HYPOTHESIS-GUIDED TESTING BEHAVIOR: THE ROLE OF GENERATION, METACOGNITION, AND SEARCH

A Dissertation
Presented to
The Academic Faculty

by

David A. Illingworth

In Partial Fulfillment
of the Requirements for the Degree
DOCTOR OF PHILOSOPHY in the
SCHOOL OF PSYCHOLOGY

Georgia Institute of Technology
May 2020

HYPOTHESIS-GUIDED TESTING BEHAVIOR: THE ROLE OF GENERATION, METACOGNITION, AND SEARCH

Approved by:

Dr. Rick Thomas, Advisor
School of Psychology
Georgia Institute of Technology

Dr. Dobromir Rahnev
School of Psychology
Georgia Institute of Technology

Dr. Jamie Gorman
School of Psychology
Georgia Institute of Technology

Dr. Karen Feigh
School of Aerospace Engineering
Georgia Institute of Technology

Dr. Christopher Hertzog
School of Psychology
Georgia Institute of Technology

Date Approved: June 17, 2019

Value is a peculiar construct. It can be defined subjectively, or defined by markets. Its meaning bends to context when invoked to define risk, personal belief, or monetary worth. One uniquely insightful way to define value is to consider that which is sacrificed to acquire a resource: The opportunities missed, the assets traded, the old goods discarded to make room for the new. This dissertation is dedicated to all that lay in the wake of this document that may better characterize the value of the following pages than the words that fill them.

ACKNOWLEDGEMENTS

I am grateful for the experiences I have gathered working with a diverse and talented array of scholars on this and other related projects. I would especially like to thank Dr. Rick Thomas, the chair of my committee. This work would not have been possible without his guidance or the investment he has made in my development as a scientific thinker. I want to express my gratitude to my committee members, Dr. Jamie Gorman, Dr. Christopher Hertzog, Dr. Dobromir Rahnev, and Dr. Karen Feigh, for the patience and insight they have afforded me throughout every stage of this project.

I would like to offer special thanks to Dr. William Shadish, who, although no longer with us, inspired and encouraged me to pursue my current path.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

Hypothesis testing is the act of acquiring information to challenge or promote a decision-maker's beliefs (i.e., hypotheses) in diagnostic tasks. To date, theorists have conceptualized this behavior as a consequence of implementing one of many possible heuristics for selecting tests, each tailored to optimize some task-relevant goal (e.g., reduce the likelihood of an erroneous diagnosis). Heuristics can account for a number of observed testing phenomena (e.g., pseudo-diagnostic search), but have difficulty explaining more nuanced testing behavior such as decisions to terminate data acquisition. Moreover, current theory has yet to address how updating a decision-maker's beliefs influences test preference, as hypothesis testing is often studied independent of other events inherent to hypothesis evaluation (e.g., hypothesis generation).

The theoretical perspective evaluated in this dissertation incorporated both environmental factors and cognitive mechanisms into the evaluation of information sources. That is, test selection was conceptualized as a consequence of a decision-maker's experience and the limitations of their cognitive abilities, as well as contextual constraints such as the cost to access data and incentives for performing a task accurately. I cast hypothesis testing as a special case of generalized cognitive search. Thus, selection of a test occurs because of the perceived tradeoffs between the value of available information depositories and costs associated with exploiting those depositories. The key facet of this theoretical perspective was the hypothesis-guided testing hypothesis, which posited that the state of one's beliefs is critical to changes in testing behavior over time.

The empirical testbed outlined in this document implemented a wide range of manipulations that initiated three adjacent, yet independent, lines of inquiry into nuanced

hypothesis testing behavior. Experiments 1 and 2 bridged the gap between hypothesis generation and test selection. Alternative accounts of information valuation assume that subjective estimates manifest independent of belief states and, as such, are static over time. Experiment 1 pitted that perspective against a hypothesis-guided valuation process by systematically manipulating cues to prompt the consideration of differential sets of hypotheses in a simulated diagnostic task. Although no relation was found between this manipulated and test selection, simulations showed that a measurement-level representation of memory (a core feature of the HyGene architecture) could account for the recorded behavior. To further evaluate the relation between generation dynamics and test selection, Experiment 2 manipulated the presence of time pressure in a diagnostic task—a factor known to truncate hypothesis generation. Once more, hypothesis set cuing had no effect on patterns of hypothesis testing. No effect emerged for time pressure, nor was one detected via simulation of the task.

Experiments 3 and 4 explored the role of beliefs in decisions to terminate testing behavior. The field lacked a comprehensive account of termination decisions in sequential data acquisition. Moreover, current theory has invoked untenable psychological mechanisms to explain how people valuate information and terminate information acquisition (e.g., Ficic & Buckman, 2015; Nelson, 2005; Nelson, McKenzie, Cotrell, & Sejnowski, 2010). Related work in metacognition has shown that self-monitoring judgments predict task switching and latencies to terminate information acquisition (Glucksbert & McCloskey, 1981; Klin, Guzman, & Levine, 1997; Kolers & Palef, 1976; Singer, 1984). Experiment 3 investigated metacognitive self-monitoring within the context of hypothesis testing and termination decisions. Specifically, self-monitoring judgments

were elicited and test outcome diagnosticity was manipulated to test the hypothesis that beliefs, as measured by judgments of knowing, play a crucial role in decisions to terminate testing behavior. A strong relation between metacognitive self-assessment and the duration of testing was found. Model fitting suggested that participants adopted a conservative threshold for terminating testing in Experiment 3 and were sensitive to the information inherent to the task. Experiment 4 exposed hypothesis testing to ecological factors ubiquitous in search environments to evaluate test selection and termination decisions within a general cognitive search perspective. Specifically, frame, acquisition cost, and cost experience were manipulated to explore if and how these ecological factors influence termination decisions and subsequent diagnostic decisions. Participants' sensitivity to cost was such that testing behavior was abbreviated as the environment imposed higher expenses for acquiring data. Moreover, the impact of the cost manipulation was dependent on experience. Participants who experienced low costs early in the study generally engaged in more testing behavior than those who experienced high costs in the early stages of the experiment.

The reported studies provided limited evidence in support of the hypothesis-guided testing hypothesis. No statistical model found an effect related to manipulations of hypothesis set cuing. However, simulations of the task suggest that participants were sensitive to the diagnostic properties of tests—an indication that memory was playing a role in their test selection process. Experiments 3 and 4 provided insight regarding the stopping rule used to terminate testing behavior. Both internal and external factors were shown to be related to termination decisions.

Taken together, the reported work advanced the nature of empirical evaluations of hypothesis testing theory by imposing complex environmental structures and recording patterns in decisions to terminate testing. The experiments initiated three lines of research that necessitate further inquiry into the role of memory, metacognitive self-assessment, and ecological factors in hypothesis testing.

CHAPTER 1. INTRODUCTION

Many instances of human judgment occur in environments that are high in uncertainty, such that multiple hypotheses or beliefs are initially considered as candidate explanations for a set of observations. Decision-makers will often engage in data acquisition under such circumstances for the purpose of informing and improving the accuracy of their judgments. Thus, the process by which people decide to engage their environment to test how well their mental representations match reality is critical to understanding the outcomes and dynamics of judgment and decision-making behavior. Birthed in philosophy of science to describe strategies for empirical inquiry, *hypothesis testing* has emerged as an important psychological construct—a ubiquitous cognitive mechanism invoked in a wide array of human behaviors.

The study of *hypothesis testing* has taken many forms since it first appeared in the psychological literature in the 1960s. Poletiek (2001), for instance, conceptualized hypothesis testing as a number of stages, beginning with the generation of hypotheses and culminating in the integration of information and revision of beliefs. The focus of this dissertation is the intermediary step defined by Poletiek as the selection or design of a test "whose outcome is expected to reveal something about the truth status of [a] hypothesis" (p. 2).

1.1 Statement of the Problem

Theoretical accounts of information search and hypothesis testing have grown in complexity, such that contemporary models of these phenomena address the utility of potential answers in addition to the nature of the queries people formulate (e.g., confirming, falsifying) prior to making decisions (Johnson-Laird & Byrne, 1991; Nelson, 2005; Nelson,

McKenzie, Cottrell, & Sejnowski, 2010). However, current theory has yet to fully integrate the impact of one's beliefs on the perceived value of tests, ignoring the reason people engage in testing at all: To evaluate hypotheses under consideration. Moreover, those who study hypothesis testing typically limit its context to the hypothesis evaluation process described by Poletiek (2001), ignoring rich bodies of literature emerging from adjacent research programs. For instance, hypothesis testing has seldom been discussed in relation to the information foraging literature, which has ties to generalized accounts of search that cover a range of behaviors spanning from the time a squirrel spends searching for nuts in a tree to the hyperlinks people click while surfing the web. In essence, hypothesis testing is a special case of information foraging and, as such, can be defined as a strategy for information acquisition that reveals new data for the purpose of evaluating one's beliefs.

To date, researchers investigating hypothesis testing continue to conduct their science within a narrow scope that limits the knowledge that can be gleaned from experimentation and inhibits the development of more holistic explanations for observed phenomena. The empirical work reported here expands and evaluates a memory-based perspective of hypothesis evaluation that accounts for the cognitive processes underlying observable testing behavior while acknowledging the constraints placed on data acquisition in complex decision environments. Specifically, this work leveraged and expanded upon the HyGene cognitive architecture (Thomas, Dougherty, Springer, & Harbison, 2008) to evaluate a straightforward thesis: The generation of beliefs during diagnostic tasks, and the memory dynamics that govern that process, played a primary role in the formation and exploitation of tests, as well as in decisions to terminate testing behavior.

1.2 Research questions

The goal of this dissertation was to address the gap in the hypothesis testing literature regarding the role of competing hypotheses by investigating hypothesis generation processes within the context of data acquisition. The work carried out to achieve this goal was designed to answer three broad research questions relating generation processes to test selection and decisions to terminate testing.

1.2.1 What is the nature of the relation between hypothesis generation and test preferences?

Prior research began to tie hypothesis testing behavior to generation processes by showing that the number of hypotheses considered by a decision-maker determines whether or not diagnostic tests are preferred (Lange, Thomas, & Dougherty, 2010). Lange et al. have argued that a positive test bias occurs when decision-makers generate only a single hypothesis. However, the HyGene architecture (Hypothesis-Generation; Thomas, Dougherty, & Buttaccio, 2014; Thomas, Dougherty, Harbison, & Sprenger, 2008) predicts that the specific set of hypotheses under consideration by a decision-maker should account for test preferences beyond the mere number of hypotheses believed to be in contention—hypothesis-guided testing. In other words, the HyGene architecture posits a dynamic, hypothesis-guided testing heuristic, where the value or preference exhibited for a test changes as a consequence of a decision-maker's beliefs. Experiments 1 and 2 systematically controlled the hypotheses cued by information present in the decision environment to test the hypothesis-guided testing hypothesis.

1.2.2 Are decisions to terminate testing behavior related to belief states?

Few have ventured to account for termination decisions within the context of information acquisition (c.f., Ficic & Buckman, 2015). In fact, the bulk of both theoretical

and empirical investigations of human search termination has focused on memory retrieval (Harbison, Dougherty, Davelaar, & Fayyad, 2009; Hills, Jones, & Todd, 2012; Hills & Pachur, 2012; Levy & Baddeley, 1971; Metcalfe & Murdock, 1981; Miller, Weidemann, & Kahana, 2012; Murdock & Okada, 1970; Raaijmakers & Shiffrin, 1981). An analogous body of literature investigating processing relevant to stopping decisions suggests that mechanisms related to metacognitive self-assessment may play a role in search termination, as it has been shown to predict restudy decisions, task switching, and exit latencies for general knowledge recollection (Glucksbert & McCloskey, 1981; Klin, Guzman, & Levine, 1997; Kolers & Palef, 1976; Singer, 1984). Experiment 3 is designed to initiate inquiry regarding the role metacognitive self-assessment in search termination, while simultaneously evaluating the predictions of numerous stopping rules for information acquisition (Ficic & Buckman, 2015). The result of this study should explicate the nature of any observed relation between belief states and termination behavior. Moreover, results should provide some clarity regarding the validity of existing stopping rules and information utility theory.

1.2.3 Are decisions to terminate hypothesis testing sensitive to the same ecological factors that influence general cognitive search?

One reason for limited theoretical advancement within the field of hypothesis testing may be a failure to reconcile understanding of testing behavior with broader research programs investigating information foraging and cognitive search. Few studies have explored hypothesis testing behavior within a foraging context, where the perceived value of a test is conceptualized as a function of its expected utility and the costs associated with acquiring test results. The goal of Experiment 4 was to address this gap in the

literature by investigating how factors ubiquitous in applied decision domains—such as costs of gathering information (e.g., time, monetary expenses), risks taken when pursuing unreliable sources of information, and changes in task context (i.e., the framing of outcomes as gains or losses)—influence decisions to terminate data acquisition.

1.3 Significance of the study

The results of the reported empirics were assessed with respect to novel computational modeling with the intent of elucidating the mechanisms involved in hypothesis testing. The model components evaluated with respect to the experiments reported in this document represent the latest advancement for a psychological construct rooted in the historical context that gave rise to cognitive science, building upon a number of theoretical accounts and computational models that have shaped the manner in which hypothesis testing is studied. The purpose of this dissertation was to evaluate the predictions of the HyGene architecture regarding the role of a decision-makers' beliefs in the formation and selection of hypothesis tests. The studies reported in this document have produced data necessary to explicate nuanced testing behavior, which should serve to challenge current accounts of this and related behavior (e.g., resource valuation). The scope of these studies affords an understanding of both internal (e.g., memory) and external (e.g., costs) factors that shape hypothesis testing behavior. In sum, this dissertation has evaluated the most holistic account of hypothesis testing to date.

Table 1. Notable advancements in hypothesis testing theory.

Achievement	Model	Authors
Psychology of hypothesis testing was first investigated.	n/a	Wason (1960)
Tradeoff of costs and resources determine the rate at which resources are acquired.	Marginal Value Theorem	Charnov (1976)
Formal theory of foraging behavior	Optimal Foraging Theory	Stephens & Krebs (1986)
Subjective utilities should be incorporated into study of hypothesis testing	n/a	Manktelow & Over (1990)
The results of tests have value, not tests themselves	n/a	Johnson-Laird & Byrne (1991)
Accounted for multiple goals of hypothesis testing: accumulating data in the face of costs and assessing the value of potential information	Probability Value Model	Poletiek (1995)
Utilized an instantiation of animal foraging theory to account for human foraging behavior	ACT-IF	Pirolli & Card (1999)
Integrated hypothesis testing with the hypothesis generation processes	HyGene-HT	Lange, Thomas & Dougherty (2010)

Table 1 lists notable achievements in the development of hypothesis testing theory. These studies represent a broad body of research that spans six decades and was distributed across a number of literatures, including several domains within ecological biology and psychology. This work has been reviewed in the following section, highlighting the application of findings from diverse settings to hypothesis testing.

CHAPTER 2. LITERATURE REVIEW

The seminal work of Wason (1960; 1966) is credited as the first foray into the psychology of hypothesis testing. Although the applicability of his earliest finding to contemporary hypothesis testing research is limited, the influence of this work on theoretical accounts of testing behavior and the nature of empirical investigations warrants a brief discussion. Wason invented two tasks to investigate the presence of falsification strategies in test selection: The rule discovery task and the Wason card task. These are widely regarded as tools better suited for studying reasoning; however, the results of these early experiments raised questions for which researchers spent decades in search of answers.

2.1 Hypothesis Testing

Popper (1959; 1963) proposed that the best scientific inquiries were those that could potentially falsify or refute a theoretical perspective. Wason (1960) perceived Popper's approach to empiricism and human decision-making as analogous endeavors and treated falsification strategies as a normative standard in his decision-making tasks. In his rule discovery task, for example, participants tasked with deciphering the rule that generated some number sequence (e.g., 2 – 4 – 6) by generating their own exemplars to be evaluated by an experimenter would behave normatively if they attempted to generate sequences that failed to adhere to the rule. He found that very few participants engaged in a falsification strategy, which was most frequently observed in those who correctly identified the rule early in the experiment. Most participants generated three-number sequences consistent with the rule under consideration, a strategy that came to be known as *confirmation bias*.

This initial discovery was further explored in Wason's (1966) card selection task, where participants were provided a rule to test regarding the stimuli they might find on either side of a set of four cards. Once more, Wason observed evidence of confirmation bias, which led him to conclude that people are inclined to adopt confirmatory test

strategies. Simple as they are, Wason's tasks inspired a large body of work investigating circumstances under which people complete such tasks in a manner consistent with falsification strategies, and what it is about how the tasks are presented that changes one's approach (for a review, see Poletiek, 2001). Ultimately, the worth of Wason's work lies in the questions he raised that persist in the literature: What is the nature of confirmation bias? If the nature of testing behavior changes in response to the context in which these problems are presented, what is the process by which people recognize the need for falsifying strategies?

Although many investigators began to deviate from Wason's tasks, evidence of confirmation bias or positive testing strategies remained a common finding (Beyth-Marom & Fischoff, 1983; Mynatt, Doherty, & Tweney, 1977; Trope, Bassok, & Alon, 1984). An important insight emerged from the work of Klayman and Ha (1987), who pointed out that Wason set up his task such that positive testing would mislead a participant since a number of reasonable hypotheses (e.g., values increasing by 2) could be embedded within the correct rule (e.g., increasing values) provided it was sufficiently broad. One can envision a scenario where a generated hypothesis only overlaps with the true state of the world, where any number of positive tests (exemplars that fit within the hypothesized rule) could falsify the hypothesis. Klayman and Ha used equation 1 to compute the probability that a positive test will falsify a hypothesis (z^+).

$$z^+ = \frac{p(\bar{t})}{p(t)} * z^- \qquad (1)$$

The ratio pits the probability that any exemplar falls within the correct set ($p(t)$) against the probability that any exemplar falls outside of the correct set ($p(\bar{t})$). Without knowing anything about the probability that a negative prediction will falsify the hypothesis (z^-),

one can see that the probability that a positive test will falsify a hypothesis (z^+) increases as the set of exemplars that fall outside of the correct set is large or the exemplars that fall within the set are rare. Thus, Klayman and Ha demonstrate that the severity of any testing strategy depends heavily upon the configuration of the environment, and, in some instances, positive test strategies are more advantageous than falsifying strategies.

2.2 Information Utility

While Klayman and Ha (1987) demonstrated the importance of the environment in hypothesis testing tasks, others have investigated the cognitive processes involved in testing. The most influential of these was recognition of the importance of valuation judgments related to information. Manktelow and Over (1990), for instance, were first to argue that subjective utilities needed to be accounted for in hypothesis testing tasks in place of the objective standards advanced by Wason and others. That is, normative perspectives of hypothesis testing (e.g., falsification or confirmatory strategies) ignored the subjectivity involved in evaluating potential tests.

Kirby (1994) recognized that the set size effect predicted by Klayman and Ha (1987) and the utility of tests described by Manktelow and Over (1990) were related constructs. Over four experiments, he demonstrated that an exemplar had less utility and, thus, was less likely to be used as a test when the probability of observing that exemplar increased. Kirby found that the frequency with which participants engaged in positive testing (as observed in Wason's selection task) was reduced in environments where the hypothesis set was large. Kirby argued that participants will seek out falsifying tests when the probability of observing results inconsistent with a considered hypothesis is greater than the probability of a consistent result. Recent tests of this phenomenon have branched outside of Wason's original selection task. Conceptualizing hypothesis size by the literal

size of a ship in a game of Battleship, Hendrickson, Navarro, and Perfors (2016) showed that people perceived misses or negative information (i.e., data indicating where ships are not located) as having greater utility for the purpose of locating large ships than hits. Alternatively, participants preferred positive information when ships were small.

Mental model theories of human reasoning have also argued against the formal rules introduced by Wason. For instance, Johnson-Laird and colleagues (Johnson-Laird, 2010; Johnson-Laird & Byrne, 1991) argued that the results of tests—not tests themselves—possess value, and it is the decision-maker's assessment of those values that determine their testing behavior. They reported a number of experiments demonstrating that people do not utilize formal rules to draw inferences from observable data. Instead, mental models of possible outcomes appear to drive such judgments. Johnson-Laird and Byrne suggest that null outcomes (i.e., those that may falsify a hypothesis) are fuzzy and, thus, are unlikely to be used to drive test selection.

Linking test selection to utilities gave rise to a number of models attempting to capture the information metric people adopt to estimate the utility of tests, as an increasing number of researchers rejected the notion that test selection employs confirming or falsifying strategies (Evans & Over, 1996; Kirby, 1994; Over & Evans, 1994; Over & Evans, 1996). Over and Evans (1994; 1996), for example, conceptualized the utility of a test as its capacity to make a hypothesis more probable. The algorithm to compute their information metric—*probability gain*—is given in equation 2.

$$probability\ gain = \frac{p(H|d)}{p(H)} \tag{2}$$

$$probability\ gain = p(H|d) - p(H) \tag{3}$$

As can be seen in equation 2, tests have greater *probability gain* as the probability of the hypothesis increases after observing the resulting data from the test (d). This metric is has appeared in alternative forms in the literature (Baron, 1985; Savage, 1954), as seen in equation 3. Despite the difference in form, the nature of the metric remains the same: the utility of a test increases with its ability to increase the probability of the hypothesis.

Oaksford and Chater (1996) suggested that the utility of a test could also be represented by its ability to reduce uncertainty. Equation 4 computes *information gain* in the form of a Kullback-Liebler (KL) distance, which can be conceptualized as the test's ability to change one's beliefs (Nelson, 2005).

$$KL\ distance = \sum p(d) * \sum p(H|d) * log \frac{p(H|d)}{p(H)} \quad (4)$$

$$information\ gain = \sum entropy(H) - entropy(H|d) \quad (5)$$

$$entropy = \sum p(H) * log \frac{1}{p(H)} \quad (6)$$

Oaksford and Chater demonstrated that *information gain* (Equation 5) and KL distance generate equivalent estimates of the utilities of available tests.

Some researchers take an alternative perspective, positing that people do not explicitly consider the possible outcomes prior to selecting a test (Poletiek, 1996). The degree to which the outcome of a test falsifies or confirms a hypothesis is only considered after the new piece of data has been acquired. Poletiek (2001) also theorized that the strategy utilized to select tests may be a function of which stage of hypothesis testing the decision-maker has reached. When decision-makers are early in the process, she believes positive testing strategies are used until confidence in their considered hypothesis increases. Decision-makers switch strategies, selecting falsifying tests to further increase

their confidence that their considered hypothesis is correct, once they have reached some threshold of confidence in their current belief—a pattern observed in Wason's (1960) early work in rule detection.

In an attempt to integrate much of the work reviewed thus far, Poletiek (1995; Poletiek & Berndsen, 2000; Poletiek, 2001) developed the probability value model of hypothesis testing (Equation 7). Dissatisfied with the psychological implausibility of severity and falsification strategies, Poletiek conceptualized hypothesis testing as emerging from the management of two unique information acquisition problems that she referred to as the symmetric and asymmetric problems. The symmetric problem supposes that the decision-maker is seeking out as much new information about hypotheses as possible, but must incur costs to do so. Thus, optimal decision-makers select the tests that are likely to provide the most information for the least costs. The asymmetric problem concerns the decision errors a decision-maker is willing to tolerate. This is captured with Poletiek's value parameter, which represents the test's ability to confirm or disconfirm a hypothesis.

$$SEUt = p(c) * v(c) + p(d) * v(d) - costs \qquad (7)$$

To illustrate the probability value model, assume a physician believes that her patient is suffering from appendicitis and, consequently, can predict that the patient will exhibit an upset digestive system and a mild fever. She possesses the resources to test either of these predictions, but can't differentiate the tests in terms of utility or costs (i.e., the likelihood of either sign is equal given that the patient has appendicitis, as is the effort needed to acquire either piece of information). Suppose that, a priori, an upset digestive system is a less probable observation and, thus, would provide greater evidence in support of the hypothesis, while a mild fever is more probable and less supportive. The physician

must now make a gamble of sorts, trading off between the probability of acquiring the desired information and the value of the information sought.

