of diagnostic hypothesis generation: Effects of information, base rates, and experience. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19,* 1151 – 1164.

Windschitl, P.D., & Wells, G.L. (1998). The alternative-outcomes effect. *Journal of Personality and Social Psychology, 75,* 1441-1423.