Equation 7 computes the subjective expected utility of a test by taking the difference between the test's ability to confirm (c) or disconfirm (d) a hypothesis and the costs associated with acquiring the information. The physician may place greater weight on the probability of observing the desired outcome by testing the mild fever prediction or place a greater weight on the value of the outcome by testing the upset digestive system prediction.

Hypothesis testing theory has grown considerably since its inception, transitioning from gross scale testing strategies (i.e., falsification, confirmation) to accounting for the probabilistic nature of acquired information relative to belief states and attempting to integrate goal-related preferences of decision-makers. However, hypothesis testing is analogous to a number of tasks that involve search of some kind, many of which have been studied in more varied contexts than hypothesis testing. Theoretical accounts of comparable behaviors, such as animal foraging, consider many more factors that have yet to be accounted for in the hypothesis testing literature.

Test selections are often assumed to occur independent of each other, which ignores the iterative nature of hypothesis testing in applied domains. The state of a decision-maker's beliefs is typically treated as the outcome criteria, despite the fact that hypothesis testing is always nested within some broader task. The physician, for example, may initially test for a mild fever in her appendicitis patient, but follow that up with a test of the upset digestive system before making treatment decisions that could result in surgery. This would suggest that larger-scale task outcomes, such as the accuracy of a diagnosis or

effectiveness of a treatment, influence testing decisions. Hypothesis testing theory should account for multiple test selections and sequential data acquisition, and address how the ultimate goals of the decision-maker impact testing; thus, hypothesis testing must adapt to account for a wider array of environmental variables. Luckily, this work can draw inspiration from much of the foundational work completed within other domains of search.

2.3 General Cognitive Search

Hills (2006) makes the bold claim that goal-directed cognition evolved following the development of an *area restricted search* mechanism, such that "what was once foraging in a physical space for tangible resources became, over evolutionary time, foraging in cognitive space for information related to those resources" (p. 4). Hills theorized that area restricted search emerges in all environments where the location of resources is correlated, or where they appear in clusters. Thus, organisms can exploit the presence of clusters by deploying strategies that allow them to maximize consumption of resources by optimizing the trade-off between those gains and the losses incurred by the act of acquiring said resources.

Hills (2006) argues that search behavior, and goal-directed behavior more generally, appears to be linked to dopaminergic systems important for signaling the detection of objects that may be of importance to organisms. For example, an influx of dopamine in the environment of microorganisms results in *tumbles* (i.e., perseverative turning in circular patterns), a behavior typically observed when microorganisms encounter a source of food (Stock & Surette, 1996). Similar mechanisms appear to influence search behavior in people. For instance, increased levels of dopamine cause enhanced control of visuomotor focus in visual search tasks as a consequence of perseveration of saccadic

movement around targets (Barrett, Bell, Watson, & King, 2004; Dursun, Wright, & Reveley, 1999). Thus, much like non-human animals will move in circular patterns after encountering patches of food, discovering important objects in a visual environment initiates area restricted search strategies for people. Dopamine-related pathologies corroborate the extension of area restricted search to humans, as overactive production of dopamine (e.g., drug addiction, schizophrenia) results in perseveration of thought (i.e., obsessions) and too little (e.g., ADD, ADHD) results in failures to control attention.

Information, berries, websites, and phone numbers can be found clustered in books, bushes, hyperlinks, and memories respectively. Several bodies of literature have been devoted to studying processes by which people seek these resources, revealing similar patterns of behavior across all domains of search. I reviewed these findings in the following sections, focusing on the computational models that continue to exhibit explanatory power as they are generalized to different areas of research.

2.3.1 Animal Foraging.

The body of literature investigating foraging behavior in non-human animals has a rich history, and is far more developed than the information search literature to be discussed later in this review. Optimal foraging theory (Stephens & Krebs, 1986) has evolved from a poorly specified verbal model based on (mostly) observational research to a complicated mathematical theory tested in a multitude of environments (Nonacs, 2001). Models spawned in the animal foraging literature now serve an integral role in building generalized theory for cognitive search.

Animal foraging is not unlike human decision-making, as animals must weigh the costs and benefits related to the selection of one option available to them relative to its

alternatives. Moreover, the resources that animals seek appear in clusters (e.g., fruit-bearing tree, berry bushes, herds of prey), which is a necessary property of an environment for exploitation via area-restricted search (Hills, 2006). To behave optimally, animals are predicted to choose patches of higher quality over those of poorer quality (Stephens and Krebs, 1986). Moreover, animals must maximize their patch residence time (PRT; or the duration of time spent searching for food at one location) for high-quality patches, and terminate their search in favor of a new patch once the cost of retrieving another morsel is greater than the cost of moving to a new location.

One of the most prolific and often cited models of optimal foraging is the Marginal Value Theorem (MVT; Charnov, 1976). This seminal work is the foundation of more complex models that span a diverse array of scientific domains such as memory research. MVT exemplifies how most models implement cost-benefit tradeoffs that occur in search, by expressing the utility of foraging as a function that maximizes gains (e.g., energy, information) relative to costs endured in the process. In the case of MVT, this trade-off (i.e., the rate of gain; $g_i(T_i)$) is represented as the product of the energy gained from a specific patch (net energy; E_n) and the total time spent traveling to and within this patch (T_i; see Equation 8). Thus, one may compute the optimal PRT for an animal foraging in any given patch (Equation 9).

$$g_i(T_i) = E_n * T_i \tag{8}$$

$$T_i = \frac{g_i(T_i)}{E_n} \tag{9}$$

MVT has a history of successfully predicting qualitative results from myriad foraging behaviors, including patch selection in armadillos and guinea pigs (Cassini et al, 1990), starlings (Cuthill et al., 1994), and pigeons (Hansen, 1987); reproduction in

hummingbirds (Pyke, 1978); and mating behavior in gibbons (Grether et al., 1992). However, a review of these and similar studies found that the quantitative patterns predicted by MVT are rarely produced empirically and, more often than not, animals in these studies spent less time than predicted foraging in patches of good quality and more time than predicted foraging in patches of poor quality (Nonacs, 2001). Such deviations from ideal behavior predicted by MVT would suggest that, on average, adopted foraging strategies are sub-optimal and potentially maladaptive. However, Nonacs (2001) proposed that MVT's errorful predictions for PRT were indicative of the model's failure to incorporate incidental fitness costs that appear in foraging tasks that influence rational foraging behavior.

Risk-sensitive foraging models better predict PRT, as they incorporate many of the potential fitness costs involved in traversing patches nested in ecological habitats. Predation, for example, accounts for much of the variability in PRT, as the best areas for foraging are also likely to be the most dangerous provided that predators and prey are attracted to the same sources of nourishment (Lima & Dill, 1990). Thus, in this context, optimal foragers must balance the benefit of optimal energy gain with the cost of increased predation risk when selecting patches from which to feed. As predicted by risk-sensitive accounts of foraging, various species of animals have been observed engaging in what MVT would categorize as sub-optimal foraging strategies in the presence of predators (Real & Caraco, 1986; Verdolin, 2006).

In a comprehensive review of risk-sensitive foraging theory, McNamara and Houston (1992) argue that no single model can sufficiently account for the numerous scenarios a foraging animal may encounter. Optimal PRT can fluctuate with respect to

available energy resources, quality of food, extrinsic time pressure (e.g., daylight), weather, predation, and the biological imperative of the food. Thus, models ought to be designed for specific foraging environments, such that the predictions generated from any given model result from those state-dependent parameters.

A model of overnight survival proposed by Stephens (1981) illustrates the argument presented by McNamara and Houston (1992). If overnight survival depends on an animal reaching a critical threshold of nutrients (x_c), the optimal time spent at the selected patch ($X(T)$) is a function of the reserves possessed by the animal (x_0) at the beginning of the foraging period (see Equation 10).

$$P(X(T) > x_c) = \Phi\left[\frac{x_0 + \mu_i T - x_c}{\sigma_i \sqrt{T}}\right] \qquad (10)$$

Assuming that all other parameters are kept constant, the quality of the patch and the time needed to reach the threshold will decrease as the reserves at the beginning of the foraging period increase. Thus, optimal patch selection and PRT are dependent on the state of the animal's reserves prior to foraging.

The important takeaway for decision theorists is that models of optimal behavior often omit factors important for, but not specific to, the focal task. When the study of animal foraging is limited in scope—such that the only parameters of importance are assumed to relate to the quality of a patch, the time spent traveling between patches, and the time spent within a patch—the intricate details of foraging dynamics are likely to be overlooked, thus, truncating the predictive strength of models derived from such assumptions. When models such as these are utilized to assess the fitness of strategies adopted in ecological habitats, complex patterns of behavior are unfairly labeled as maladaptive.

As mentioned previously, animal foraging theory is, by far, the most developed literature within the domain of cognitive search. The sophistication of the models accounting for patch selection and switching between patches has impacted a number of other scientific fields. While all of the models discussed here have the potential to be highly impactful once adopted into other fields of work, none has been as influential as the MVT. The magnitude of its influence is clear in light of its role in the development of models to capture the process by which people retrieve information from memory.

2.3.2 Memory search.

Searching for information in memory is yet another analogous task to hypothesis testing. Additionally, search in memory shares environmental properties with animal foraging, such as the presence of clustered resources. Memory, however, has proved to be a more challenging environment for examining search processes. Unlike environments that host animal foraging behavior (i.e., wide open spaces wherein movements of the specimen indicative of search behavior can be freely observed), memory must be studied indirectly. One useful tool to achieve this end has been paradigms used in the study of verbal fluency.

Verbal fluency is a long established psychological paradigm where subjects must generate a series of words that all follow the same rule (Newcombe, 1969). Typically, subjects are given 60 seconds to perform the task and are scored based on the number of terms that correctly follow the rule (e.g., all terms must be animals). However, in a series of experiments designed to investigate the underlying cognitive processes that drive performance on verbal fluency tasks, Troyer, Moscovitch, and Winocur (1997) found that clustering and cluster switching were also important components in navigating semantic

memory spaces. In this context, clustering refers to the act of listing terms that are highly associated with one another successively. For example, when tasked with listing animal names, participants are likely to cluster dog, cat, and hamster together given that they can be sub-categorized as household pets. Cluster switching refers to the act of terminating search in one sub-category (e.g., household pets) in favor of cuing recall from a new sub-category (e.g., farm animals). People who utilize the largest clusters and exhibit more cluster switching generate the most correct terms while completing a verbal fluency test (Troyer, Moscovitch, & Winocur, 1997), demonstrating the benefits of area restricted search.

Alternatively, some have argued that this pattern of results reflects the usefulness of semantic associates as retrieval cues (Levy & Baddeley, 1971). While this explanation may accurately describe the process by which performance is enhanced on this task, as such it may only serve as a proximate reason for why this strategy would be adaptive. The link to area restricted search appears to be a better fit, as knowledge—like food—appears to be hierarchically clustered such that large categories of information (patches) consist of many sub-categories (clusters). Thus, optimal memory queries would need to balance the use of both global (i.e., categorical) and local (i.e., semantic associate) cues for memory search; such dynamics have been implemented in a number of successful computational models of memory (Davelaar, 2015; Gronlund & Shiffrin, 1986; Metcalfe & Murdock, 1981; Raaijmakers & Shiffrin, 1981).

BEAGLE—a recently developed model of semantic memory search—builds off of the MVT to predict cluster switching during free recall from natural categories (Hills, Jones, & Todd, 2012). The optimal time spent in a cluster—as derived from their average

resource intake algorithm (Equation 11), which is a ratio between gain per time unit spent retrieving within a cluster and the sum of time spent traveling within and between clusters—is calculated as the product of average resource intake (*R*; this is referred to as net energy in MVT) and the cumulative gain within a patch (g^*; Equation 12).

$$R = \frac{g(t_w)}{t_w + t_b} \tag{11}$$

$$t^* = R * g^* \tag{12}$$

Ultimately, BEAGLE is MVT with a new nomenclature indicative of memory research. Tests of its predictive value have demonstrated that BEAGLE performs better than static models that do not make use of both local and global retrieval cues (Hills, Jones, & Todd, 2012). More importantly, BEAGLE accurately predicted that people cluster switch once they've nearly depleted a cluster. That is, in the same way that animals will move to a new patch once the likelihood of finding more food has diminished, people will transition to a new global cue once they have reported nearly all the words related to a subcategory.

The applicability of MVT to recall illustrates the generalized nature of search, and showcases the importance of cost-benefit tradeoffs for explaining when it becomes preferable to terminate search in one location (e.g., a patch, cluster of memories) and transition to search elsewhere. However, as I pointed out at the start of this section, the search environment is constrained to a space that is difficult to observe. People may very well encode items such that they are accessible via overlapping global cues, which could proffer explanations for specific global transitions. Thus, the simplicity of the model that explains foraging in semantic space may be superficial, and it would be difficult to explore how more contextualized models—such as those deployed in the animal foraging

literature—apply to internal processes. Alternatively, the costs associated with memory foraging aren't likely to be as complex as those encountered by animals while foraging. Mating opportunities and predation, for example, aren't likely to have analogous counterparts in recall, where time and opportunity are likely to be the most impactful costs associated with continued search. Moreover, the incentive structures for the two domains of search appear to be innately different given that the sought-after resource in animal foraging can take any number of forms (e.g., prey, fruit, nuts), each with its own utility (e.g., nutritional value). Memories are all of a kind despite varying with respect to their relevance to any given attempt at recollection. Thus, environmental pressures (e.g., task domain) are likely to drive differential utility of memories.

However, that which differentiates some instantiations of search becomes a source of commonality between others, as information is the resource sought in both recollection and hypothesis testing. One obvious, yet critical, difference is the location of the desired information: internal (memory) or external (information repositories). When information awaits exploitation in the environment, search resembles animal foraging due to the various forms information can take (e.g., books, webpages). This can be seen in the application of MVT-like models to data collected in the information search literature.

2.2.3 Information search. Hypothesis testing is, in essence, a special case of information search, making information search the most relevant body of literature within the domain of general cognitive search. The distinction between hypothesis testing and other forms of information search may be superficial in nature, as one can imagine acquired information of all kinds serving the purpose of examining any number of hypotheses (e.g., PCs are superior to Apple computers, Natalie Portman has voiced a recurring character on

The Simpsons, it will rain in Atlanta tomorrow afternoon). This issue, however, lies outside of the scope of this review, and I assumed that people seek information for reasons other than testing hypotheses.

College students, for example, regularly engage in information search for the purpose of acquiring knowledge that will assist them in learning material, performing well on exams, and writing term papers. This information is available in the form of books, websites, and academic journals, each of which is linked to costs and benefits that a student must weigh to optimize their study practices. Perhaps the most common and frequent human foraging endeavor this century, searching for information via the World Wide Web has been studied at length in terms of time costs, resource costs, and opportunity costs (Pirolli & Card, 1999).

Consistent with Hills' (2006) notion of area restricted search, information on the World Wide Web is organized in clusters such that important websites containing search-relevant information will often provide links to other relevant websites. Hyperlinks often appear as words, phrases, or sentences that some have argued can serve as proximate cues that emit "information scent" (i.e., hints), which cues the forager to the existence of distal information patches (Pirolli & Card, 1999). The ACT-IF (information foraging) model, developed by Pirolli and Card (1999), is a spread activation model (see Anderson, 1993) that predicts how people navigate these information spaces. Specifically, the ACT-IF model anticipates cluster selection behavior, where a cluster is a collection of words or links on a computer screen. The clusters preferred by information foragers are those that have the strongest scent, which is essentially the strength of the semantic association of

words contained in the information display (e.g., website) and the probe or desired information.

Potential information gains resulting from any cluster of words (g) are a function of the expected ratio of information activation (A_i) to the time costs of information search (T; Equation 13). Thus, a cluster's appeal will vary depending on how closely related the cluster is to the desired information and the amount of time it would take to filter through that cluster.

$$g(c,s) = exp\left(\frac{\sum_{i \in Q} A_i}{T}\right) \quad (13)$$

Optimal information foraging, therefore, involves strategic allocation of attentional resources, such that clusters with the greatest potential ought to be selected most often. More importantly, ACT-IF can predict optimal search time with respect to the rate of gain in a manner reminiscent of MVT's capacity to predict optimal PRT. ACT-IF calculates the rate of gain as a function of the ratio of information gain for a specific cluster to the time spent finding and filtering through that cluster (Equation 14).

$$R_D = \frac{\sum_{i=1}^{k} g(i,s)}{t_B + t_w} \quad (14)$$

Empirical tests of this model have demonstrated that ACT-IF is capable of accurately predicting cluster selection, as well as the amount of time spent foraging within information clusters (Pirolli & Card, 1999). These results suggest that human external search mechanisms are sensitive to costs and operate in such a way that mirrors optimal foraging strategies adopted by animals.

Pirolli (2005; see also Wu & Pirolli, 2007) allows ACT-IF to flexibly adapt to different searching environments (e.g., variations in web design, difficulty of accessing

information), understanding that optimal search strategies are likely to change as the human-computer interface changes. This is an important trait for a model of this nature to possess given that information search on the Web is accurately described as an ill-structured problem that can take on many forms, and it is the first example of an information foraging model that addresses ecological complexities in a manner somewhat similar to the animal foraging models discussed previously. This flexibility, however, does not address the nature of information in environments with great uncertainty, as is often the case in hypothesis testing. Hypothesis testing more closely resembles diet selection (Stephens & Krebs, 1986), where proximate cues are unavailable for assessing what is likely to be gleaned from a test (i.e., test results may be related to more than one hypothesis). Moreover, hypothesis testing (or hypothesis evaluation, more broadly) can be conceptualized as a well-defined problem, as the goals for such tasks are relatively concrete (e.g., accurate diagnosis).

The three bodies of literature discussed in this section (animal foraging, memory search, information search) share a number of features that make it possible for numerous models founded on a singular idea (e.g., maximizing gains per unit cost) to account for a wide array of behaviors. Moreover, the evidence in support of a generalized cognitive search mechanism lends credence to the notion that hypothesis testing (an information seeking behavior) ought to be conceptualized as a special case of search. As such, hypothesis testing theory should include contextual mechanisms such as those incorporated into models of animal foraging, and conceptualized the perceived utility of information repositories as a function of their association with a probe or the purpose for search (i.e., a hypothesis within this context) as has been done in models of information foraging.

Another necessary mechanism that has yet to be implemented in hypothesis testing research are rules for termination testing behavior. That is, we do not yet have a theoretical account for when is it no longer worth the effort (or costs, more generally) to continue acquiring information and, instead, act on currently held beliefs. Generally speaking, very few accounts of termination rules have been published in all search-related literature. The sparse nature of our understanding of search termination is a critical limitation of the field and is especially important in information search as it represents the moment when a decision-maker is satisfied with their understanding of the environment and ready to act on that understanding.

2.4 Hypothesis Generation

Revealing the cognitive mechanisms underlying decision-making has been an important advancement in cognitive decision theory, steering the field away from the heuristics and biases perspective of human judgment (Kahneman & Tversky, 1996) and toward a process-oriented field of study. The ability to reduce human judgment to foundational cognitive processes affords unifying theory of higher order cognitive processes, as has been exemplified by the Hypothesis Generation (HyGene) cognitive architecture (Dougherty, Thomas, & Lange, 2010; Thomas, Dougherty, & Buttaccio, 2014; Thomas, Dougherty, Sprenger, & Harbison, 2008).

The HyGene cognitive architecture was originally developed for the purpose of integrating long-term memory and working memory systems for the purpose of explaining variation in probability judgments (Thomas, Dougherty, Springer, & Harbison, 2008). In its present format, HyGene accounts for the generation, maintenance, and testing of hypotheses by implementing 3 memory modules: working memory, episodic memory, and

semantic memory (Dougherty, Thomas, & Lange, 2010). HyGene treats the hypothesis generation process as a general case of cued recall, such that observed data cue the activation of a subset of episodic memories highly associated with the data. Hypotheses are generated from semantic memory when the conditional intensity of activated memories exceeds a threshold, and are maintained in a Set of Contenders (SOC) limited in size by both cognitive (e.g., individual differences in working memory) and task (e.g., time pressure) constraints. Hypotheses maintained in the SOC are available to be used as input for additional tasks (e.g., probability judgment, hypothesis testing).

Global matching models like HyGene have been used to account for a number of memory and probability judgment phenomena (Dougherty, Gettys, & Ogden, 1999; Hintzman, 1984; Hintzman, 1988). However, HyGene accounts for four unique findings in the probability judgment literature: subadditivity, strength of alternatives effects, working memory capacity effects, and time pressure effects. HyGene computes conditional probabilities by invoking support theory (Equation 15; Tversky & Koehler, 1994), but transforms it into a process-driven probability estimation by replacing *support* with memory activation (i.e., support for a hypothesis in a subset of activated memory; Equation 16). This is essentially a ratio of the intensity of active memories in support of the focal hypothesis to the total intensity of active memories.

$$P(A,B) = \frac{s(A)}{s(A) + s(B)} \tag{15}$$

$$P(H_i|D_{obs}) = \frac{I_{C_i}}{\sum_{i=1}^{w} I_{C_i}} \tag{16}$$

According to HyGene simulations, the probability judgment effects listed above emerge as a consequence of the number of hypotheses generated in response to the

observed data (Thomas et al., 2008). The magnitude of subadditivity (i.e., the degree to which the sum of objective probabilities is less than the sum of judged probabilities), for example, was thought to be related to working memory capacity (Dougherty & Hunter, 2003b), such that those with high working memory capacity exhibited less subadditivity than those with low working memory capacity. HyGene simulations demonstrate that when many hypotheses (e.g., 4) are contained in the SOC (as has been observed for high-span individuals), subadditivity is low because more hypotheses are present to account for their share of the activated memories in the episodic store. Probability estimates increase substantially when few hypotheses are contained in the SOC (e.g., 2), resulting in high subadditivity. A similar pattern of behavior emerges from the model when time allowed to generate hypotheses is restricted (i.e., simulated time pressure), as fewer hypotheses are generated, probability judgments increase, and subadditivity increases.

The importance of HyGene is its explanatory power regarding higher-order cognitive phenomena, such as the probability judgments reviewed above. Since its initial application to probability judgments, HyGene has been extended to both visual search and hypothesis testing. A number of studies have shown that the contents of working memory guide saccadic movement in visual search tasks (Beck, Hollingworth, & Luck, 2012; Olivers, Peters, Houtkamp, & Roelfsema, 2011; Soto & Humphreys, 2007), which suggests HyGene is well suited for modeling visual search given its theoretical account for how hypotheses are generated into working memory. Buttaccio, Lange, Thomas, and Dougherty (2015) examined first fixations in a visual search task to examine the potential influence of hypotheses generation on visual attention. A HyGene model that assumed the first hypothesis to be generated influenced visual search alone was found to fit the data

best, suggesting that generated hypotheses can account for the allocation of overt visual attention—a finding consistent with prior research indicating that overt visual attention is typically guided by a single item contained in working memory (Olivers et al., 2011).

In an extension of HyGene intended to capture hypothesis testing behavior (HyGene-HT), Dougherty, Thomas, and Lange (2010) implemented a number of memory-related heuristics for evaluating information depositories and guiding information search. These included the memory-strength heuristic (i.e., select the cue associated with the most highly activated hypothesis), dissimilarity heuristic (i.e., select the cue that maximizes the dissimilarity between the focal hypothesis and the strongest competitor), memory strength difference heuristic (i.e., select the cue that maximizes the difference in memory strength between the two leading hypotheses), and Bayesian diagnosticity (i.e., select the cue with the highest likelihood ratio). This early attempt to model information search using basic memory phenomena proffered a theoretical foundation for bias towards positive-test selection (a commonly observed outcome in human reasoning discussed earlier in this review; Wason, 1968), suggesting that tendencies to a single hypothesis accounts for the observed pattern of positive-test bias. That is, there is no incentive to value information depositories that could reveal falsifying data with respect to the hypothesis under consideration if a person has failed to generate a competing hypothesis.

Thomas, Lange, and Dougherty (as cited in Lange et al., in press) found evidence in support of HyGene-HT's predictions, as people sought diagnostic tests when observed data was suggestive of two hypotheses and preferred associative tests when observed data was suggestive of one hypothesis. Thus, self-generated hypotheses were observed to elicit

preferences for tests under different ecological conditions, suggesting that hypotheses influence the perceived value of tests.

The generalizability of HyGene-HT, however, is severely limited given that some of the heuristics outlined above are too simplistic to be tenable in complex decision-making environments. Additionally, the heuristics only accommodate the value of the available tests, thus, ignoring the costs incurred by exploiting each test. Moreover, HyGene-HT shares a number of features with the adaptive toolbox model for stopping rules reported previously (Ficic & Buckman, 2013), as it assumes that many information heuristics are available to decision-makers and individual differences or some set of contextual properties dictate which of these heuristics are used by a decision-maker. Most notably, HyGene-HT lacks a formal stopping rule and, thus, cannot account for termination decisions.

Although hypothesis generation models appear to provide a sound foundation for hypothesis testing theory, current models are ill-equipped to address hypothesis testing as defined in this review. Visual search models appear to be unique with respect to hypothesis generation, as visual constraints limit the number of hypotheses that could potentially influence search behavior independent of the additional constraints represented in the HyGene architecture (Buttaccio et al., 2015). Despite incorporating memory processes, HyGene-HT exhibits a narrow perspective of hypothesis testing indicative of much of the earliest work in the field. The model lacks much of the infrastructure necessary to investigate general search mechanisms, such as a way to estimate the perceived cost of exploiting a test.

2.5 Hypothesis-Guided Search

The computational modeling carried out to simulate the empirical studies reported in this dissertation implemented components of the conceptual model I described below, which aims to advance hypothesis testing theory by incorporating both environmental factors and cognitive mechanisms into the evaluation of information sources. That is, I conceptualized test selection as a consequence of a decision-maker's experience and the limitations of their cognitive abilities, as well as contextual limitations like the availability of data and the incentives for performing a task accurately. I accounted for costs associated with acquiring data in a manner that was consistent with generalized theories of search. The perceived utility of information depositories was subjective, and costs were interpreted relative to the contexts in which they arose. Most importantly, I linked the perceived value of tests to the hypotheses entertained by decision-makers.

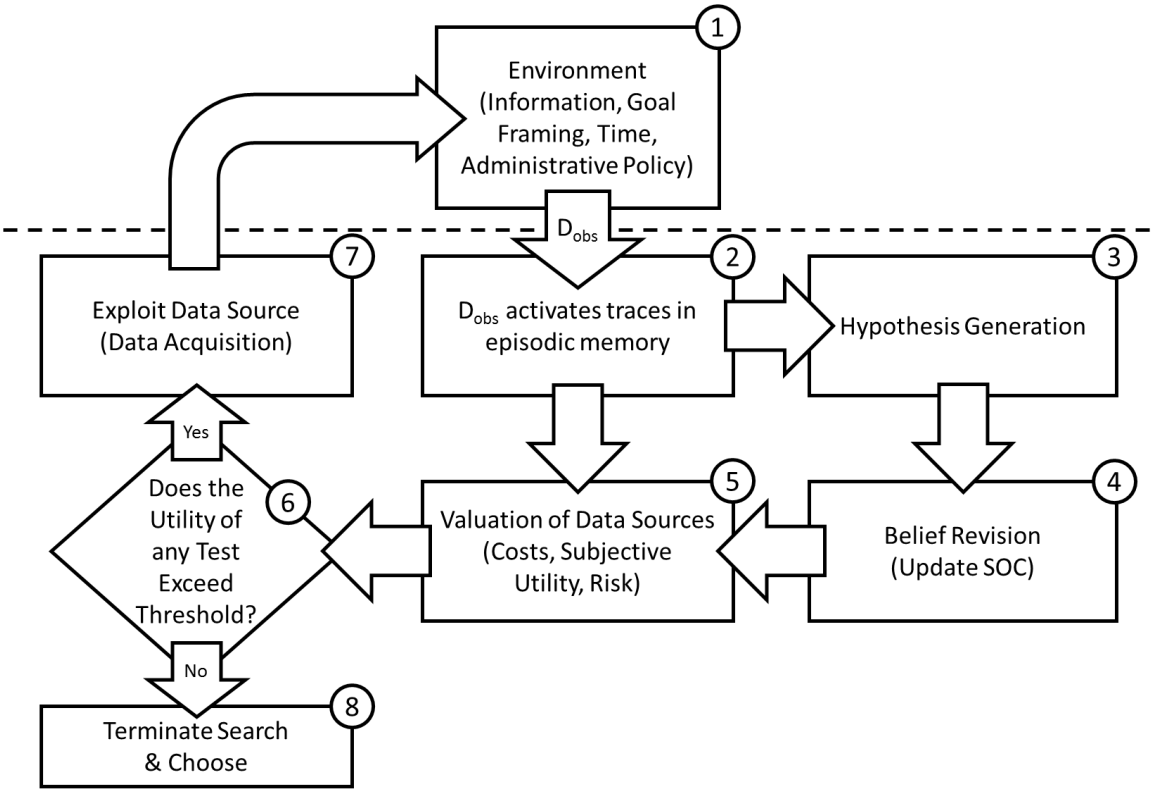


Figure 1. The hypothesis-driven valuation model of hypothesis testing. Boxes above the dashed line represent factors that exist outside of the cognitive system, while boxes below the dashed line occur within the mind.

Figure 1 illustrates the hypothesized cognitive account of hypothesis testing behavior. The dashed line in Figure 1 separates information that originates outside of the cognitive system (above the line) and the processes theorized to occur within the mind (below the line). Steps 1 through 3 are consistent with previous conceptualizations of hypothesis generation proposed by Thomas, Dougherty and their colleagues (Dougherty, Thomas, & Lange, 2010; Thomas, Dougherty, & Buttaccio, 2014; Thomas, Dougherty, Sprenger & Harbison, 2008). In essence, hypothesis generation is a special case of cued recall, where data present in the environment (step 1) serves as cues to engage in retrieval from memory. I expand on Thomas et al.'s interpretation of a decision environment to include information that restricts or frames the nature of a decision. The environment contains the initial cues for diagnosis, contextual limitations (e.g., time constraints), and incentive structures (e.g., goal framing). This enters the cognitive system as observed data (D_{obs}), internalized by the decision-maker.

In step 2, the observed data activate traces in episodic memory with which they share common features. Hypothesis generation (step 3) includes a number of events described in greater detail elsewhere (Thomas et al., 2008). Information is aggregated from the activations in step 2, forming a probe to match against semantic representations of known hypotheses. This process is an iterative one. The probe is continuously matched against the contents of semantic memory until the decision-maker reaches a maximum number of failed attempts to sufficiently match a hypothesis (Harbison, Dougherty, Davelaar, & Fayyad, 2009).

As in Thomas et al.'s (2008) model, generated hypotheses populate the set of contending hypotheses (SOC; step 4). The SOC is, essentially, a representation of working memory and, as such, is limited in its capacity to actively maintain information (hypotheses in this case). Thus, hypotheses compete for the limited space available within the SOC, such that only those with the strongest activation or the most support remain in the SOC. These remaining hypotheses play an important role in judging the value of available tests (Step 5) by modifying valuation judgments with respect to posterior belief distribution (Gettys & Fisher, 1979). That is, tests associated with the most activated hypotheses maintained in the SOC will be judged as more valuable than those associated with hypotheses with weaker activations.

Valuation judgments are also assumed to be sensitive to information available in the environment (arrow from box 2 to box 5 in Figure 1). For instance, the domain may require decision-makers to adhere to specific protocols for selecting tests. These environmental pressures provide additional information for valuating available tests. While the tests required by protocol and those suggested by the contents of the SOC may overlap, this model provides an explanation for violations of administrative policy without invoking insubordination or a failure to recall the protocols. That is, decision-makers may circumvent policy when the hypotheses within the SOC drive up the valuation of tests not listed in mandated protocols.

Decisions regarding the exploitation of sources will emerge from cost-benefit assessments of continued search. In step six, people consider the valuation judgment of a test and costs associated with the test relative to a threshold for carrying out tests to formulate exploitation decisions. Thresholds, perceived costs, and valuation judgments are

likely to vary by context and by individual, and the nature of thresholds could potentially take many different forms (e.g., improve potential gains by some amount, reach a specified level of confidence, reduce uncertainty by some amount). Thus, this model is capable of anticipating the conditions under which decision-makers may select tests dominated by alternatives with respect to value.

It is worth noting that this model accounts for associative search (pseudo-diagnostic testing) in a manner consistent with prior versions of HyGene, as decision-makers will perceive such tests as valuable only when a single hypothesis has been generated (Lange, Thomas, & Dougherty, 2010). Consequently, decision-makers begin to appear as though they have adopted diagnostic strategies as a result of considering multiple generated hypotheses. This model, however, expands upon HyGene-HT by accounting for observed differences in testing strategies without changing the mechanism by which tests are evaluated.

When threshold is exceeded, this model posits that the most appealing information depository will be exploited (Step 7). The decision to exploit a source of information will result in sampling more data from the environment (Step 1), which will initialize another iteration of belief revision. This process illustrates how the HyGene architecture can account for nuanced test preference. Additionally, it formalizes a mechanism by which the information sought out by decision-makers results in downstream effects of information preference (Smith, Huber, & Vul, 2013), as the information realized by a testing event is likely to result in changes in the perceived value of remaining information depositories.

A critically important contribution of this model is the inclusion of a termination rule, which is assumed to be emergent from the cost-benefit assessments of continued

search (Step 6). Simply put, testing is terminated when the gains to be made from exploiting the most useful test available fail to exceed the threshold (Step 8). The process itself, however, can be exceedingly complex because what it means for a test to be unwarranted varies as a consequence of the environment, individual differences, internal processing, or any combination thereof. For instance, the threshold will differ on an individual basis, where conservative testers will require that the perceived utility of future states be far more advantageous than their current state (i.e., high threshold) and more liberal testers will run tests that provide additional information of any value (i.e., low threshold). Conservative testers will be more likely to terminate quickly after selecting few tests, while liberal testers will terminate after many tests have been exploited. It may be the case that thresholds are adaptive in nature, responding either to changes in the environment or the decisiveness of the decision-maker (Kruglanski & Webster, 1996).

As stated previously, just about any factor present in the model can have an impact on decisions to terminate. The generation process itself can exhibit a lot of influence over termination decisions, as can the posterior distribution of beliefs over generated hypotheses. Termination is likely to occur early when posterior beliefs are asymmetrical (i.e., there are few strong candidate hypotheses) as is the case when a single hypothesis is generated, and late when beliefs are symmetrical (hypotheses under consideration have similar probabilities) as is more likely when many hypotheses are generated. Imagine a patient arriving at a hospital, presenting with lower-right abdomen pain, high fever, and vomiting indicative of an upset digestive system. The physician's belief in an appendicitis diagnosis may be so strong that no test is valued sufficiently to warrant running any test.

The structure placed on the task by the environment can impact termination in a number of ways. Costs, for example, may vary with respect to environmental constraints, such as time pressure. Should the decision-maker perceive that their time is limited for any reason (e.g., high workload, patient in critical condition), the time needed to run tests would ultimately be too costly to warrant testing. Goal framing, or changes in the incentive structure, can impact the manner in which thresholds are placed or modify the valuation process. Placing an emphasis on accurate diagnoses, for example, can either increase the perceived value of tests or reduce the threshold. Either of these changes will result in an increase in the number of tests run, affording the decision-maker more information upon which to base their decisions. Alternatively, placing an arbitrary cut-off on the resources available to the decision-maker (e.g., cap on tests run, cap on monetary expense) is likely to result in earlier termination either because few tests warrant use of those resources or decision-makers will conserve their resources until they are most needed.

The conceptual model introduced above can account for hypothesis testing phenomena (e.g., pseudo-diagnostic search, diagnostic search, early exit) without invoking more than a cued memory process that drives resource valuation judgments, and a termination rule. Not only is this a more parsimonious approach to building hypothesis testing theory in comparison to much of the work that has been done to date, but it is the first theory of hypothesis testing that accounts for the decision-maker's beliefs for the purposing of explaining test preference.

The most shocking limitation of current hypothesis testing theory is the inability for any contemporary model to account for the manner in which hypotheses under the consideration of decision-makers govern observable hypothesis testing behavior. As

defined by Poletiek's (2001) account of hypothesis evaluation, testing cannot occur until hypotheses have been generated by decision-makers. Thus, hypotheses under consideration are, in essence, the motivating factor in engaging the environment for information to clarify the decision-maker's understanding of their surrounding environment.

CHAPTER 3. EMPIRICAL STUDIES

3.1 Medical Diagnosis Game

The array of empirical studies reported in this section utilize a single experimental paradigm, adapted to address the various research questions posed previously: The Medical Diagnosis Game (MDG; Illingworth & Thomas, 2015). The MDG is a forced-choice, simulated diagnosis task modeled after experience-based category learning paradigms (e.g., Hoff & Rehder, 2010; Posner & Keele, 1969). This paradigm has been validated for investigating sequential data acquisition (e.g., hypothesis testing), diagnosis, and tradeoffs inherent to foraging tasks (Illingworth & Thomas, 2015; 2016; 2017).

The MDG typically consists of two phases. I designed the first phase of the game to facilitate internalization of the probabilistic relation between disease states and test outcomes, as participants complete numerous blocks of trials diagnosing patients with different disease-test outcome configurations. The number of diseases, tests, and test outcomes vary by experiment. To illustrate the task, assume that there are three diseases, four tests, and three possible outcomes per test. In a learning trial, outcomes from the four medical tests associated with a fictitious patient ailed by one of three mutually exclusive diseases are presented to participants. Test results take the form of circular, black and white images of familiar medical tests (e.g., computed tomography scans (CAT), chest cavity x-rays (X-RAY)). The pattern of outcomes (images) obtained from medical tests is controlled by the statistical relation between diseases and medical tests. Test labels (e.g., CAT, X-RAY) and corresponding images are randomly assigned to each test in the experiment for each participant.

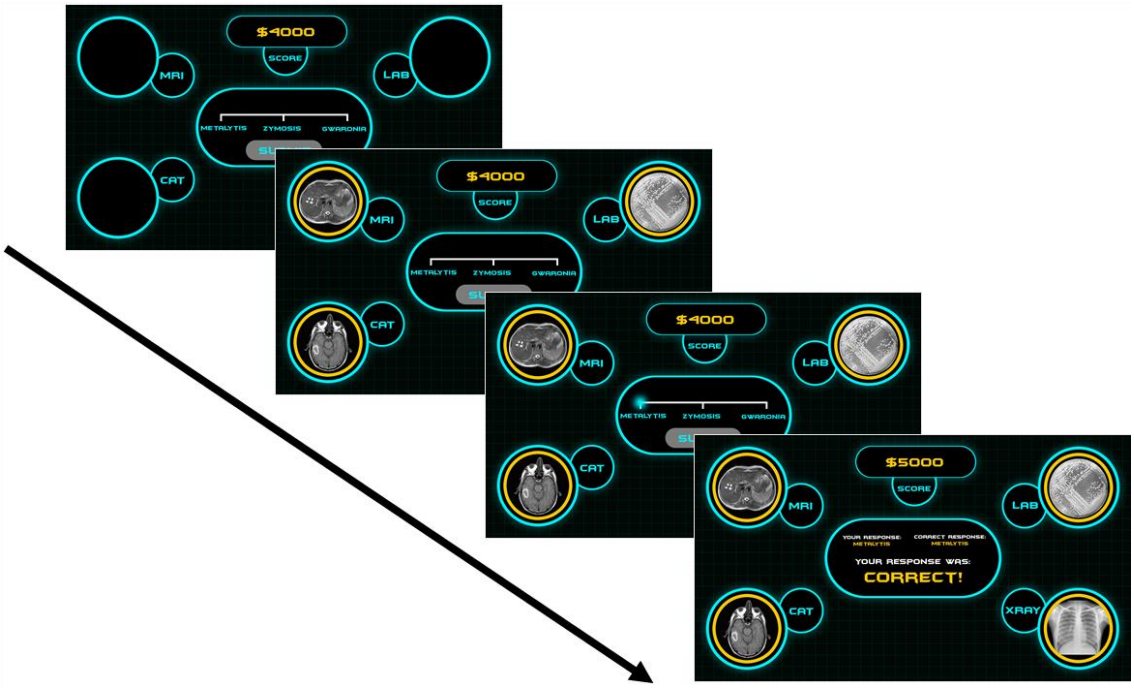


Figure 2. Illustration of a learning trial for the MDG experimental paradigm. All test results appear after 1500ms during learning trials. Test trials differ such that results only appear after participants click on the desired test. Diagnoses are issued at the participants' discretion, after which feedback is provided.

Figure 2 illustrates a learning phase trial. The experimental interface consists of four circular widgets located in the four corners of the computer screen where the results of medical tests appear, as well as a scale—centered in the display—to submit diagnoses. Medical tests are randomly assigned to one of the four widget locations, and remained in the same location throughout the experiment, for each participant. Participants submit their diagnoses at their own discretion by selecting one of three fictitious diseases—Metalytis, Zymosis, or Gwaronia. Diagnoses can be changed indefinitely until participants submit their response. Feedback is provided after each diagnosis such that the word "CORRECT!" appears after participants diagnose the patient with the appropriate disease and "INCORRECT" appears after an erroneous diagnosis.

The gamification of the task is implemented in the feedback following completion of the diagnostic task, as all participants earn 1000 points ($) for each correct response during the learning phase. The points participants earn accumulate in their "bank" of resources that they later use to complete the second phase of the experiment.

The second phase of the experiment houses all environmental manipulations other than the statistical structure of test outcomes. The most important feature of phase 2 is the occlusion of test results until participants explicitly request to view the outcomes of specific tests. This affords measurement of test preference, as well as an opportunity to measure relative valuation and tradeoffs via the manipulation of information access costs. Such manipulations can delay the presentation latency of those outcomes (i.e., a time cost) or impose a monetary expense (i.e., sacrificing previously earned points to view outcomes).

Outcome presentation can also be constrained such that test outcomes can be controlled after specific test selections either with respect to the sequence of selection (e.g., first test outcome is always positive) or the specific test requested (e.g., CAT outcome is always positive). Such manipulations afford control over objective posterior belief distributions following each datum acquired, which may be important in decisions to terminate testing. Finally, the incentive structure of the task can be manipulated via the payoff mechanism for diagnostic accuracy. Participants can continue to earn 1000-point payoffs for correct diagnoses, or those points can be deducted from banks after incorrect diagnoses to frame the task within a loss context. Additionally, payoffs can be yoked to confidence judgments, affording examination of participants' calibration to their own performance on the task.

3.2 Experiment 1 - Hypothesis Generation and Test Preference

Prior work linking test selection to hypothesis generation has focused on single test selection in binary choice tasks (Thomas, Dougherty, & Lange, 2010). Most testing behavior, however, occurs in environments wherein multiple tests are exploited and data is accumulated over time. To date, no study has systematically manipulated the hypothesis set cued by information available at the beginning of the hypothesis evaluation process for the purpose of investigating sequential testing behavior. The purpose of Experiment 1 was to investigate test preference as it relates to cued hypothesis sets and changes to beliefs over time.

Information utility models often invoke hypotheses when computing test value (Manktelow & Over, 1990; Over & Evans, 1994; 1996; Oaksford & Chater, 1996); however, such metrics are not suited for sequential data acquisition to evaluate more than a single hypothesis. Probability gain, for example, conceptualizes the utility of a test as its capacity to make a hypothesis more probable. Not only does this definition of utility lose clarity in decision environments with greater uncertainty (e.g., more than one hypothesis is considered), it presupposes that the goal of testing is to confirm a specific hypothesis or that decision-makers are aware of the correct state of the world ahead of testing.

Alternatively, the HyGene architecture assumes that preference for information manifests as a byproduct of belief. The need to seek diagnostic information manifests when more than one hypothesis is considered as a possible account for a dataset (Thomas, Dougherty & Lange, 2010). Otherwise, pseudo-diagnostic or positive search is an adequate strategy for informing a decision-maker. HyGene, however, makes more nuanced predictions regarding information preference as it relates to the set of contending hypotheses. Provided that more than one hypothesis is considered by a decision-maker,

preference should reflect the contents of working memory such that information depositories with a history of differentiating between contending states of the world are exploited at a higher rate.

The purpose of Experiment 1 was to test this prediction by measuring test preference in response to cues intended to bias participants towards maintaining differential belief distributions. The statistical structure of the experimental environment detailed below was designed to exact control over the hypotheses most strongly considered by participants. The general hypothesis was that presenting cue would predict test preference where patterns of testing behavior would appear markedly different depending on which presenting cue was presented.

3.2.1 Method.

Undergraduate students enrolled at the Georgia Institute of Technology were recruited to participate in this study via an online experiment management system (SONA Systems). In total, 31 participants completed the experiment. All participants received partial course credit for their involvement in the study.

Hypothesis-guided testing behavior assumes that the contents of one's beliefs (i.e., the hypotheses considered by a decision-maker) drive test preference. Thus, the tests selected by participants when decision environments cue differential hypothesis sets should vary considerably. This manipulation was implemented by presenting participants with a presenting sign that patients exhibited prior to the selection of any tests. The ecology that defined the relation between disease states and presenting cues is outlined in Table 2. The values presented in Table 2 represent the probability that one of the four possible diseases accounted for the emergence of each presenting symptom. Not only did the presenting

symptoms cue different sets of hypotheses, they also differed with respect to the number of hypotheses a decision-maker with perfect knowledge of this ecology considered. Manipulating the presenting cue in this way afforded an investigation of pseudo-diagnostic search (as in Thomas, Dougherty, & Lange, 2010) in addition to the hypothesis-guided testing behavior predicted by the HyGene architecture. That is, pseudo-diagnostic search was hypothesized to occur when there is a single strong hypothesis considered, while diagnostic search was hypothesized when multiple hypotheses were candidate explanations.

Table 2. Presenting sign ecology for Experiment 1.

	Presenting Symptoms			
	Cue 1	Cue 2	Cue 3	Cue 4
Hyp 1	.65	.32	.14	.07
Hyp 2	.12	.14	.32	.31
Hyp 3	.12	.14	.32	.31
Hyp 4	.12	.32	.14	.31

Table 3 lists the probability that each of four possible diseases (Hyp 1-4) accounts for a patient's ailment conditional on the patient presenting with one of four symptoms (Test 1-4). The tests available to decision-makers are designed to map closely to the cue configuration presented in Table 2. Note that Cue 1 was strongly associated only with Hypothesis 1 (Hyp 1 manifests in 65% of cases presenting Cue 1). The only hypothesis for which Test 1 exhibits an informative outcome is Hypothesis 1 (see top left of Table 3). Similarly, Cue 2 has strong associations with Hypotheses 1 and 4, while outcomes for Test 2 are most useful for disambiguating these same two hypotheses. The critical property of the tests listed in Table 3 was the even distribution of diagnosticity across the four tests. That is, when a decision-maker believes that all four hypotheses are equally likely, the diagnostic values of all tests are approximately equal (diagnosticity ≈ 2).

Table 3. Test outcome ecology for Experiment 1.

		Diagnostic Tests			
		Test1	Test2	Test3	Test4
Hypothesis 1	Outcome 1	.80	.65	.30	.30
	Outcome 2	.10	.20	.40	.40
	Outcome 3	.10	.15	.30	.30
Hypothesis 2	Outcome 1	.30	.30	.65	.60
	Outcome 2	.40	.40	.20	.20
	Outcome 3	.30	.30	.15	.20
Hypothesis 3	Outcome 1	.30	.30	.15	.20
	Outcome 2	.40	.40	.20	.60
	Outcome 3	.30	.30	.65	.20
Hypothesis 4	Outcome 1	.30	.15	.30	.20
	Outcome 2	.40	.20	.40	.20
	Outcome 3	.30	.65	.30	.60
Diagnosticity		2.09	2.09	2.09	2.03

However, once data was observed by the decision-maker (e.g., a presenting symptom), their belief distribution should change as would the diagnostic value of the tests. For instance, assuming Cue 1 is observed at the start of a trial, the diagnostic value of Tests 1 and 2 become higher relative to Tests 3 and 4. If we assume that decision-makers will select the most diagnostic test, Test 1 is most likely to be selected after Cue 1 is observed. The remaining cues and tests possess the same relation.

The learning phase of Experiment 1 was completed over 24 blocks of 20 trials (each disease hypothesis was equally represented in those 20 trials), resulting in 480 learning trials. Participants were instructed to use Phase 1 trials to internalize the statistical ecology that defined the relation between disease states and information available in the world. Presenting symptoms were one of four common medical conditions: fever, rash, migraine, and ache. These labels were randomly assigned to the cues detailed in Table 2. Data from 4 medical tests derived from a fictitious patient ailed by one of four mutually exclusive diseases were presented to participants during each trial. Stimuli were circular, black and

white images of computed tomography scans (CAT), chest cavity x-rays (X-RAY), bacterial cultures (LAB), and abdominal magnetic resonance images (MRI). Test labels and corresponding images were randomly assigned to each test appearing in Table 3 for each participant. After 1500ms elapsed following the onset of a learning trial, the outcomes of all tests were presented simultaneously. Participants submitted their diagnoses at their own discretion by selecting one of three fictitious diseases—Metalytis, Zymosis, Gwaronia, or Descolada. Diagnoses could be changed indefinitely until the participant submitted their response. Feedback was provided after each diagnosis such that the word "CORRECT!" appeared after participants diagnosed the patient with the appropriate disease and "INCORRECT" appeared after an erroneous diagnosis. Accuracy was incentivized during the learning phase of Experiment 1, such that participants were awarded points ($1000) that were deposited into a bank that grew with each subsequent correct diagnosis. Participants were instructed to make accumulating as many points as possible the primary goal of their task.

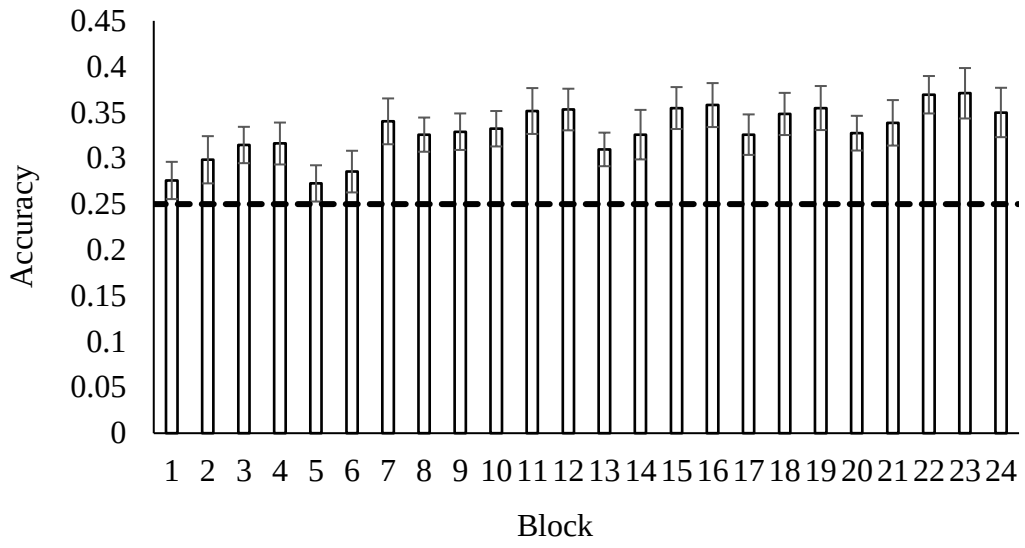
The test phase of Experiment 1 was completed over 4 blocks of 20 trials each, for a total of 80 test trials. The disease hypotheses were equally represented in each block of test trials. Test phase trials differed from those of the learning phase such that test results did not automatically appear after a short latency. Rather, participants had to explicitly click on the test widget before the corresponding outcome would appear. Tests were, thus, selected sequentially. The number of tests viewed was left to the participant's discretion, where termination of search (i.e., submission of their diagnosis) could occur after viewing between none and all of the test outcomes. Completion of each test trial was self-paced and, as was the case in learning trials, diagnoses could be changed indefinitely until the

participant submitted their response. Feedback was withheld from participants during the test phase of the experiment.

3.2.2 Results.

Learning. A logistic regression evaluated how well accuracy was predicted by block. Block was not found to be predictive of learning phase accuracy (χ^2 (n=31,23) = 23.77, p = 0.42), suggesting that performance did not improve over the course of the 24 blocks. However, participants were 1.41 times more likely to submit a correct diagnosis in the final block as they were in the first (β = -0.01, SE = 0.13, p <.01). Moreover, there was a significant linear trend relating accuracy to block (χ^2 (n=31,23) = 8.70, p = 0.003) These mixed results imply that knowledge of the task environment was highly variable in the sample, and may serve as an important individual difference when considering performance across the varied metrics of the task.

A



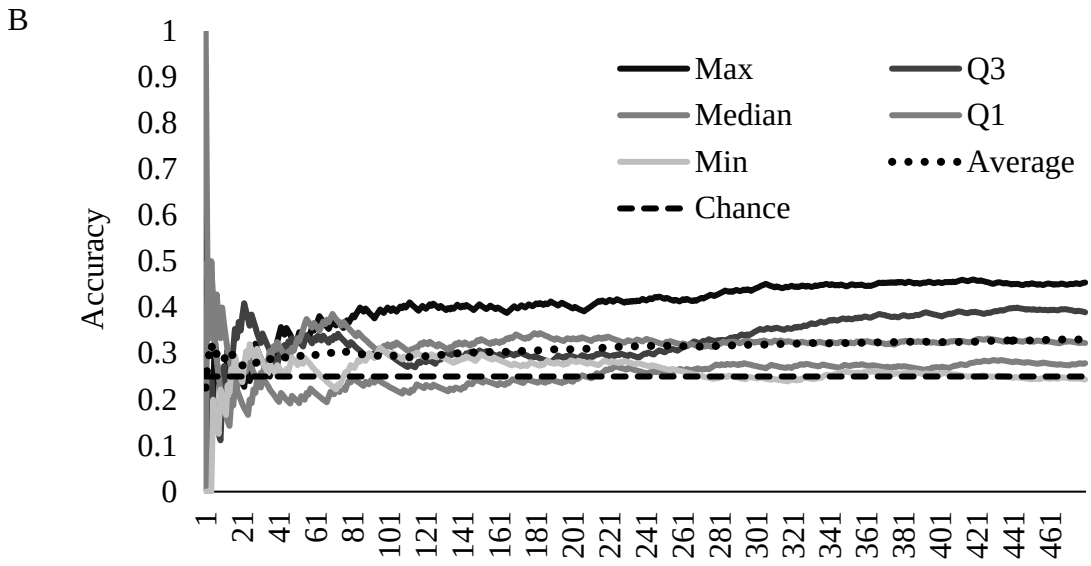


Figure 3. Experiment 1 learning. Panel A illustrates Experiment 1 learning phase accuracy broken out by block. The dotted line represents chance performance (25%). Error bars represent standard error. Panel B tracks proportion correct for 5 participants across all trials of Phase 1, as well as sample average (black, dotted line) and chance performance (black, dashed line). The worst performer exhibited chance accuracy, while the best performer approximately doubled the rate of correct responses.

Accuracy. A logistic regression evaluated how well accuracy was predicted by block. Block was not found to be predictive of learning phase accuracy (χ^2 (n=31,1) = 0.94, p = 0.70), suggesting that performance did not improve over the course of the test phase. Participants were nearly equally likely (O = 0.96) to submit a correct diagnosis in the final block as they were in the first (β = -0.04, SE = 0.1 , p =.70). Moreover, accuracy in the first block (M = 0.33, SE = 0.02) and the second block (M =0.35, SE = 0.02) were nearly identical to learning phase performance.

Stopping. A multinomial regression tested for differences in the total number of tests selected by presenting cue. Presenting cue did not predict the total tests selected per trial (χ^2 (n=31,3) = 3.04, p = 0.38). As can be seen in Figure 4, the number of tests selected appear evenly distributed across all presenting cue conditions. A follow-up analysis

explored the possibility that differences in learning phase performance could shed light on what if any sensitivity termination decisions exhibited in response to presenting cue. A second multinomial regression showed that neither presenting cue (χ^2 (n=31,3) = 2.61, p = 0.45) nor learning (χ^2 (n=31,1) = 1.22, p = 0.27) predicted the number of tests selected. More importantly, the interaction between presenting cue and learning did not predict total tests selected (χ^2 (n=31,3) = 2.92, p = 0.40). These analyses fail to provide any evidence that either the presenting symptom or the participants' knowledge of the task environment influenced decisions to terminate test selection.

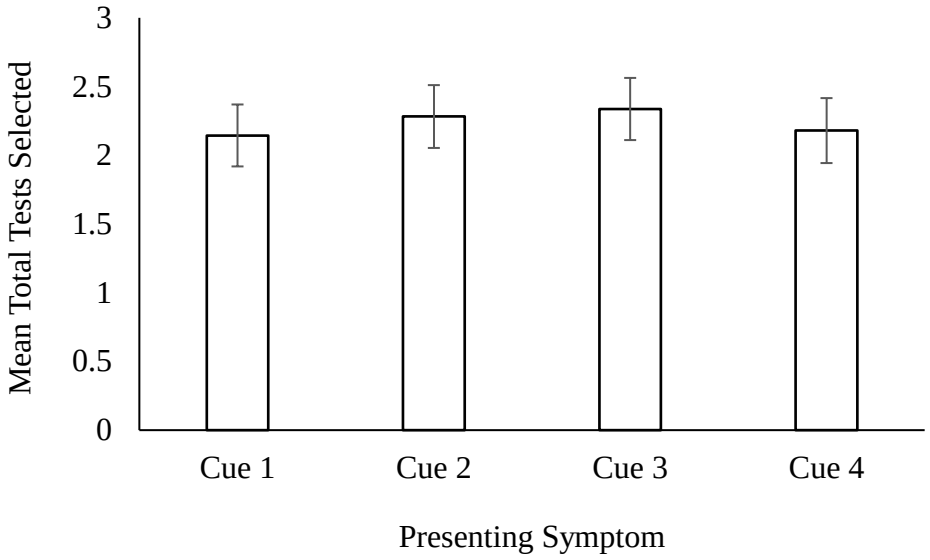


Figure 4. Experiment 1 total testing. The figure illustrates mean total tests selected by presenting symptom. The total number of tests selected did not vary by presenting symptom. Error bars illustrate standard errors.

Test selection. The frequency with which each available test was selected was evaluated by running a series of binomial logistic regression analyses. Presenting cue was not found to be predictive of Test 1 selection (χ^2 (n=31,3) = 0.83, p = 0.84). Learning phase performance was added to the model to evaluate whether or not knowledge of the task environment could elucidate Test 1 selection. Neither presenting cue (χ^2 (n=31,3) = 1.49,

$p = 0.68$) nor learning (χ^2 (n=31,1) = 1.10, $p = 0.29$) predicted Test 1 selection. Moreover, the interaction term exhibited no relation to whether or not Test 1 was selection (χ^2 (n=31,3) = 0.88, $p = 0.83$).

The same pair of analyses were conducted for Tests 2 through 4. Presenting cue was not predictive of Test 2 selection (χ^2 (n=31,3) = 6.53, $p = 0.09$), nor was it a significant predictor of Test 2 selection after learning was added to the model (χ^2 (n=31,3) = 3.14, $p = 0.37$). Presenting cue was not predictive of Test 3 selection before (χ^2 (n=31,3) = 2.62, $p = 0.45$) or after learning was added to the model (χ^2 (n=31,3) = 7.21, $p = 0.07$). Learning did not predict Test 3 selection (χ^2 (n=31,1) = 0.16, $p = 0.43$), and the interaction term did not predict selection either (χ^2 (n=31,3) = 7.14, $p = 0.07$). Test 4 results followed the same pattern of results, as presenting cue did not predict Test 4 selection on its own (χ^2 (n=31,3) = 1.26, $p = 0.74$), nor after learning was included in the model (χ^2 (n=31,3) = 2.56, $p = 0.47$). Neither learning (χ^2 (n=31,1) = 1.00, $p = 0.32$) nor the interaction term (χ^2 (n=31,3) = 2.24, $p = 0.52$) predicted Test 4 selection either.

Figure 5 illustrates the test selection analyses reported above, parsing participants into high learning and low learning groups via a median split on their learning phase performance. Low learners exhibited consistency in their test selection behavior across nearly all presenting cue conditions. High learners exhibited much more variability in their testing behavior, including predicted higher rates for Tests 2 and 3 after presented with Cues 2 and 3 respectively. Though these results suggest that these patterns fall short of statistical significance, these differences hint at the possibility that knowledge of the task environment influences sensitivity to presenting cues and test selection.

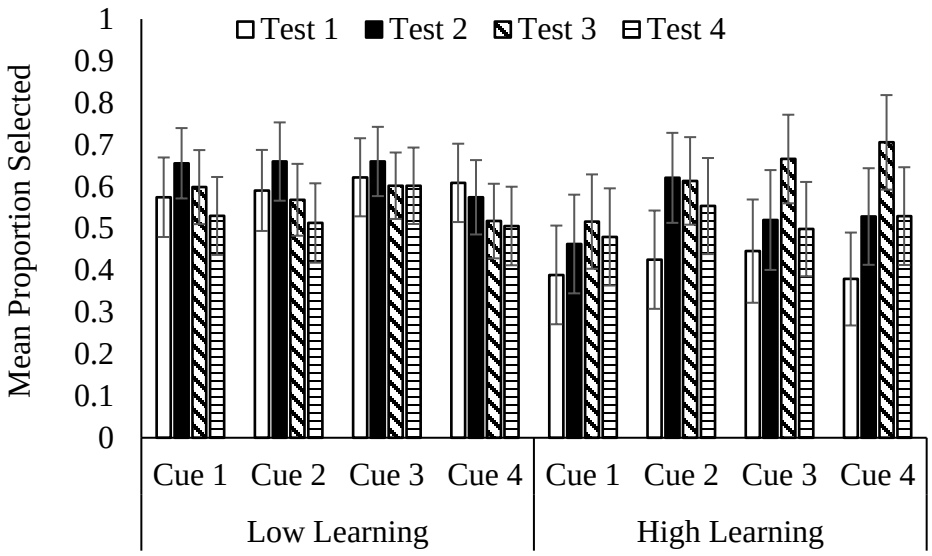


Figure 5. Experiment 1 test selection. The figure illustrates mean proportion of trials for which each test was selected. Selection rate is parsed out by presenting cue and learning performance. The total number of tests selected did not vary by presenting symptom. Error bars illustrate standard errors.

Test preference. Another way to interpret test selection is to consider the order in which tests are selected, denoting those tests selected earlier in a sequence as preferred to those exploited later. Presumably, a test selected earlier in a sequence has been prioritized by the decision-maker for reasons that reflect the perceived properties of the information expected to manifest. For the following analyses, tests selection has been scored with respect to order. Tests selected first were scored as 4, while those selected second, third, or fourth were scored as 3, 2, or 1 respectively. Tests not selected were scored as 0.

A multinomial logistic regression evaluated whether or not presenting cue predicted preference for Test 1 (see Figure 6). The result suggests that preference for Test 1 was not influenced by the presenting cue (χ^2 (n=31,3) = 0.45, p = 0.93). Consistent with the previous set of analyses, learning phase performance was added for an exploratory analysis of the influence of task environment knowledge. Neither presenting cue (χ^2 (n=31,3) =

1.44, $p = 0.70$) nor learning (χ^2 (n=31,1) = 0.93, $p = 0.33$) accounted for Test 1 preference. The interaction term did not predict Test 1 preference either (χ^2 (n=31,3) = 0.98, $p = 0.81$).

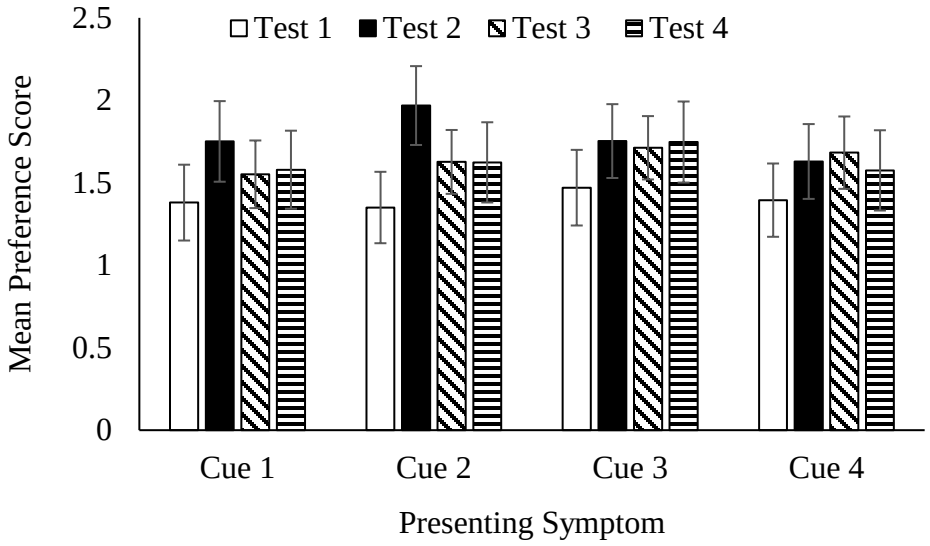


Figure 6. Experiment 1 test preference. The figure illustrates mean preference score for each test broken out by presenting cue. Test preference was not found to be related to presenting cue. Error bars illustrate standard errors.

The same pair of analyses were conducted for Tests 2 through 4. Presenting cue was not predictive of preference for Test 2 (χ^2 (n=31,3) = 5.86, $p = 0.12$), Test 3 (χ^2 (n=31,3) = 1.48, $p = 0.69$), or Test 4 (χ^2 (n=31,3) = 1.41, $p = 0.70$). Test 2 preference was not predicted by presenting cue after learning was included in the model (χ^2 (n=31,3) = 3.22, $p = 0.36$). Learning did not predict Test 2 preference either (χ^2 (n=31,1) = 0.15, $p = 0.70$), nor did the interaction term (χ^2 (n=31,3) = 1.88, $p = 0.60$). Test 3 preference was not related to presenting cue after learning was added to the model (χ^2 (n=31,3) = 6.93, $p = 0.07$). Phase 1 learning (χ^2 (n=31,1) = 0.04, $p = 0.83$) and the interaction term (χ^2 (n=31,3) = 6.96, $p = 0.07$) were also found to be unrelated to Test 3 preference. Both presenting cue (χ^2 (n=31,3) = 1.44, $p = 0.70$) and learning were not predictive of Test 4 preference. Moreover, the

interaction between presenting cue and learning did not predict Test 4 preference (χ^2 (n=31,3) = 2.46, p = 0.48).

Model fitting. The statistical models that evaluated the observed behavior in Experiment 1 failed to demonstrate any link between presenting cue and patterns of testing behavior. A link between hypothesis generation and test selection would be equally challenging to detect in light of that result, suggesting that full-scale HyGene modeling of this experimental task would be of limited value. Instead, simulation of Experiment 1 was carried out to further explore participant sensitivity to the statistical structure of the task environment. Two approaches to test exploitation were formalized in a computational model of the experimental task in an effort to better understand the patterns of testing behavior observed in Experiment 1.

An ideal observer model simulated a strategy that reflected perfect knowledge of the task environment, where the test with the highest diagnostic value was always selected. Bayesian diagnosticity was the information metric capitalized upon by the model and was estimated for each test (T) consistent with Equation 17. The value of a test was computed as the average diagnosticity of each possible test outcome (d_j). Posterior belief distributions were perfectly calibrated to the result of each selected test, and the diagnostic values of remaining tests were computed after each test selection. Equation 18 was used to compute posterior belief. The random model simulated a strategy for selecting medical tests without any consideration of their informative properties. Model behavior was recorded across all possible presenting cue and test result combinations, as well as all possible trials where between 1 and 4 tests were selected.

$$diagnosticity(T) = \frac{\sum_{d_j} P(d_j) * \max\left(\frac{p(d_j|H_1)}{p(d_j|H_0)}, \frac{p(d_j|H_0)}{p(d_j|H_1)}\right)}{d} \quad (17)$$

$$P(H_1|d_j) = \frac{p(d_j|H_1)p(H_1)}{p(d_j|H_1)p(H_1) + p(d_j|H_0)p(H_0)} \quad (18)$$

I used the softmax learning rule (Fu & Anderson, 2006; Pleskac, 2012; Sutton & Barto, 1998) to introduce noise to the ideal observer model's perfect sensitivity to test diagnosticity. Equation 19 modified the probability of each possible action the model could take—selecting a test (a_i). The model iterated through 3 levels of τ (low, moderate, high). When τ was low, the most diagnostic tests were selected at a rate just short of the ideal observer. When τ was high, test selection approached patterns that mimicked the random model.

$$p(a_i) = \frac{e^{p(a_i)/\tau}}{\sum_{n=1}^{j} e^{p(a_i)/\tau}} \quad (19)$$

Likelihoods were computed for each trial of participant behavior across all models. Specifically, the order in which participants selected tests were used to evaluate participant performance for the purpose of estimating a fit to each model. Likelihoods were aggregated for each participant and used to compute G^2 statistics. The G^2 for each model was compared against the random model and corrected for parameters to compute a Bayes Information Criterion (*BIC*). Table 4 lists aggregate *BIC* statistics for all instantiations of the ideal observer model with respect to the random model, where more negative values represent better fit. *BIC*s larger than zero indicate that the random model was a better fit.

Table 4. Experiment 1 aggregate fit statistics for all models.

		Tau (τ)			
	Ideal	0.2	0.8	1.4	Random
G^2	14419.93	4106.03	1668.49	1400.52	1106.79
BIC	-13319.57	-3011.10	-574.56	-306.58	0.00

The ideal observer model had the best fit as far as the aggregate totals are concerned. While the results reported thus far suggest that participant performance was far from ideal, the model fitting results suggest that participants were sensitive to information available in the decision environment. This finding was further supported because the ideal observer was the best fit for 27 of 31 participants. The remaining 4 participants were fit by the softmax model with a Tau of 0.2. Thus, all 31 participants were best fit by models with high access to the diagnostic value of the tests available in the task. Fit statistics for each participant were tabled in Appendix A.

3.2.3 Discussion.

Analysis of the behavior captured in Experiment 1 has failed to detect evidence that the presenting cue had any influence on hypothesis testing behavior. Thus, Experiment 1 has provided no support for the theoretical model illustrated in Figure 1. The hypothesis-guided testing hypothesis explicitly anticipates that pre-testing cueing of differential hypothesis sets would result in predictable patterns of testing behavior. If properly encoded, the value of available tests during Phase 2 of the task should shift dramatically, as their association with the information activated in response to any presenting cue should

vary substantially. At face value, the results of Experiment 1 cast doubt on the hypothesis-guided testing hypothesis.

These results are incompatible with previous work that found evidence to support the notion that diagnosticity can be detected and utilized by participants in a sequential hypothesis-testing task (Illingworth & Thomas, 2015). In their study, Illingworth and Thomas reported that there was a strong preference for high diagnosticity information sources regardless of the shape of cost distributions (symmetric vs. asymmetric) and the distinctiveness of the test outcomes. Nelson (2005; Nelson et al., 2010) has repeatedly found evidence in support of informed information acquisition. Although his argument primarily regarded evidence for the probability gain metric of information, a number of his participants appeared to engage in diagnosticity-sensitive information acquisition.

Experiment 1 also failed to replicate findings that demonstrated the conditions under which participants engage in pseudo-diagnostic and diagnostic search. Lange, Thomas, and Dougherty (2010) reported evidence suggesting that pseudo-diagnostic search was a direct byproduct of the generation process such that it manifested only when a single hypothesis was considered during hypothesis testing. When evidence of multiple generated hypotheses was present, participants engaged in more diagnostic testing practices. The presenting cue matrix detailed in Table 2 was designed to elicit such behavior by cueing varying numbers of diseases with each presenting symptom. Not only did test selection go unaffected by this manipulation, but the total number of tests selected remained constant with respect to presenting cue—a sign that foraging behavior was unresponsive to uncertainty inherent to the task.

While the results of Experiment 1 nullified the value of extensive HyGene process modeling, the task simulation proved to be insightful. Specifically, the results of the simulation provide evidence to suggest that participants exhibited behavior that was sensitive to the diagnosticity manipulation. Clearly, this result must be considered within the context of the null results reported in the statistical analyses of Experiment 1 performance. Nothing about the behavior recorded in this study indicate that participants approached perfect understanding of the statistical ecology implemented in this experiment. If anything, the model fits can be used to conclude that participants were not selecting tests entirely at random and had internalized some information available in the experimental task.

It is worth noting that one participant did not engage in any testing behavior at all and a few others did so on fewer than 5 trials. These individuals are included amongst those that were best fit by highly informed models. Additionally, a number of participants were recorded as having completed trials in under 2 seconds. It is highly unlikely that a decision-maker carefully considering which of the available tests would best support their diagnoses could achieve such a feat, advancing concern that participants were not meaningfully engaged in the task.

Difficulties internalizing the complex task environment is one possible substantive cause for the poor performance exhibited by participants during this experiment. In a comparable task, Illingworth and Thomas (2015) reported that participants completed the learning phase of their experiment having been correct on approximately 56% of trials—23 points above chance performance. Participants in Experiment 1 averaged a rate of 33%

correct responses—only 8 points above chance performance—after a learning phase that was nearly four times as long in terms of trials.

Perhaps more troubling is the fact that inclusion of Phase 1 learning in the analyses of test selection and test preference also failed to unveil an influence of presenting cue or an interaction between presenting cue and learning. At a minimum, participants at the higher end of the learning distribution, who presumably possessed a better understanding of the task environment relative to other participants—would be expected to show signs of sensitivity to presenting cue if the hypothesized generation processes played any role in test selection. It is worth noting, however, that the best performer completed Phase 1 with a 46% accuracy rate—far short of the mean performance exhibited by participants in prior experiments.

Although the findings reported for Experiment 1 provide little support for any of the HyGene processes predicted to drive hypothesis-testing behavior, it seems premature to conclude that the predictions of the hypothesis-driven valuation model (Figure 2) were fallacious. One specific design modification that would benefit future study of hypothesis-guided testing is to reduce the demand on participants during Phase 1. Although compelling arguments have been made for avoiding forced-choice tasks with few alternatives and binary outcomes for information depositories (Illingworth & Thomas, 2015; Weber & Milliman, 1997), the complexity implemented in the environmental ecology of this experiment is not necessary for rigorous evaluations of depository exploitation. Illingworth and Thomas (2015), for example, implemented an ecology with one less hypothesis and tests with higher diagnosticity to investigate sensitivities to diagnosticity and cost in a sequential data acquisition task.

I argue that additional study of hypothesis set cueing and subsequent hypothesis testing patterns is warranted in spite of the outcome of Experiment 1. The lack of evidence in support of the memory mechanisms hypothesized to be foundational to testing behavior should not be off-putting when paired with sparse evidence of an effort spared to learn task-relevant information or produce a carefully considered diagnosis. Examining testing patterns within the context of hypothesis generation still possesses the potential to elucidate the underlying mechanisms of the behavior.

3.3 Experiment 2 - Time Pressure, Generation, and Test Preference

A core tenet of the HyGene architecture posits that generated hypotheses are maintained and governed by working memory dynamics (Dougherty, Thomas, & Lange, 2010; Thomas et al., 2008; Thomas, Dougherty, & Buttaccio, 2014). Numerous studies have found support for the relation between working memory and hypothesis generation, demonstrating the downstream importance of working memory capacity (Dougherty & Hunter, 2003a; Dougherty & Hunter, 2003b), divided attention (Sprenger, et al., 2011), time pressure (Dougherty & Hunter, 2003b), and temporal working memory dynamics (Lange, Buttaccio, Davelaar, & Thomas, 2014; Lange, Thomas, Buttaccio, Illingworth, & Davelaar, 2013; Lange, Thomas, & Davelaar, 2012). Working memory constraints have generally been shown to influence the size and quality of hypothesis sets considered by decision-makers, as detected by their effects on diagnosis and probability judgment. Other downstream decision behavior related to belief states, such as hypothesis-guided testing, are likely subject to the same working memory processes that influence generation.

Dougherty and Hunter (2003a) were first to argue that the process of generating hypotheses was time-consuming, suggesting that more hypotheses are generated as more

time is allowed for generation. They would subsequently show that probability judgments are more subadditive when estimated under time constraints—a sign of judgments being derived from an impoverished set of hypotheses. Dougherty and Hunter concluded that time pressure abbreviates the generation process and results in poorly calibrated judgments caused by suboptimal sets of hypotheses. Experiment 2 extended study of working memory constraints and hypothesis generation to evaluate the hypothesis-guided testing hypothesis by investigating how those dynamics influence testing behavior. Specifically, time pressure was implemented within the MDG to test the relation between truncated hypothesis generation and test preference.

The prediction tested with Experiment 2 is deficient test selection (i.e., greater utility loss) under time constraints. When Cue 4 is presenting, for example, Test 4 would be preferred earlier in the testing sequence in trials without time pressure than under time constraints. Once more, data collected for this experiment were evaluated via computational modeling.

Evidence of deficits in testing behavior as a consequence of constraining the time allotted for completion of the MDG task would support the hypothesis-guided testing hypothesis. Such a finding would support the claim of a shared cognitive mechanism for test selection with probability judgment and diagnosis, as all would exhibit deficits under time-limited conditions.

3.3.1 Method.

Undergraduate students enrolled at the Georgia Institute of Technology were recruited to participate in this study via an online experiment management system (SONA

Systems). In total, 37 participants completed the experiment. All participants received partial course credit for their involvement in the study.

The learning phase of Experiment 2 was completed over 24 blocks of 20 trials, resulting in a total of 480 learning trials. Learning trails were designed as has been described for Experiment 1. Accurate performance was incentivized by awarding correcting responses with points ($1000) that would be deposited in a bank that accumulated points over the course of the learning phase.

The ecological structure of this experiment was designed to exploit the possible effects of time pressure on the generation process. Table 5 lists the hypothesis-presenting cue relations that controlled stimuli presentation within the task. Relative to Experiment 1, twice as many cues were strongly associated with three hypotheses (Cues 3 and 4). The remaining cues (1 and 2) suggested one strong hypothesis but were also associated with one slightly weaker competitor. The presence of multiple cues associated with many hypotheses enhanced the number of opportunities to detect any deficits in testing behavior that may have emerged as a consequence of considering incomplete sets of hypotheses.

Table 5. Presenting sign ecology for Experiment 2.

	Presenting Symptoms			
	Cue 1	Cue 2	Cue 3	Cue 4
Hyp 1	.55	.05	.35	.05
Hyp 2	.25	.05	.35	.35
Hyp 3	.05	.25	.35	.35
Hyp 4	.05	.55	.05	.35

Presentation of test outcomes respected the test-hypothesis ecology outlined in Table 6. As was the case in Experiment 1, the diagnostic properties of the tests were meant to mirror the cue-hypothesis table. For example, Cue 1 was suggestive of Hypotheses 1 and 2, while Test 1 outcomes differentiated between the same two hypotheses.

Table 6. Test outcome ecology for Experiment 2.

		Diagnostic Tests			
		Test1	Test2	Test3	Test4
Hypothesis 1	Outcome 1	.65	.30	.60	.30
	Outcome 2	.20	.40	.20	.40
	Outcome 3	.15	.30	.20	.30
Hypothesis 2	Outcome 1	.15	.30	.20	.60
	Outcome 2	.20	.40	.60	.20
	Outcome 3	.65	.30	.20	.20
Hypothesis 3	Outcome 1	.30	.15	.20	.20
	Outcome 2	.40	.20	.20	.60
	Outcome 3	.30	.65	.60	.20
Hypothesis 4	Outcome 1	.30	.65	.30	.20
	Outcome 2	.40	.20	.40	.20
	Outcome 3	.30	.15	.30	.60
Diagnosticity		2.09	2.09	2.03	2.03

The ecology of Experiment 2 was designed to emulate that for Experiment 1 such that the diagnosticity for each test was approximately equal when the hypotheses were equally likely. The diagnosticity of each test became disparate after belief distributions were altered by the presenting cue and the outcomes of selected tests. Unlike Experiment 1, the test outcome ecology for Experiment 2 was uniformly manipulated such that Tests 1 and 2 possessed the exact same diagnostic properties (i.e., patterns of test outcomes) but disambiguated different pairs of hypotheses. Tests 3 and 4 also shared the exact same diagnostic properties but were designed to differentiate separate triads of hypotheses.

The test phase of Experiment 2 was completed over 4 blocks of 20 trials each, resulting in a total of 80 test trials. Selection of tests during the test phase of the experiment occurred as described for Experiment 1, where a presenting cue appeared at the onset of a trial and was followed by sequential test selection. A 2 (time constraint: present, not) x 2 (cue: 1 hypothesis, 3 hypotheses) x 2 (counter-balanced order of time constraint conditions) mixed design was implemented, where both time constraint and cue type was manipulated

within-subject. Time constraint trials were completed in 2 consecutive blocks. The order in which pairs of blocks were experienced was manipulated between-subjects.

In time constraint trials, participants had 2 seconds after the onset of a trial to exploit the tests they deemed most informative prior to submitting a diagnosis. In other words, participants observed the presenting symptom, generated hypotheses, and selected tests before the 2-second time frame was complete. Participants were locked out of additional testing and diagnoses were elicited after the 2-second time interval was exhausted. Alternatively, participants were entirely self-paced in the time constraint absent condition. Diagnoses were not submitted under time pressure regardless of condition.

3.3.2 Results.

Learning. A logistic regression evaluated whether accuracy was predicted by block. Block was not found to be predictive of learning phase accuracy (χ^2 (n=37,23) = 25.12, p = 0.34), suggesting that performance did not improve over the course of the 24 blocks. However, participants were 1.44 times more likely to submit a correct diagnosis in the final block as they were in the first (β = 0.44, SE = 0.12, p <.001). There was also a significant linear trend relating block and accuracy (χ^2 (n=37,1) = 6.33, p = 0.01). These results are illustrated in Figure 7. Once again, mixed learning results imply that knowledge of the task environment varied across participants, and can potentially aid in elucidating performance during Phase 2.

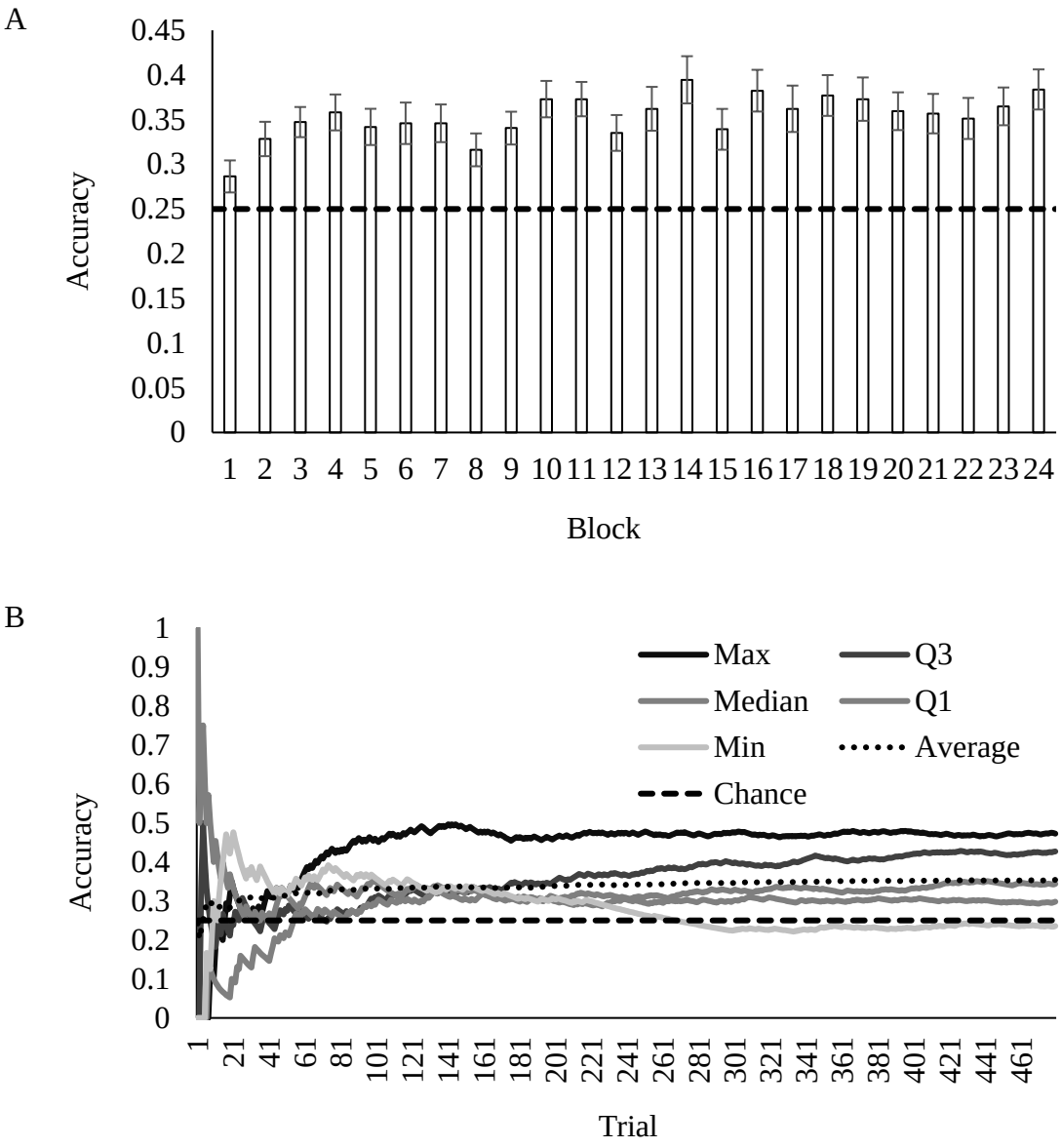


Figure 7. Experiment 2 learning. Panel A illustrates Experiment 2 learning phase accuracy broken out by block. The dotted line represents chance performance (25%). Error bars represent standard error. Panel B tracks proportion correct for 5 participants across all trials of Phase 1, as well as sample average (black, dotted line) and chance performance (black, dashed line). The worst performer scored below chance accuracy, while the best performer fell shy of 50% accuracy.

Accuracy. A logistic regression evaluated how well accuracy was predicted by block. Block was not found to be predictive of learning phase accuracy (χ^2 (n=37,1) = 0.73, p = 0.39), suggesting that performance did not improve over the course of the test

phase. Participants were nearly equally likely (O = 0.93) to submit a correct diagnosis in the final block as they were in the first (β = -0.07, SE = 0.08 , p =.39). Moreover, accuracy in the first block (M = 0.37, SE = 0.02) and the second block (M =0.38, SE = 0.02) were approximately equal to learning phase performance.

Stopping. A multinomial logistic regression tested for the influence of presenting cue and time pressure on the number of tests selected during Phase 2. Neither presenting cue (χ^2 (n=37,3) = 5.71, p = 0.13) nor time pressure (χ^2 (n=37,1) = 0.12, p = 0.73) was predictive of the number of tests selected by participants. The interaction between presenting cue and time pressure also failed to predict the number of tests selected (χ^2 (n=37,3) = 0.54, p = 0.91). As was the case in Experiment 1, the main manipulations of the experimental task were not found to have had any effect on the stopping rules adopted by participants. Total testing behavior for Experiment 2 is illustrated in Figure 8.

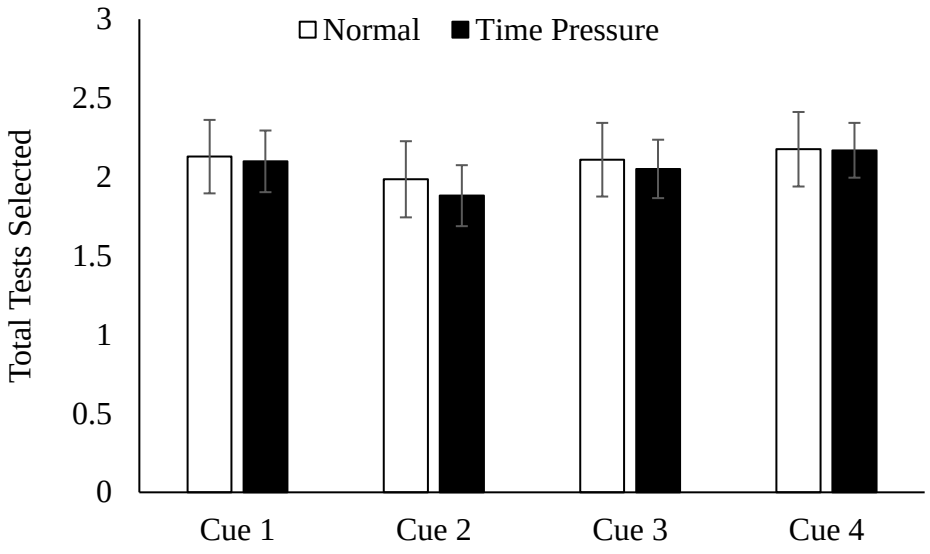


Figure 8. Experiment 2 total testing. Figure illustrates mean total tests selected broken out by presenting cue and time pressure. Participants were consistent in their foraging behavior regardless of the experimental conditions. Error bars represent standard errors.

Provided that the analysis of Phase 1 performance illustrated varied knowledge of the task environment, another regression was run to explore what if anything can be learned by including learning in the analysis. Total tests selected was regressed on presenting cue, time pressure, learning, and all possible interaction terms. Presenting cue (χ^2 (n=37,3) = 5.10, p = 0.16), time pressure (χ^2 (n=37,1) = 1.98, p = 0.16), and learning (χ^2 (n=37,1) = 0.46, p = 0.50) were not found to predict total tests selected. Moreover, the two-way interactions between presenting cue and time pressure (χ^2 (n=37,3) = 0.37, p = 0.95), learning and presenting cue (χ^2 (n=37,3) = 3.43, p = 0.33), and learning and time pressure (χ^2 (n=37,1) = 0.75, p = 0.39) also failed to predict total tests selected. The three-way interaction between presenting cue, time pressure, and learning did not predict total tests selected either (χ^2 (n=37,3) = 0.35, p = 0.95). Thus, inclusion of learning phase performance in the model failed to find evidence that said learning interacted with any manipulated variable to account for search termination behavior.

Test selection. The frequency with which each test was selected was analyzed in a series of binomial regressions to evaluate what influence, if any, presenting cue and time pressure exerted on patterns of test selection (see Figure 9). A binomial regression found that neither presenting cue (χ^2 (n=37,3) = 1.04, p = 0.79) nor time pressure (χ^2 (n=37,1) = 0.38, p = 0.54) predicted selection of Test 1. The interaction term also fell short of predicting Test 1 selection (χ^2 (n=37,3) = 6.46, p = 0.09). Test 2 selection was regressed on presenting cue (χ^2 (n=37,3) = 2.93, p = 0.40) and time pressure (χ^2 (n=37,1) = 0.29, p = 0.59), finding that neither variable was a significant predictor of selection. The interaction between presenting cue and time pressure also failed to predict Test 2 selection (χ^2 (n=37,3) = 2.14, p = 0.54).

Test 3 selection was also regressed on presenting cue and time pressure. Neither manipulated variable was related to Test 3 selection (χ^2 (n=37,3) = 2.93, p = 0.40 and χ^2 (n=37,) = 0.29, p = 0.59 respectively). The interaction between presenting cue and time pressure was not a significant predictor of Test 3 selection (χ^2 (n=37,3) = 2.14, p = 0.54). Analysis of Test 4 selection took followed the same pattern as the previous three tests. Neither presenting cue (χ^2 (n=37,3) = 2.93, p = 0.40) nor time pressure (χ^2 (n=37,1) = 0.29, p = 0.59) were predictors of Test 4 selection. The interaction of these two variables also failed to predict Test 4 selection (χ^2 (n=37,3) = 2.14, p = 0.54).

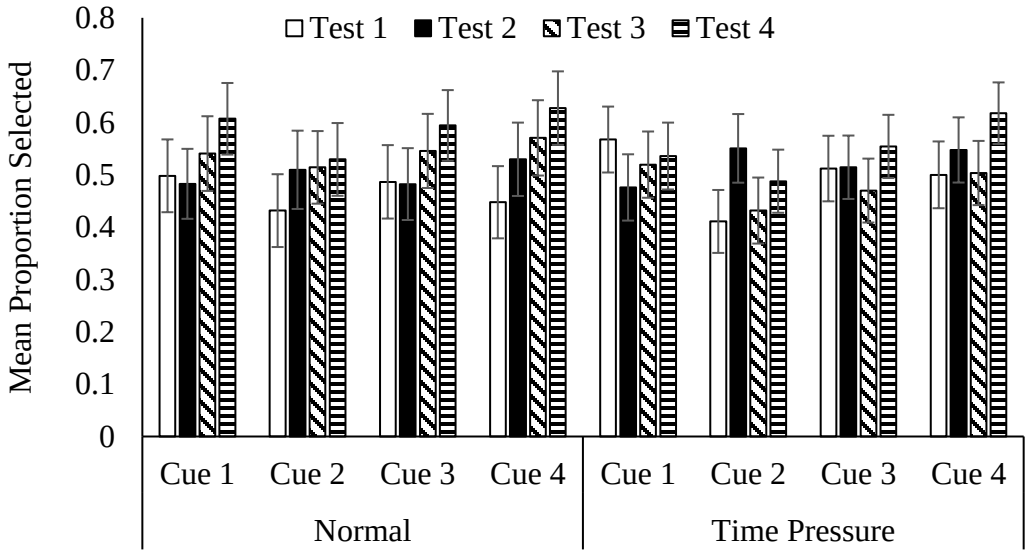


Figure 9. Experiment 2 test selection. Figure illustrates mean test selection broken out by presenting cue and time pressure. Participants were consistent in their foraging behavior regardless of the experimental conditions. Error bars represent standard errors.

The series of test selection analyses were repeated to incorporate learning phase performance for the purpose of exploring the degree to which knowledge of the task environment influenced testing patterns. Test 1 was not predicted by presenting cue (χ^2 (n=37,3) = 3.98, p = 0.26), time pressure (χ^2 (n=37,1) = 0.89, p = 0.35), or learning (χ^2

(n=37,1) = 0.39, p = 0.53). None of the interaction terms predicted Test 1 selection, including all of the two-way inter actions. This included that between presenting cue and time pressure (χ^2 (n=37,3) = 0.10, p = 0.99), presenting cue and learning (χ^2 (n=37,3) = 4.55, p = 0.21), and time pressure and learning (χ^2 (n=37,1) = 0.28, p = 0.60), as well as the three-way interaction between presenting cue, time pressure, and learning (χ^2 (n=37,3) = 0.51, p = 0.92). Similarly, Test 2 selection was not predicted by presenting cue (χ^2 (n=37,3) = 0.11, p = 0.99), time pressure (χ^2 (n=37,1) = 0.97, p = 0.32), or learning (χ^2 (n=37,1) = 0.01, p = 0.93). The two-way interactions, including that between presenting cue and time pressure (χ^2 (n=37,3) = 1.24, p = 0.74), presenting cue and learning (χ^2 (n=37,3) = 0.47, p = 0.92), and time pressure and learning (χ^2 (n=37,1) = 0.37, p = 0.54), all failed to predict Test 2 selection. The same was true for the three-way interaction, which was not a significant predictor of Test 2 selection (χ^2 (n=37,3) = 1.31, p = 0.73).

Test 3 selection found a predictor in time pressure (χ^2 (n=37,1) = 3.97, p = 0.046), suggesting that Test 3 was selected at a higher rate under time pressure than without. However, a follow up analysis showed that the mean proportion of trials was not significantly greater in the time pressure condition (z = 1.60, p = 0.11). Neither presenting cue (χ^2 (n=37,3) = 0.97, p = 0.81) or learning (χ^2 (n=37,1) = 1.75, p = 0.19) predicted Test 3 selection. The two-way interactions were not predictive of Test 3 selection: presenting cue by time pressure (χ^2 (n=37,3) = 1.42, p = 0.70), presenting cue by learning (χ^2 (n=37,3) = 1.05, p = 0.79), and time pressure by learning (χ^2 (n=37,) = 2.55, p = 0.11); neither was the three-way interaction between presenting cue, time pressure, and learning (χ^2 (n=37,3) = 1.88, p = 0.60). Presenting cue (χ^2 (n=37,3) = 4.59, p = 0.20), time pressure (χ^2 (n=37,1) = 0.24, p = 0.62), and learning (χ^2 (n=37,) = 0.08, p = 0.78) did not predict Test 4 selection.

None of the interaction terms predicted Test 4 selection, including all two-way interactions—including presenting cue by time pressure (χ^2 (n=37,3) = 2.82, p = 0.42), presenting cue by learning (χ^2 (n=37,3) = 4.70, p = 0.20), and time pressure by learning (χ^2 (n=37,1) = 0.05, p = 0.83)—and the three-way interaction between presenting cue, time pressure, and learning (χ^2 (n=37,3) = 2.79, p = 0.43).

Test preference. Consistent with the analyses for Experiment 1, selection behavior was transformed into a preference score (e.g., first selected test was scored 4, fourth selected test was scored 1, unselected tests were scored 0). A suite of multinomial logistic regression analyses were run to evaluate if the presenting cue and time pressure predicted the preference score for each available test. Test 1 preference was not predicted by presenting cue (χ^2 (n=37,3) = 0.90, p = 0.82) or time pressure (χ^2 (n=37,1) = 1.64, p = 0.20). The interaction between presenting cue and time pressure also failed to reach statistical significance (χ^2 (n=37,3) = 5.42, p = 0.14). A similar pattern of results emerged for Test 2 preference. Neither presenting cue (χ^2 (n=37,3) = 0.45, p = 0.93), time pressure (χ^2 (n=37,3) = 0.45, p = 0.93), nor the interaction between the two variables (χ^2 (n=37,3) = 0.45, p = 0.93) predicted Test 2 preference.

Test 3 preference was not predicted by presenting cue (χ^2 (n=37,3) = 2.28, p = 0.52) or time pressure (χ^2 (n=37,1) = 0.91, p = 0.34). Neither did the interaction between presenting cue and time pressure did not reach statistical significance for predicting Test 3 preference (χ^2 (n=37,3) = 3.00, p = 0.39). The results for Test 4 preference followed the trend established by the previous three tests, as presenting cue (χ^2 (n=37,3) = 0.45, p = 0.93), time pressure (χ^2 (n=37,3) = 0.45, p = 0.93), and the interaction term (χ^2 (n=37,3) = 0.45, p = 0.93) all failed to predict the variable. The manipulated variables for Experiment

2 predicted preference for none of the tests available in the task. These scores were illustrated in Figure 10.

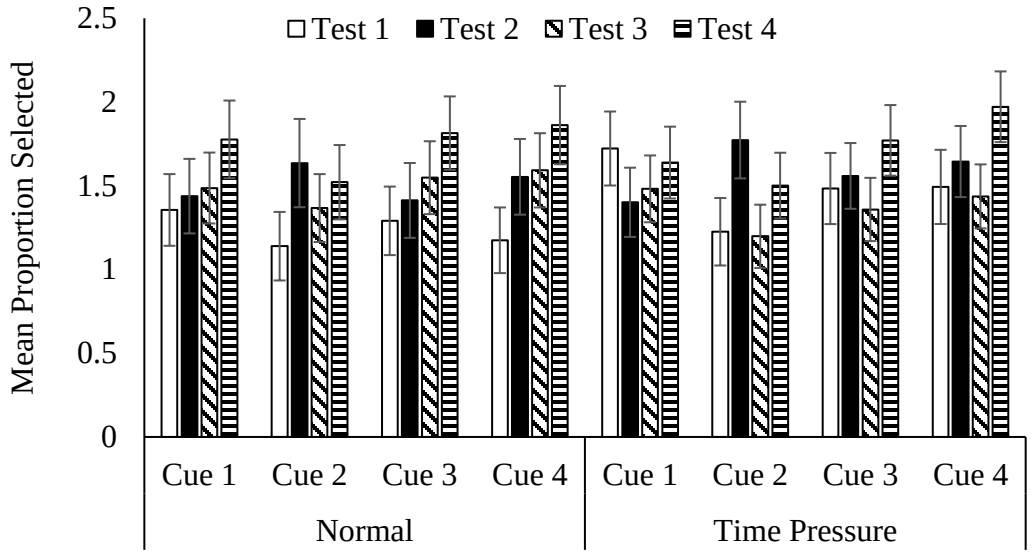


Figure 10. Experiment 2 test preference. Test preference scores are broken out by presenting cue and time pressure. The manipulated variables did not account for the order in which participants selected tests. Error bars represent standard errors.

A second suite of multinomial regressions were run to evaluate how learning phase performance may enlighten test preference. Test 1 preference was not predicted by presenting cue (χ^2 (n=37,3) = 2.41, p = 0.49), time pressure (χ^2 (n=37,1) = 0.72, p = 0.40), or learning (χ^2 (n=37,1) = 0.13, p = 0.71). None of the two-way interactions were predictive of Test 1 preference, including that between presenting cue and time pressure (χ^2 (n=37,3) = 0.45, p = 0.93), presenting cue and learning (χ^2 (n=37,3) = 3.00, p = 0.39), and time pressure and learning (χ^2 (n=37,1) = 0.07, p = 0.80). The three-way interaction between presenting cue, time pressure, and learning did not predict Test 1 preference (χ^2 (n=37,3) = 1.33, p = 0.72).

Analysis of Test 2 preference revealed a similar pattern of results. Presenting cue (χ^2 (n=37,3) = 0.52, p = 0.91), time pressure (χ^2 (n=37,1) = 0.78, p = 0.38), and learning

(χ^2 (n=37,3) = 0.06, p = 0.81) all fell short of predicting test selection. None of the interaction terms were predictive of Test 2 preference. These included the two-way interactions between presenting cue and time pressure (χ^2 (n=37,3) = 1.41, p = 0.70), presenting cue and learning (χ^2 (n=37,3) = 1.08, p = 0.78), and time pressure and learning (χ^2 (n=37,1) = 0.15, p = 0.70), and the three-way interaction between presenting cue, time pressure, and learning (χ^2 (n=37,3) = 1.45, p = 0.69). Test 3 preference was predicted by time pressure (χ^2 (n=37,1) = 4.60, p = 0.03), such that Test 3 was preferred at a higher rate under time pressure than without (z = 2.25, p = 0.025). Neither presenting cue (χ^2 (n=37,3) = 2.88, p = 0.41) nor learning (χ^2 (n=37,1) = 1.83, p = 0.18) predicted Test 3 preference. None of the two-way interactions predicted Test 3 preference, including that for presenting cue and time pressure (χ^2 (n=37,3) = 3.50, p = 0.32), presenting cue and learning (χ^2 (n=37,3) = 3.03, p = 0.39), and time pressure and learning (χ^2 (n=37,3) = 3.71, p = 0.054)—though the last of these approached significance. The three-way interaction between presenting cue, time pressure and learning was also found not to predict Test 3 preference (χ^2 (n=37,3) = 4.36, p = 0.22).

Test 4 preference was not predicted by presenting cue (χ^2 (n=37,3) = 4.83, p = 0.18), time pressure (χ^2 (n=37,1) = 0.15, p = 0.70), or learning (χ^2 (n=37,1) = 0.06, p = 0.81). Neither was it predicted by any of the interaction terms. This included all of the two-way interactions—presenting cue by time pressure (χ^2 (n=37,3) = 2.79, p = 0.42), presenting cue by learning (χ^2 (n=37,3) = 5.16, p = 0.16), and presenting cue by learning (χ^2 (n=37,1) = 0.19, p = 0.66)—and the three-way interaction between presenting cue, time pressure, and learning (χ^2 (n=37,3) = 2.93, p = 0.40). Inclusion of learning in this series of analysis did nothing to change the outcome of the results. Generally speaking, none of the

manipulated variables were found to have had any impact on testing patterns as measured by the selection of the available tests.

Model fitting. As was the case for Experiment 1, statistical evidence for a link between presenting cue and testing behavior eluded Experiment 2. The modeling endeavor undertaken to evaluate Experiment 2 mimicked that of Experiment 1 for the purpose of exploring participant sensitivity to the experimental task and the influence of time pressure. The same two approaches to test exploitation formalized for the Experiment 1 simulation were implemented in a computational model Experiment 2. The ideal observer model was perfectly calibrated to the statistical structure of the disease-test result matrix and always selected the most diagnostic, available test. The posterior belief distribution for the ideal observer were always perfectly calibrated to the information state reached after selection of each test. Consistent with the simulation for Experiment 1, Equation 17 was used to estimate the diagnostic value of each test (T) and Equation 18 was used to update posterior belief distributions after each testing event. A random model was designed to select tests without any consideration of their informative properties. The softmax learning rule (Equation 19) was implemented across three levels of τ to interfere with the ideal observer model's perfect sensitivity to diagnosticity. Again, sensitivity to diagnosticity declined as τ increased.

Likelihoods were computed for each trial of participant behavior across all models. Specifically, the order in which participants selected tests were used to evaluate participant performance for the purpose of estimating a fit to each model. Likelihoods were aggregated for each participant and used to compute G^2 statistics. The G^2 for each model was compared against the random model and corrected for parameters to compute a Bayes

Information Criterion (*BIC*). Table 7 lists aggregate *BIC* statistics for all instantiations of the ideal observer model with respect to the random model, where more negative values represent better fit. *BIC*s larger than zero indicate that the random model was a better fit.

Table 7. Experiment 2 aggregate fit statistics for all models.

		Tau (τ)				
		Ideal	0.2	0.8	1.4	Random
No Pressure	G^2	14299.06	2908.14	1012.80	841.75	660.50
	BIC	-13645.90	-2262.24	-1012.80	-195.85	0.00
Time Pressure	G^2	12959.91	2820.64	1022.32	849.22	660.50
	BIC	-12306.70	-2174.73	-376.42	-203.32	0.00

Each participant was fit twice: Once for all behavior exhibited under normal conditions and once more for all behavior exhibited under time pressure. As can be seen in Table 7, the ideal observer model exhibited the best aggregate fit in the no time pressure condition and the time pressure condition. If the time pressure manipulation was having its predicted effect, models indicative of less sensitivity to the diagnosticity in the environment would exhibit better fit under time pressure.

Table 8. Experiment 2 proportion of participants fitting each model.

	Tau (τ)				
	Ideal	0.2	0.8	1.4	Random
No Pressure	0.84	0.00	0.00	0.11	0.05
Time Pressure	0.92	0.03	0.0	0.05	0.00

The fits for individual participants mirror the aggregate fits, as the ideal observer model best fit the majority of participants. However, the highest tau model best fit 4

participants and the random model best fit an additional 2 participants, indicating there were a few participants who were exhibiting near-random test selection behavior. The time pressure model fitting exhibited the same general trend, as even more participants fit the ideal observer model under time pressure than without. Only two participants showed the predicted pattern of fitting higher information models without time pressure and lower information models under time pressure. Fit statistics for each participant were reported in Appendix B.

3.3.3 Discussion.

As was the case in Experiment 1, the behavior reported for Experiment 2 has failed to find evidence in support of the hypothesis-guided valuation model—the notion that presenting cue elicited the generation of differentially activated hypotheses, which would give rise to markedly distinctive patterns in hypothesis testing behavior. Specifically, the patterns in test selection and test preference scores were found to be unrelated to presenting cue.

As was discussed following Experiment 1, a number of studies have reported instances when participants were responding to the diagnostic value of information sources (Illingworth & Thomas, 2015; Lange, Thomas, & Dougherty, 2010; Nelson, 2005; Nelson et al., 2010). The ecological/statistical structure of the task in Experiment 2 differs from previous studies such that all sources of information were equal provided that participants believed all diseases were equally likely at the start of a trial. However, trials were structured in a way that would never allow for a flat belief distribution across all diseases, as the presenting cue was always available to participants at the start of each trial. That is, any participant who had learned the cue structure detailed in Table 4 should have exhibited

differential test preference on a per-trial basis. Moreover, the total number of tests selected remained constant with respect to presenting cue. Cues 1 and 2 were designed to relate strongly to two hypotheses while Cues 3 and 4 were designed to relate strongly to three hypotheses. As uncertainty generally increases with the number of hypotheses under consideration, it was predicted that more tests would be exploited in response to Cues 3 and 4 than Cues 1 and 2.

The simulation results for Experiment 2 provide some indication that participants were responsive to the statistical structure of the task, but should be interpreted cautiously given that they are not supported by the statistical analyses reported previously. The majority of participants were best accounted for by models that were highly sensitivity to the diagnostic value of the available tests, even when time constraint was present. Only 2 participants showed signs of less sensitivity to information under time pressure—the pattern of behavior consistent with the time pressure effects reported by Hunter and Dougherty (2003a).

Generation processes remain a possible account for the reported results. As mentioned previously, generating hypotheses is a time-consuming process (Dougherty & Hunter, 2003a). To whatever degree possible—in spite the poor performance exhibited during the task—time-constrained hypothesis generation could explain why a couple of participants wound up better fit by a random acting model. Similarly, Kerstholt (1994) explains time pressure deficits in decision-making as a result of limits on the time needed to internalize and interpret novel data. In a sequential hypothesis testing task such as that deployed in Experiment 2, it may be the case that test selections were poorly mapped to

presenting cue because participants were incapable of properly interpreting the implications of the presenting symptom (i.e., the correct posterior distribution).

The diagnostic task in Experiment 2 can be examined with respect to its properties as a dynamic task, as beliefs were expected to change over time in response to each novel piece of information acquired by the decision-maker. Framed within the context of risky decisions making paradigms—where sequential sampling of information either enhances or diminishes the probability of earning the payoff—sequential sampling theories such as decision field theory would posit that test selection under time pressure would change as a consequence of threshold changes (Dror, Busemeyer, & Basola, 1999). That is, whether or not additional testing takes place depends on the risk inherent in the decision environment, where high-risk environments lend to increased risk-taking and low-risk environments lead to reduced risk-taking. Ultimately, sequential sapling models should be disregarded as potential accounts for the observed data as the rate of testing did not vary with respect to time pressure.

Continued study may be necessary to account fully the behavior measured during Experiment 2. It is questionable, however, what influence these data should impart on future research design as the behavior exhibited by participants in this task case doubt on the validity of any conclusion drawn from the results. As was the case in Experiment 1, a fair number of participants either engaged in no testing behavior (n =7) or engaged in testing for fewer than 5 trials (n = 7). Once more, several participants were detected having completed trials in under 2 seconds (unreported data). The experiments implemented to test the hypothesis-guided testing hypothesis were demanding tasks. Both the learning phase of the task and the subsequent test phase necessitated attention to detail for actions

exhibited by participants to result in successful completion of the task as measured by diagnostic accuracy. Evidence of such behavior remained concealed from any analysis conducted for this experiment.

As was true for Experiment 1, poor learning likely played a role in the insensitivity to the statistical structure of the task environment. Participants in Experiment 2 averaged a rate of 35% correct responses with the best performer maxing out at about 47% accuracy in the learning phase of the study. Though previously unreported, accuracy during the test phase did not improve. The best performer completed the test phase with 55% accuracy, while the worst performer finished with 21% accuracy—four points worse than the worst performer in the learning phase.

Any change to the design implemented to address the shortcomings of Experiment 1 would also facilitate better learning for Experiment 2. Put simply, it is possible for participants to learn in this task, and likely that they will behave in interesting ways when they do. Examining testing patterns within the context of hypothesis generation and time pressure retains potential to inform theory development for hypothesis testing.

3.4 Experiment 3 - Metacognition and Terminating Testing Behavior

Although there is a body of work evaluating the effects of perceptual and cognitive fluency on decision-making (for a review see Schwarz, 2004), virtually no work has investigated the influence of explicit self-assessment in behavioral decision-making tasks. The one exception is work in overconfidence, but the confidence judgments typically concern pre-existing knowledge and there is no experimentally-controlled learning component (c.f., Dougherty, 2001; Rich & Gureckis, under review). Experiment 3 was designed to exploit this gap in the literature by investigating the influence of metacognitive

assessment on test selection and search termination decisions. Specifically, the MDG paradigm was outfitted with intra-trial self-assessment prompts to track changes in confidence for the sake of evaluating its relation to test selection, search termination, and retroactive confidence judgments.

Of particular interest is the relation between judgment-of-knowing (JOK) estimates and search behavior, as there is a scarce amount of research investigating the metacognitive mechanisms that govern data acquisition strategies. Only recently have researchers begun to elucidate rules for terminating search in memory for cued recall tasks (Dougherty & Harbison, 2007; Harbison, Dougherty, Davelaar, & Fayyad, 2009; Hills et al., 2012; Miller, Weidemann, & Kahana, 2012). In a related literature, self-assessment research has shown that exit latencies are empirically tied to JOK estimates for general knowledge, such that low JOKs come quicker than relatively higher judgments (Glucksbert & MCCloskey, 1981; Klin, Guzman, & Levine, 1997; Kolers & Palef, 1976; Singer, 1984). The cross-section of self-evaluative judgments and search presents a rich area of research considering few studies have directly investigated the influence of metacognitive mechanisms on search behavior despite how frequently researchers invoke them in the literature (Dougherty & Harbison, 2007; Ficic & Buckman, 2015; Harbison, Dougherty, Davelaar, & Fayyad, 2009; Illingworth & Thomas, 2015).

The results of Experiment 3 were intended to evaluate the probabilistic stopping rule postulated in the hypothesis-driven valuation model (Harbison, Dougherty, Davelaar, & Fayyad, 2009). Specifically, the stopping rule predicts that termination decisions depend on belief states and vary with respect to the information gleaned from test outcomes. As beliefs give rise to expectations regarding the outcome of available tests, they facilitate

prediction of the expected value of continued search. When a decision-maker's expectations do not exceed their threshold for warranting more testing, search would be terminated and judgments submitted. Computational modeling was carried out to evaluate this prediction of the model.

Experiment 3 was designed as a transitional study with the purpose of migrating my work on hypothesis testing towards metacognitive processing by testing a general hypothesis motivated by threshold stopping rules: That belief states, as measured by JOK estimates, would relate to decisions to terminate hypothesis testing.

3.4.1 Method.

Undergraduate students enrolled at the Georgia Institute of Technology were recruited to participate in this study via an online experiment management system (SONA Systems). In total, 36 participants completed the experiment. All participants received partial course credit for their involvement in the study.

The learning phase of Experiment 3 was completed over 24 blocks of 16 trials, for a total of 384 learning trials. Otherwise, the learning component of Experiment 3 was identical to Experiments 1 and 2, where participants were incentivized for accurate responses with points ($1000) that were deposited in a bank that accumulated points over the course of Phase 1. Table 9 lists the environmental ecology that defined cue-hypothesis associations. Unlike previous experiments, all cues were strongly associated with one hypothesis and one slightly weaker alternative. Thus, each Cue should have led participants to favor one hypothesis over all others, but maintain a secondary alternative.

Table 9. Presenting sign ecology for Experiment 3.

	Presenting Symptoms			
	Cue 1	Cue 2	Cue 3	Cue 4
Hyp 1	.50	.10	.10	.30
Hyp 2	.30	.50	.10	.10
Hyp 3	.10	.30	.50	.10
Hyp 4	.10	.10	.30	.50

The test-hypothesis associations reported in Table 10 controlled the presentation of test outcomes. As was the case for the previous two experiments, the tests were designed to map to the cue-hypothesis relations from Table 9. Hypotheses 1 and 2 are strongly associated with Cue 1, for example (see top of Cue 1 column in Table 6); and Test 1 is the best available test to disambiguate those two hypotheses. The tests available in Experiment 3 all possessed the same pattern of outcome probabilities, but they applied to different hypotheses. The diagnostic value for all tests was controlled such that they were equivalent without the presenting symptoms (*diagnosticity* = 2.46)

Table 10. Test outcome ecology for Experiment 3.

		Diagnostic Tests			
		Test1	Test2	Test3	Test4
Hypothesis 1	Outcome 1	.65	.30	.30	.10
	Outcome 2	.25	.40	.40	.25
	Outcome 3	.10	.30	.30	.65
Hypothesis 2	Outcome 1	.10	.65	.30	.30
	Outcome 2	.25	.25	.40	.40
	Outcome 3	.65	.10	.30	.30
Hypothesis 3	Outcome 1	.30	.10	.65	.30
	Outcome 2	.40	.25	.25	.40
	Outcome 3	.30	.65	.10	.30
Hypothesis 4	Outcome 1	.30	.30	.10	.65
	Outcome 2	.40	.40	.25	.25
	Outcome 3	.30	.30	.65	.10

The test phase of Experiment 3 was completed over 4 blocks of 16 trials, for a total of 64 test trials. During the test phase of the experiment, participants selected tests as has been the case in each of the previous instantiations of the MDG, but JOKs were elicited after each testing event. This change in the procedure also forced participants to select at least one test before recording the first JOK. Participants were asked to rate their confidence, from 0 to 100, that they knew the correct disease. Responses were recorded by a sliding scale that would appear in the center of the display; only whole numbers between 0 and 100 could be selected on the scale. Participants were instructed that 100 meant they thought there was a 100% chance that they knew the correct disease, and that 0 meant they thought there was a 0% chance that they knew the correct disease. This prompt appeared following each selection of a medical test. Access to additional medical tests was blocked while participants were tasked with responding to the JOK prompt. After participants submitted their JOK, they were given the option to terminate the trial and submit a diagnosis or select another test. The point at which a diagnosis was submitted was self-paced, such that participants were free to select all or none of the tests available on any given trial. A JOK elicitation always appeared prior to the diagnosis submission and subsequent termination of the trial.

3.4.2 Results.

Learning. A binomial regression analysis regressed Phase 1 accuracy on block, finding that block had no relation to the accuracy of diagnosis in the learning phase (χ^2 (n=36,23) = 25.50, p = 0.33). Moreover, participants were not any more likely to submit a correct diagnosis in block 24 than they were in block 1 (β = 0.19, SE = 0.12, p = 0.19). However, there was a significant linear trend relating block and accuracy (χ^2 (n=36,1) =

6.25, $p = 0.01$). Figure 11 illustrates that the distribution of learning phase performance is closer to chance performance than it had been for the previously reported studies. Mixed learning results imply that knowledge of the task environment varied across participants, and can potentially aid in elucidating performance during Phase 2. Notably, the bottom quadrant of participants was below the chance line.

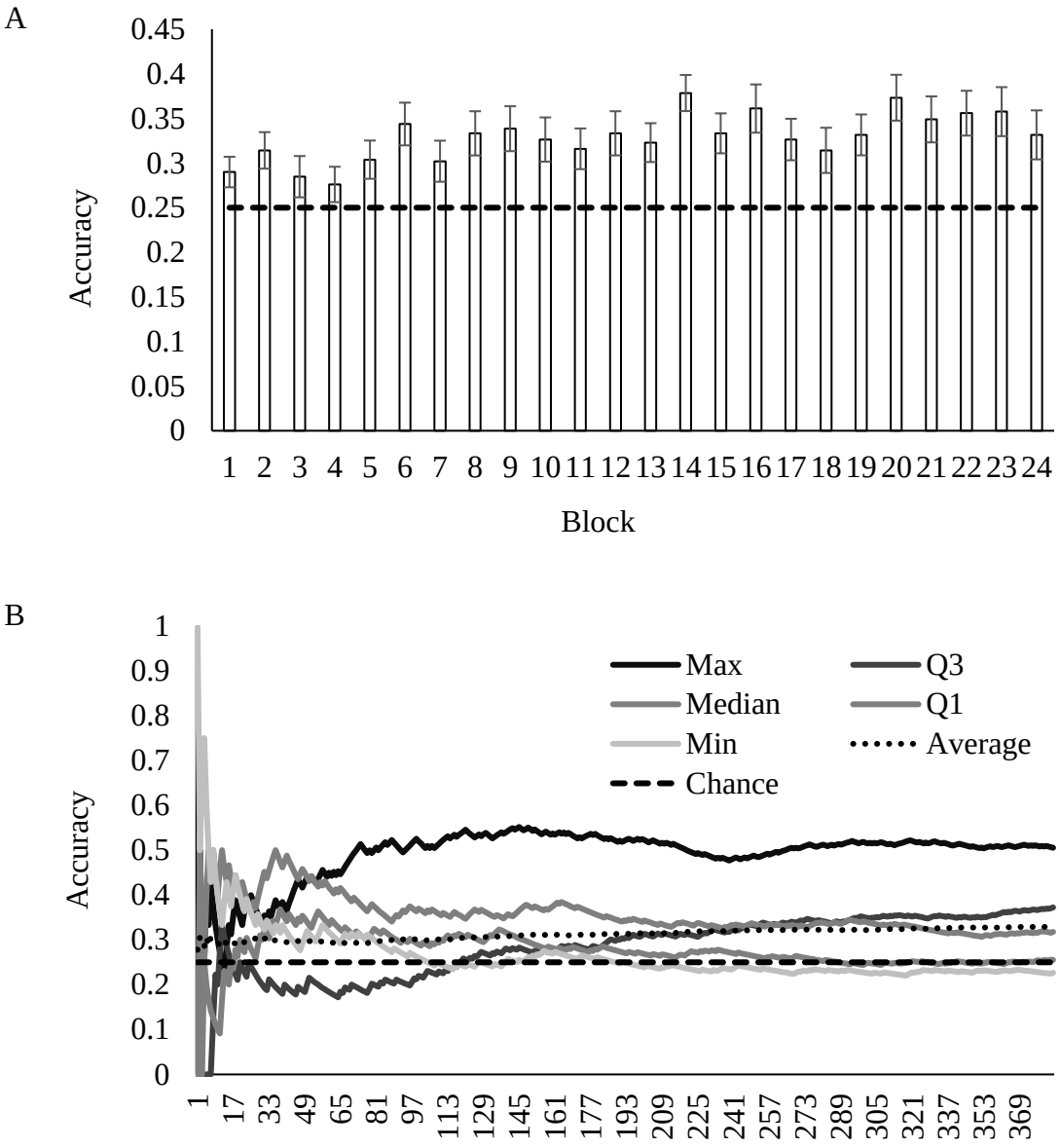


Figure 11. Panel A illustrates Experiment 3 learning phase accuracy broken out by block. The dotted line represents chance performance (25%). Error bars represent standard error. Panel B tracks proportion correct for 5 participants across all trials

of Phase 1, as well as sample average (black, dotted line) and chance performance (black, dashed line). The worst performer scored below chance accuracy, while the best performer scored just above 50%.

Accuracy. A logistic regression evaluated how well accuracy was predicted by block. Block was not found to be predictive of learning phase accuracy (χ^2 (n=36,1) = 0.32, p = 0.57), suggesting that performance did not improve over the course of the test phase. Participants were nearly equally likely (O = 0.96) to submit a correct diagnosis in the final block as they were in the first (β = -0.04, SE = 0.08 , p =.57). Moreover, accuracy in the first block (M = 0.38, SE = 0.02) and the second block (M =0.39, SE = 0.02) were approximately equal to learning phase performance.

Stopping. A multinomial logistic regression analyzed the degree to which presenting cue and participants' first JOK predicted the total number of tests selected. Recall that participants were required to select at least 1 test and respond to 1 JOK elicitation for Experiment 3. Thus, JOK 1 is the only judgment for which all participants have data. JOK 1 was found to be a strong predictor of tests selected (χ^2 (n=36,1) = 19.00, $p < 0.0001$) such that the number of tests selected was reduced as the magnitude of JOK 1 increased (β = -0.04, SE = 0.009, $p < 0.0001$). This result suggests that participants sought more information when their metacognitive self-assessment suggested uncertainty regarding the patient's disease. Neither presenting cue (χ^2 (n=36,3) = 1.14, p = 0.77) nor the interaction between presenting cue and JOK 1 (χ^2 (n=36,3) = 1.36, p = 0.72) predicted total tests selected.

JOK. To assess JOKs over the course of the experimental task, the dataset was parsed by number of tests selected. Four linear regression analyses evaluated the relation between presenting cue and JOK. For three of those analyses, the number of tests selected

prior to the JOK was included in the statistical model to test for changes in JOK magnitude over the course of trials. Presenting cue did not predict JOK magnitudes when participants selected 1 (χ^2 (n=36,3) = 5.45, p = 0.14), 2 (χ^2 (n=36,3) = 1.83, p = 0.61), 3 (χ^2 (n=36,3) = 2.09, p = 0.55), or 4 tests (χ^2 (n=36,3) = 2.04, p = 0.56). Number of tests selected, however, was a strong predictor of JOK magnitude across all analyses (see Figure 12). Tests selected strongly predicted JOK when 2 tests were selected (χ^2 (n=36,1) = 16.39, p < 0.0001), such that the difference between JOKs reported after the second test selection and the first test selection was significantly larger than zero (z = 6.35, p < 0.0001).

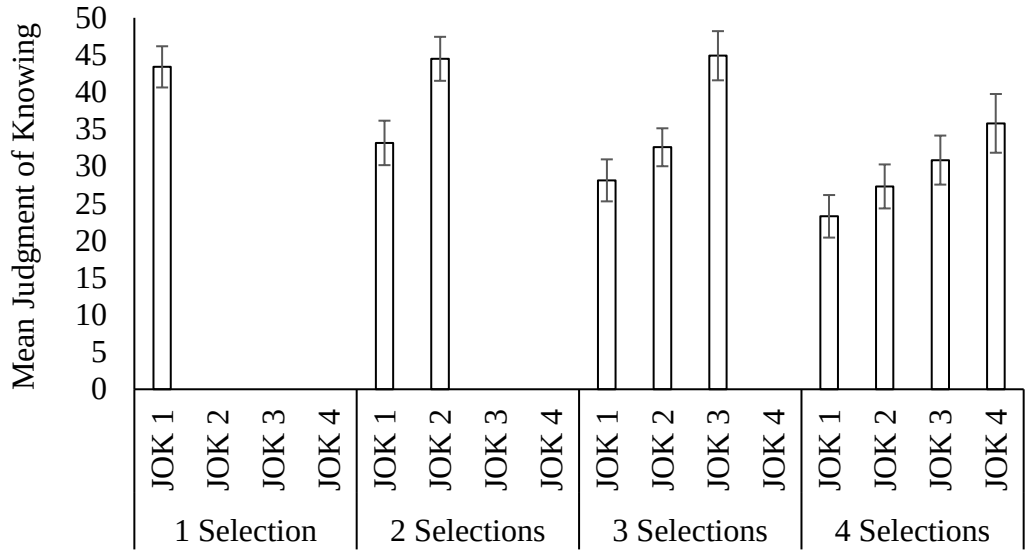


Figure 12. Experiment 3 JOK magnitude across trials. JOKs increased in magnitude across trials, showing increasing confidence in participants' capacity to diagnose patients accurately. Termination occurred after similarly sized JOKs on average when participants selected fewer than the maximum number of test.

Tests selected was predictive of JOK magnitude when three tests were selected (χ^2 (n=36,2) = 13.04, p = 0.0015), such that magnitudes grew substantially which each subsequent test selection. Participants' JOKs were of significantly higher magnitude after selecting the second test compared to the first test selection (z = 4.29, p < 0.0001), as was true when comparing JOK magnitude after selecting the third test to the second test

selection (z = 7.18, p < 0.0001). Tests selected also predicted JOK magnitude when participants selected four tests (χ^2 (n=36,3) = 7.99, p = 0.046). The same pattern persisted under these conditions where JOK magnitude grew after each test selection. JOK magnitude was significantly larger after the second test selection compared to the first test selection (z = 3.73, p < 0.001), larger after the third test selection compared to the second test selection (z = 3.55, p < 0.001), and larger after the fourth test selection compared to the third (z = 3.28, p = 0.0001).

JOK growth suggested that participants experienced an increase in confidence regarding their capacity to diagnose patients correctly while they accumulated data. Moreover, two patterns of interest emerged in the data and are visible in Figure 12. First, the relation between JOK 1 and the number of tests selected is well illustrated, as a marked drop in JOK magnitude can be seen for JOK 1 when the total number of tests selected increases. Second, participants appear to have terminated search after reporting JOKs of similar magnitude for trials when they selected fewer than the maximum number of tests. To explicitly test for differences in final JOK, a regression analysis found that presenting cue (χ^2 (n=36,3) = 1.78, p = 0.62), total tests selected (χ^2 (n=36,3) = 7.99, p = 0.046), and the interaction of these two variables (χ^2 (n=36,9) = 7.99, p = 0.046) failed to predict the final JOK elicited from participants prior to search termination. Figure 13 illustrates this result.

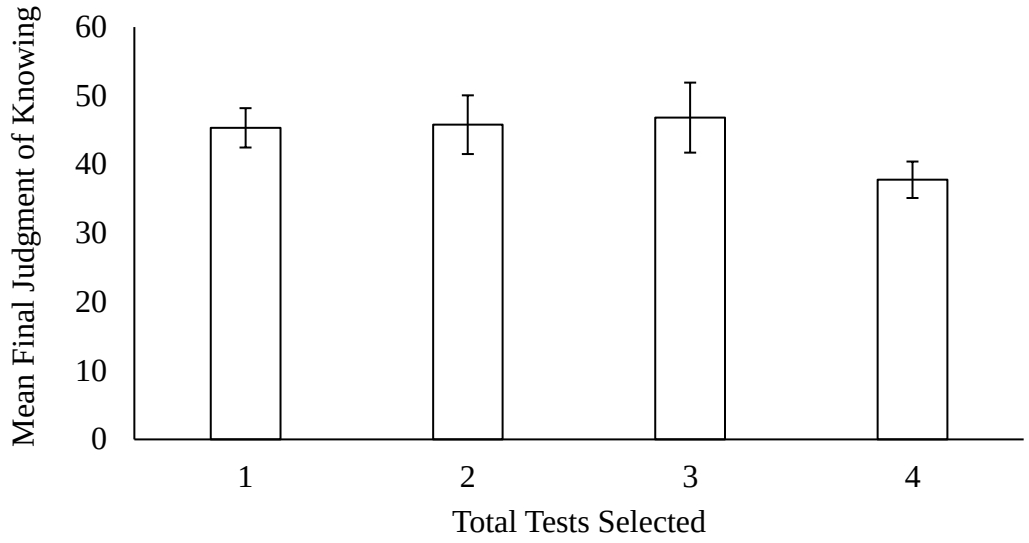


Figure 13. Experiment 3 final JOK magnitude. The figure illustrates the JOK elicited when participants terminated testing, broken out by total tests selected. Analyses failed to detect differences across the four patterns of test selection. Error bars represent standard error.

Test selection. Four binomial logistic regressions assessed selection of each of the four available tests within the context of the varied presenting cues (see Figure 14). Test 3 was predicted by presenting cue (χ^2 (n=36,3) = 9.88, p = 0.019). As expected, Test 3 selection was most closely associated with Cue 3 presentation. Test 3 was selected 1.49 times more often following presentation of Cue 3 relative to Cue 2 (β = 0.40, SE = 0.19, p = 0.029), 1.62 times more often relative to Cue 1 (β = 0.48, SE = 0.21, p = 0.023), and 1.75 times more often relative to Cue 4 (β = 0.56, SE = 0.16, $p < 0.001$). Selection of no other test was predicted by presenting cue, including Test 1 (χ^2 (n=36,3) = 5.84, p = 0.12). Test 2 (χ^2 (n=36,3) = 5.36, p = 0.15) and Test 4 (χ^2 (n=36,3) = 3.44, p = 0.33).

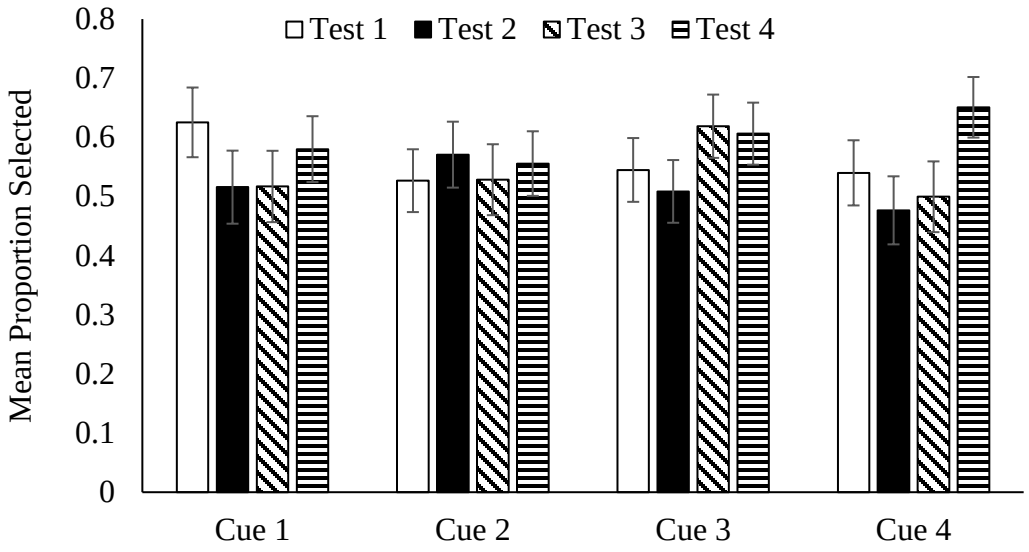


Figure 14. Experiment 3 test selection. Test 3 was the only source of information predicted by presenting cue, as it was found to be most closely associated with Cue 3 presentation. Presenting cue influenced no other tests. Error bars represent standard errors.

The test selection analyses were repeated for the sake of exploring what if anything could become known by including Phase 1 performance in the statistical model. Generally, none of the variables predicted selection for any of the four tests (see Table 11). The one exception was Test 3, for which learning was a significant predictor. Presenting cue, previously been found to predict Test 3 selection, and the interaction term approached but did not reach statistical significance. While the interaction term did not predict Test 3 selection, it may be the case that the relation between Cue 3 presentation and Test 3 selection hinges upon how well the task environment is encoded. This finding must be interpreted, however, while considering that no other test was found to be related to presenting cue.

Table 11. Experiment 3 test selection analyses including Phase 1 learning.

		n	*df*	χ^2	*p*
Test 1	Presenting Cue	36	3	2.31	0.51
	Learning	36	1	0.43	0.51
	Interaction	36	3	2.79	0.42
Test 2	Presenting Cue	36	3	2.78	0.42
	Learning	36	1	0.13	0.71
	Interaction	36	3	3.51	0.32
Test 3	Presenting Cue	36	3	6.37	0.10
	Learning	36	1	4.16	0.04
	Interaction	36	3	6.96	0.07
Test 4	Presenting Cue	36	3	6.49	0.09
	Learning	36	1	0.88	0.34
	Interaction	36	3	6.42	0.09

Test preference. To delve further into test selection behavior, selection data was transformed to reflect the order in which tests were selected. The order recorded during completion of the task was reverse scored such that a test selected first was scored a 4 and a test selected fourth was scored a 1. Any test that was not exploited during a trial received a score of 0. Four multinomial logistic regressions were run to evaluate whether or not presenting cue predicted test preference (see Figure 15). As was the case for test selection, Test 3 preference was predicted by presenting cue (χ^2 (n=36,3) = 9.93, p = 0.019). Test 3 was found to be most associated with Cue 3, as it was 1.51 times more likely to be selected earlier in the testing sequence after Cue 3 than Cue 2 (β = 0.42, SE = 0.18, p = 0.019), 1.78 times more likely to be selected early after Cue 3 than Cue 1 (β = 0.58, SE = 0.24, p = 0.015), and 1.88 times more likely to be selected early after Cue 3 than Cue 4 (β = 0.63, SE = 0.18, $p < 0.001$). Presenting cue did not predict Test 1 preference (χ^2 (n=36,3) = 5.36,

$p = 0.15$), Test 2 preference (χ^2 (n=36,3) = 4.50, $p = 0.21$), or Test 4 preference (χ^2 (n=36,3) = 4.21, $p = 0.24$).

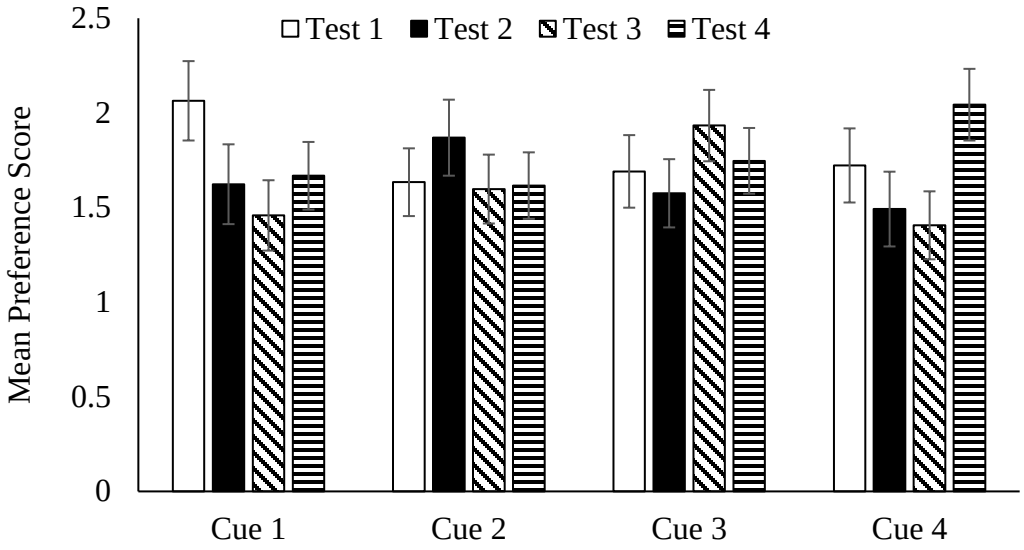


Figure 15. Experiment 3 test preference. Test 3 was the only medical test for which the order it was selected was predicted by presenting cue, as it was found to be most closely associated with Cue 3 presentation. Presenting cue influenced no other tests. Error bars represent standard errors.

A second array of multinomial logistic regressions included learning phase performance to explore if testing order could be elucidated by considering how well participants understood the task environment. The addition of learning in the statistical models reported in Table 12 did not substantially alter the result, as presenting cue did not predict preference scores for any available test. This result marks a change for Test 3, which was initially found to have a relation with presenting cue but was predicted by Phase 1 performance instead (χ^2 (n=36,1) = 4.85, $p = 0.027$).

Table 12. Experiment 3 test preference analyses including Phase 1 learning.

		n	*df*	χ^2	*p*
Test 1	Presenting Cue	36	3	4.61	0.20
	Learning	36	1	0.80	0.37
	Interaction	36	3	4.83	0.18
Test 2	Presenting Cue	36	3	1.39	0.71
	Learning	36	1	0.21	0.65
	Interaction	36	3	1.94	0.58
Test 3	Presenting Cue	36	3	6.83	0.08
	Learning	36	1	4.85	0.03
	Interaction	36	3	6.96	0.07
Test 4	Presenting Cue	36	3	4.25	0.24
	Learning	36	1	0.52	0.47
	Interaction	36	3	4.21	0.24

Model fitting. Although the statistical models evaluating the testing behavior recorded in Experiment 3 found no link to presenting cue, unlike Experiments 1 and 2 there was indirect evidence that participants were responding to the information acquiring from testing events. Thus, simulations were conducted to explore the stopping behavior participants exhibited in Experiment 3 and its link to belief revision.

Test selection and belief revision were modeled consistent with the methodology outlined for Experiment 1 and 2 simulations with the exception that softmax was not parameterized in Experiment 3 simulations. Diagnosticity was computed using Equation 17 and belief revision was estimated using Equation 18. The best available test was always selected and posterior belief distributions were always perfectly calibrated to the newly acquired datum. The stopping rule was implemented as a probabilistic function, where the probability of termination search was computed using Equation 20 (Harbison, Dougherty, Davelaar, & Fayyad, 2009).

$$P(t) = \frac{1}{1 + e^{-g(X-\theta)}} \quad (20)$$

$$X = EV_{Max} - EV_{Current} \quad (21)$$

The simulation iterated through several values of the gain parameter (g) and the theta parameter (θ). Theta represented the decision threshold, which took the form of the minimum value gained by continued search in this simulation. Gain represented sensitivity to the difference between the change in expected value brought about by continued search and the threshold. That difference (X) was estimated using Equation 21, where the expected value of the system's current state of knowledge (i.e., the probability of a correct response multiplied by the payoff for a correct response) was subtracted from the expected value anticipated for a future state brought about by continued search. A random model was included in the simulation. The random model selected tests at random and terminated search such that the probability of termination was 50% after each test selection.

Table 13. Experiment 3 aggregate fit statistics (*BIC*) for all possible parameter combinations.

		Theta (θ)					
		50	100	150	200	250	Random
Gain (g)	0.6	884.08	854.00	509.44	320.00	42.20	0.00
	0.8	526.65	812.75	733.38	411.75	723.81	
	1.0	1058.86	877.01	630.87	346.66	-352.04	
	1.2	977.99	780.04	629.94	459.48	-457.844	
	1.4	1312.46	516.93	516.96	576.76	-379.21	

Likelihoods were computed for each trial of participant behavior across all parameter combinations and the random model. Specifically, the preference scores for each test selection, total number of tests, and final JOK were used to evaluate participant

performance. Likelihoods were aggregated for each participant and used to compute G^2 statistics to fit participants to each parameter combination and the random model. The G^2 for each parameter combination was compared against the random model and corrected for parameters (2) to compute a Bayes Information Criterion (*BIC*) for each participant. Table 10 lists aggregate *BIC* statistics for all parameter combinations with respect to the random model, where more negative values represent better fit. *BIC*s larger than zero indicate that the random model was a better fit. The best fit model was the stopping rule that implemented a gain value at 1.2 and a theta value at 250 (see Table 13). This suggested that participants exhibited slightly heightened sensitivity to potential gains in value from continued search and established a threshold such that they expected novel information to improve their confidence by 25%. The aggregate fit statistics are fortified by the fact that nearly a third of participants (31%) were best fit by those parameter combinations (see Table 14).

Table 14. Experiment 3 proportion of participants fitting to parameter combinations.

		Theta (θ)					
		50	100	150	200	250	Random
Gain (g)	0.6	0.00	0.00	0.00	0.00	0.00	0.19
	0.8	0.00	0.00	0.00	0.00	0.00	
	1.0	0.00	0.00	0.00	0.00	0.33	
	1.2	0.00	0.00	0.00	0.00	0.31	
	1.4	0.00	0.00	0.00	0.00	0.17	

Over 80% of participants were best fit by the threshold value of 250. This suggests that participants across the board expected that participants exhibited a high threshold for information. Participants also exhibited high sensitivity to change, as demonstrated by the fits to gain values of 1.0, 1.2, and 1.4. A model that selected tests at random and terminated

search at random best fit nearly 20% of participants, which marks a relative reduction in the number of participants fitting a random model when considering performance in Experiments 1 and 2.

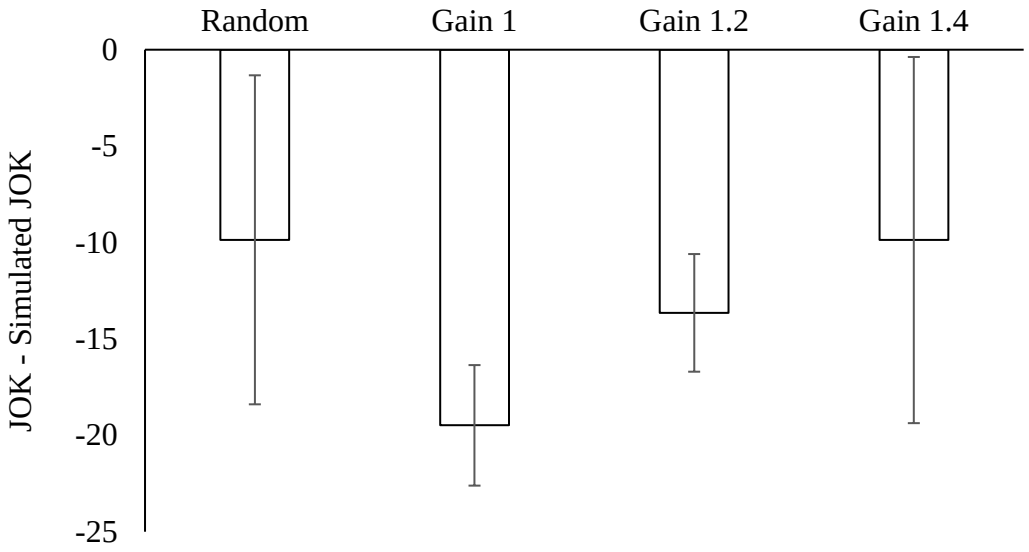


Figure 16. Experiment 3 JOK difference. No relation was found between simulated and observed JOK. This remained true when the data were parsed out by best fit model. Error bars represent standard errors.

Participant JOKs were explicitly evaluated with respect to simulated posterior asymmetry by matching each human trial to model behavior. Specifically, model asymmetry was operationalized as the maximum belief in a disease state exhibited by the best fit parameters for each participant. Behavior was matched on the content of the case. Simulated JOK was regressed over participant JOKs, revealing that there was no relation between the two (χ^2 (n=36,1) = 1.55, p = 0.21). The analysis was conducted once more, parsing out the data by best fit model. Once again, no relation was unveiled between JOK and simulated JOK for the Random model (χ^2 (n=7,1) = 1.99, p = 0.16), the Gain=1 model (χ^2 (n=12,1) = 0.39, p = 0.53), the Gain=1.2 model (χ^2 (n=11,1) = 0.23, p = 0.63), or the

Gain=1.4 model (χ^2 (n=6,1) = 3.09, p = 0.08). Figure XX illustrates the difference between observed JOKs and simulated JOKs by the best fit model.

3.4.3 Discussion

Experiment 3 was the first study to my knowledge to both elicit metacognitive self-assessment in an information acquisition task and report a relation between that self-assessment and the amount of information foraging recorded in the experiment. Specifically, participants generally increased the number of tests they exploited as their initial self-assessments decreased—a finding generally supportive of the probabilistic rule posited in the hypothesis-driven valuation model. This result suggests that participants experienced some sensitivity to the information available in the environment despite the poor accuracy performance reported in the results section. The general rise in JOK over the course of trials wherein multiple tests were selected also suggested that participants were reacting to newly acquired information. That increased confidence judgments followed data acquisition is evidence that participants recognized improvement in their knowledge about a case as its details were revealed—a sign that meaningful data was gleaned from test results.

The results of Experiment 3 also suggest that decisions to terminate search were related to beliefs. Specifically, mean JOKs appeared to approach a similar plateau before search was terminated, and the mean final JOK did not vary with respect to the number of tests selected by participants. The pattern points to changes in belief brought about by newly acquired data and the subsequent rise in confidence as important contributors to decisions to terminate search. The fit of the simulation results to patterns in the behavioral data supported this conclusion, as evaluating Experiment 3 data on a per-trial basis showed

signs that posterior belief and anticipated outcomes were directly related to decisions to terminate search. A large majority of participants were best fit by the probabilistic termination rule when the threshold parameter equaled 250—a 25% improvement in confidence. That the parameter was indicative of a high threshold for increases in expected value suggests that participants were somewhat conservative in their decisions to stop testing. Specifically, they expected that the effort expended to acquire new information would result in a relatively large improvement in their prospects for a correct diagnosis.

One caveat to consider when contemplating the result linking beliefs to decisions to terminate search is that it was not coupled with evidence of hypothesis-guided testing behavior. Moreover, behavior was modeled with respect to a system that was perfectly sensitive to diagnosticity and perfectly calibrated to novel data. What was in dispute when simulating and fitting Experiment 3 participants was whether decisions to terminate search behavior was informed and related to the knowledge state of participants or determined at random. Although the knowledge state simulated was ideal, it created a stark contrast with the random model and revealed something about the behavior exhibited by participants in this task. The fact that far more participants fit the ideal operator suggests that a majority of the participants were making informed decisions when terminating search.

Additionally, the simulation of the gain parameter's impact on the probability of terminating search could be interpreted as learning. A low gain would suggest poor sensitivity to difference between the expected value of future states and an idiosyncratic threshold and, potentially, poor understanding of the task environment. In the simulation, gain trades off with threshold such that a low gain could diminish the liberal strategy suggested by a low threshold, as it would reduce the system's ability to detect large

differences between future states and the threshold. Participants generally fit higher values of gain best, which indicates that participants experienced no deficits as it relates to detecting how the expected value was likely to change in future states of search.

Although Experiment 3 provided data sufficient to explore the nature of the probabilistic stopping rule put forth in the hypothesis-driven valuation model, participants did now show signs of sensitivity to presenting cue when selecting tests. This experiment was the last reported in this document to implement the presenting cue manipulation within the MDG paradigm. Summation of these three studies finds no support for the hypothesis-guided testing hypothesis. Experiment 3 did include a significant predictor of Test 3, as presentation of Cue 3 resulted in a higher frequency of Test 3 selection and a higher preference score for Test 3. The most likely explanation for this result is a spurious significant test given that this was the only of 12 opportunities to detect any Cue-Test relation.

As was the case for Experiments 1 and 2, participants performed poorly in the MDG task implemented in Experiment 3. Learning phase performance averaged around 33% accuracy. Moreover, the participant representing the 1st quartile completed the learning phase performing at 25% accuracy, meaning a quarter of participants were performing at or below chance—the worst of any experiment reported in this document. This poor showing in the learning phase contrasts harshly with the fact that Experiment 3 possessed the most consistent manipulation of diagnosticity across tests, and the tests possessed the highest diagnosticity of any set of tests reported thus far (2.46). The relative ease with which this statistical structure could be learned may have come across in test phase performance, as a total of 8 participants scored above 52% (*Max* = 61%) after the best

performer in the learning phase finished at 51% accuracy. In fact, the accuracy distribution for the test phase was positively skewed if not bimodal, showcasing that a number of participants were able to perform well on the task.

The inconsistency with which participants performed in Experiment 3 is yet another instance in which participant engagement may have been a limitation of this study. All but two participants completed test trials in less than 1 second (a whole second shorter than the time pressure manipulation implemented in Experiment 2). The other two participants were recorded as having minimum response times of 1.001 and 1.003 seconds. Bear in mind that information was not immediately available to participants during test phase trials. The nature of the game was such that participants were tasked with explicitly selecting the information they deemed necessary to complete the task—a time-consuming endeavor for a mindful decision-maker. As has been discussed, a number of studies previously found diagnosticity to be an adequate information metric such that participants appear capable of encoding and acting on the information presented in similar tasks. (Illingworth & Thomas, 2015; Lange, Thomas, & Dougherty, 2010; Nelson, 2005; Nelson et al., 2010). Although I previously proposed changes to the ecological structure of the task to facilitate better learning, not much can be done to intervene with participant performance when such little time and effort is spent on challenging tasks such as the MDG paradigm.

3.5 Experiment 4 - General Search Tradeoffs and Hypothesis Testing

Theoretical accounts of information search and hypothesis testing have grown in complexity, such that contemporary models of these phenomena address the utility of potential answers in addition to the nature of the queries people formulate (e.g., confirming, falsifying) prior to making decisions (Johnson-Laird & Byrne, 1991; Nelson, 2005; Nelson,

McKenzie, Cottrell, & Sejnowski, 2010). Moreover, attention has been drawn to the importance of subjective judgments of information utility (Manktelow & Over, 1990)—a perspective that was an important precursor to the emergence of memory-based accounts of utility estimation (Johnson, Haubl, & Keinan, 2007; Weber, Johnson, Milch, & Chang, 2007) and process model accounts of hypothesis testing behavior (Dougherty, Thomas, & Lange, 2010; Thomas, Dougherty, Sprenger, & Harbison, 2008).

In spite of advancements in the rigor with which search and hypothesis testing are conceptualized and empirically examined, understanding of the nuanced nature of hypothesis testing is limited. Most notably, comprehensive theories accounting for decisions to terminate hypothesis testing have yet to be developed (c.f., Ficic & Buckman, 2015). Decisions to terminate testing are critical to the understanding of information acquisition, as the amount a decision-maker searches determines the quantity and content of the gathered information as well as the total costs incurred searching (Illingworth & Thomas, 2015). Moreover, the data ultimately accessed by decision-makers is hypothesized to influence the knowledge and beliefs upon which actions are based (Lange, Thomas, Buttaccio, Illingworth, & Davelaar, 2013; Melhorn, Taatgen, Lebiere, & Krems, 2011). One reason for limited theoretical advancement may be a failure to reconcile the current understanding of hypothesis testing with broader research programs investigating information foraging and cognitive search.

Recent developments in cognitive search theory have highlighted parallel findings that emerge across a myriad of research programs investigating search, positing that search behavior shares common strategies that are processed with similar neural structures (Hills, 2006; Hills, Todd, & Goldstone, 2008). Specifically, Hills (2006) argues that these patterns

materialize because of an area-restricted search mechanism that is well suited for environments where the locations of resources are correlated. In other words, search becomes localized to exploit clustered resources, such as the information that could be gleaned from the results of a medical test. Hypothesis testing also shares characteristics with diet selection in addition to area restricted search (Stephens & Krebs, 1986), as selecting information depositories or formulating a test is analogous to exhibiting a preference for a specific class of nutrients due to increased expected gains via increasingly selective consumption (Pirolli & Card, 1999; Winterhalter, 1986).

Few studies have explored hypothesis-testing behavior within a foraging context where the perceived value of a test is conceptualized as a function of its expected utility and the costs associated with acquiring test results. The goal of Experiment 4 was to address this gap in the literature by investigating how factors ubiquitous in applied decision domains—such as costs of gathering information (e.g., time, monetary expenses), risks taken when pursuing unreliable sources of information, and changes in task context (i.e., the framing of outcomes as gains or losses)—influence decisions to terminate data acquisition.

Attempts to elucidate rules for terminating search behavior have been primarily confined to memory retrieval (Dougherty & Harbison, 2007; Harbison, Dougherty, Davelaar, & Fayyad, 2009; Hills, Jones, & Todd, 2012; Metcalfe & Murdock, 1981; Miller, Weidemann, & Kahana, 2012; Raaijmakers & Shiffrin, 1981; c.f., Lejuez et al., 2002; Pleskac, 2008). Models of memory search termination assume that retrieval termination depends on the probability of successful retrieval and the cost of additional attempts to probe memory. These models are often inspired by the animal foraging literature. In

particular, optimal foraging theory (Stephens & Krebs, 1986) assumes that decisions to leave a cluster of resources rely on a cost-benefit assessment, considering both expected benefits (energy gains) and costs (energy expenditures) of continued search. The animal literature has a rich history of investigating contextual factors (e.g., predation risk, energy reserves, mating opportunities) that influence foraging (Lima & Dill, 1990; McNamara & Houston, 1992; Nonacs, 2001)—a practice yet to be fully integrated into the study of information acquisition.

Perhaps the most studied contextual effect in decision-making is the framing effect in which people show a stronger preference for a risky prospect when described as a potential loss than when described as a potential gain (Kahneman & Tversky, 1979; Tversky & Kahneman, 1992). Kahneman & Tversky's Prospect Theory accounts for framing effects by postulating that people are risk-averse in gain contexts, but are risk seeking in loss contexts. To date, framing effects have received little attention in the hypothesis testing literature (c.f., McKenzie, 2004). Mishra and colleagues have investigated the influence of framing on terminating sequential decision-making (Mishra & Fiddick, 2012; Mishra, Gregson, & Lalumiere, 2012), but such studies differ from the scope of the current work given the focus on data acquisition and hypothesis testing. We also explore the influence of two operational definitions of risk within the framing paradigm. Both outcome variance (Sharpe, 1968) and probability of undesirable outcomes (Dror, Busemeyer, & Basola, 1999) have been tied to risk-sensitive behavior; however, the suitability of these manipulations for learning-based experiments has been a topic of some debate (Weber, 1988; Weber & Milliman, 1997).

When deciding to terminate hypothesis testing it is critical for an intelligent agent to know when the costs of further data acquisition are likely to exceed the utility realized from potential information to reduce unnecessary expenditures. In many situations, queries or tests that provide people with richer information are often more costly options. In medicine, for example, the expense of medical procedures has risen dramatically as medical professionals have come to rely on more advanced technologies (Skinner, 2013). Thus, in lieu of acquiring information from sophisticated sources, medical practitioners may opt for procedures that provide less diagnostic information, or reduce the number of tests they run, to avoid costs (Cohen, Jones, Littenberg, & Neuhauser, 1982; Cummings, Frisof, Long, & Hrynkiewich, 1981; Hoey, Eisenberg, Spitzer, & Thomas, 1982).

How people perceive cost is another factor likely to influence decisions to terminate search. It has long been argued that human judgments reflects an inconsistent scaling mechanism, highly subject to contextual influences (Kahneman & Tversky, 1979; Lopes, 1984; Stewart, Chater, & Brown, 2006). Detected across varied literatures such as perception (Rogers, 1941; Sherif, Taub, & Hovland, 1958), social judgment (Herr, 1986), and reinforcement learning (Bower, 1961), contrast effects illustrate how peripheral information bleeds into assessments of focal stimuli. Specific to the current study, the experience of expending costs of varying magnitudes is expected to influence perceptions of cost in the MDG and, subsequently, decisions to terminate testing behavior. Experiment 4 will directly manipulate the order of cost conditions and evaluate how the experience of cost influences testing behavior.

The goal of Experiment 4 was to explore hypothesis-testing behavior within the context of tradeoffs inherent to foraging tasks. To that end, the study afforded detection of

numerous contextual effects studied extensively in the decision literature that may influence testing behavior. Thus, this experiment had the potential to tie hypothesis testing to both the foraging and decision-making literatures and may serve as the foundation upon which more complex relations between generation processes and environmental constraints can be investigated. Experiment 4 investigated the prediction that increasing costs would reduce testing behavior while the framing of the problem would result in a preference for riskier tests under a loss condition.

3.5.1 Method.

Undergraduate students enrolled at the Georgia Institute of Technology were recruited to participate in this study via an online experiment management system (SONA Systems). In total, 112 participants completed the experiment. All participants received partial course credit for their involvement in the study.

The values in Table 15 defined the ecological disease-test outcome relations for Experiment 4. Note that there was no presenting cue ecology for this experiment. Level diagnosticities signify the degree to which specific test outcomes differentiated between the hypotheses, while test diagnosticity represents how well each test differentiated between the hypotheses after collapsing across all possible outcomes. Although the overall diagnosticity of the tests was controlled so as to be nearly identical, some tests were designed to be riskier than others.

Table 15. Environmental ecology for Experiment 4.

	Test 1	Test 2	Test 3	Test4
Level 1	3.54	3.56	3.57	3.28
Level 2	1.00	1.00	1.00	1.00
Level 3	3.74	3.80	3.87	4.01
Diagnosticity	3.21	2.91	2.58	2.21
D. Variance	2.33	2.41	2.48	2.46

Risk was operationally defined as the probability with which an undesired event will occur when taking a particular action (e.g., Dror, Busemeyer, & Basola, 1999). Specifically, I defined risk within the MDG as the probability of sampling a useless or non-diagnostic test outcome—information that disallows changes in a decision-maker's beliefs. Such information is undesirable because it costs resources but can neither confirm nor contradict the beliefs of the decision-maker. The tests listed in Table 8 were ordered by increasing riskiness, where the probability that Level 2 (the non-diagnostic outcome) was obtained was lowest given Test 1 (20%) and highest given Test 4 (50%).

The learning phase for Experiment 4 was completed over 8 blocks of 30 trials, resulting in a total of 240 learning trials. Learning trials were completed in a manner consistent with the methodology described for the previously reported experiments. The one deviation from the previous procedure was an instruction that the points earned during the learning phase would be sacrificed to acquire information during the test phase of the experiment.

The second phase of Experiment 4 employed a 3 (costs) x 2 (frame) mixed design. The test phase was completed over 6 blocks of 30 trials, resulting in a total of 180 test trials. Costs were manipulated within-subject across 3 levels—None, Moderate, and High. Cost conditions were experienced in a randomized order, and each condition was

completed over 2 consecutive blocks. Participants paid $0 per test in the None condition, $100 per test in the Moderate condition, and $200 per test in the High condition. Prices appeared on the display just below the test label. When a mouse-click was registered within the widget, the price of the test was deducted from the participant's bank and the outcome of the test appeared within the circular widget. The number of tests viewed was left to the participant's discretion, where termination of search (i.e., submission of their diagnosis) could occur after viewing between none and all of the test outcomes.

Frame was manipulated between-subjects. Regardless of frame, a correct diagnosis resulted in a net gain of $1000, while an incorrect diagnosis resulted in a net gain of $250 to the participant's bank. Each trial began with $1000 appearing in a temporary holding account displayed on the monitor. In the gain condition, the payoff for a correct answer was described as acquiring $1000, while the payoff for an incorrect answer was described as acquiring $250 from the temporary account. In the loss condition, the payoff for a correct answer was described as losing $0, while the payoff for an incorrect answer was described as losing $750 of the points available in the temporary account. This feedback was provided after the completion of each trial.

3.5.2 Results.

Learning. A binomial logistic regression evaluated participant learning over Phase 1 of the experiment. Block approach but did not reach significance as a predictor of Phase 1 accuracy (χ^2 (n=112,7) = 12.54, p = 0.084). However, participants were found to be 1.19 times more likely to issue a correct diagnosis in the final block of Phase 1 than they were in the first block (β = 0.19, SE = 0.12, p = 0.19). There was also a significant linear trend linking block and accuracy (χ^2 (n=112,1) = 7.05, p = 0.008). Taken together, these results

suggest that participants generally exhibited learning of the task environment by the end of Phase 1. Figure 17 illustrates the positive trend in performance across blocks of learning trials and individual participants across all trials.

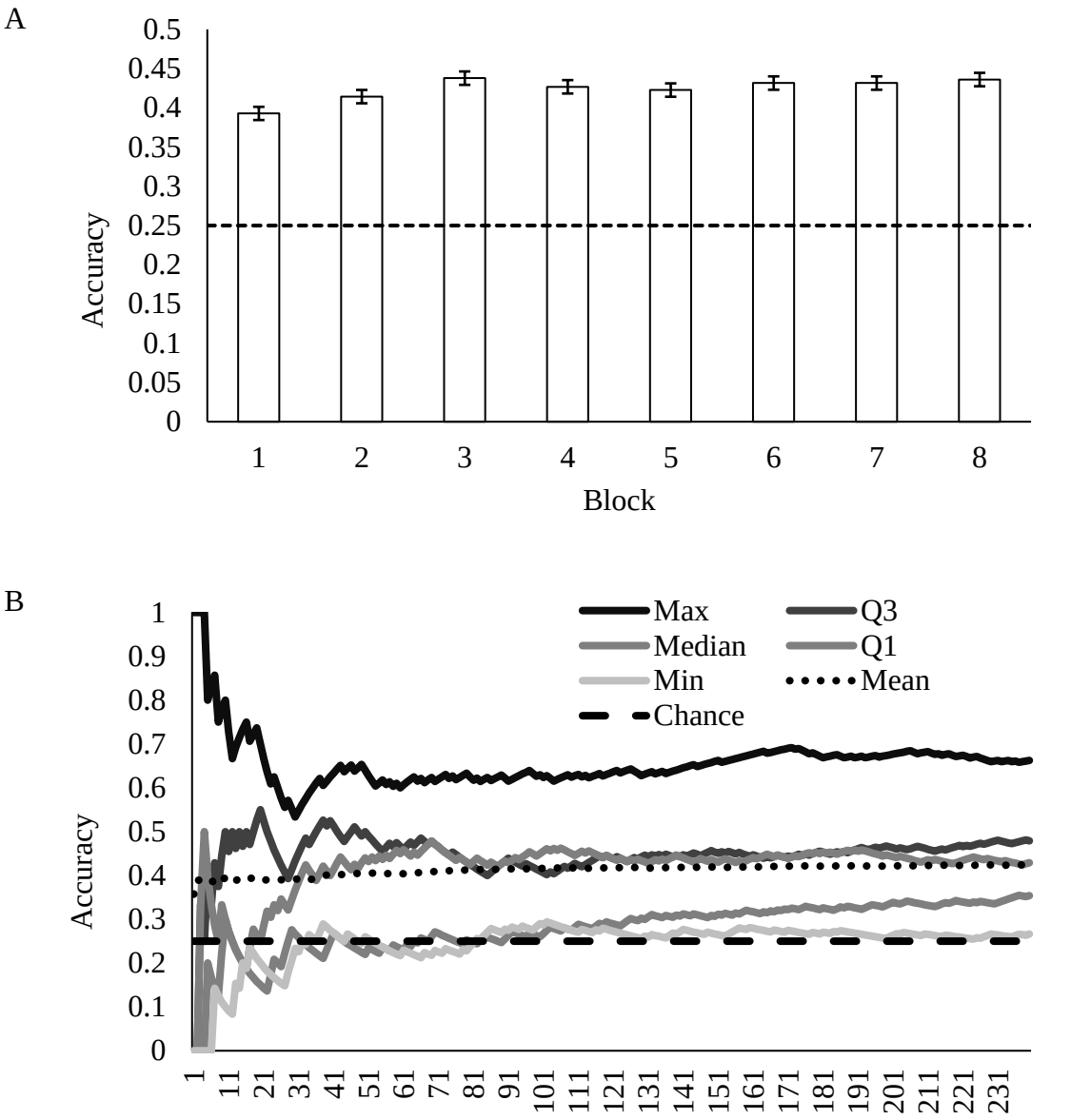


Figure 17. Experiment 4 learning. Panel A illustrates Experiment 4 learning phase accuracy broken out by block. The dotted line represents chance performance (25%). Error bars represent standard error. Panel B tracks proportion correct for 5 participants across all trials of Phase 1, as well as sample average (black, dotted line) and chance performance (black, dashed line). The worst performer performed just above chance accuracy, while the best performer was correct for well over 60% of trials.

Accuracy. A logistic regression evaluated how well accuracy was predicted by block. Block was not found to be predictive of learning phase accuracy (χ^2 (n=112,5) = 2.78, p = 0.73), suggesting that performance did not improve over the course of the test phase. Participants were nearly equally likely (O = 0.94) to submit a correct diagnosis in the final block as they were in the first (β = -0.06, SE = 0.05 , p =.26). Moreover, accuracy in the first block (M = 0.45, SE = 0.01) and the last block (M =0.44, SE = 0.01) were approximately equal to learning phase performance.

Stopping. A multinomial logistic regression analyzed the total number of tests selected across frame, cost, and first cost conditions (i.e., the cost condition the participant experienced in their first block of Phase 2). Cost was found to be a significant predictor of total test selection (χ^2 (n=112,2) = 47.85, p < 0.0001). However, this relation was superseded by a cost by first cost interaction that significantly predicted total tests (χ^2 (n=112,4) = 21.84, p < 0.001). The nature of this interaction can best be understood by comparing the $100 cost conditions when either $0 or $200 is experienced first. Participants were .82 times as likely to select more tests in the $100 condition after experiencing the $200 condition first relative to those who experienced the $0 condition first ($\beta$ = -0.20, SE = 0.34, p = 0.57)—a contrast effect illustrated by Figure 18. The pattern of testing behavior across cost conditions varies as it relates to total tests selected and the cost condition experienced first. Testing is generally abbreviated when $200 is experienced first relative to other condition orders, while the largest impact of changing costs is visible after the $0 condition is experienced first.

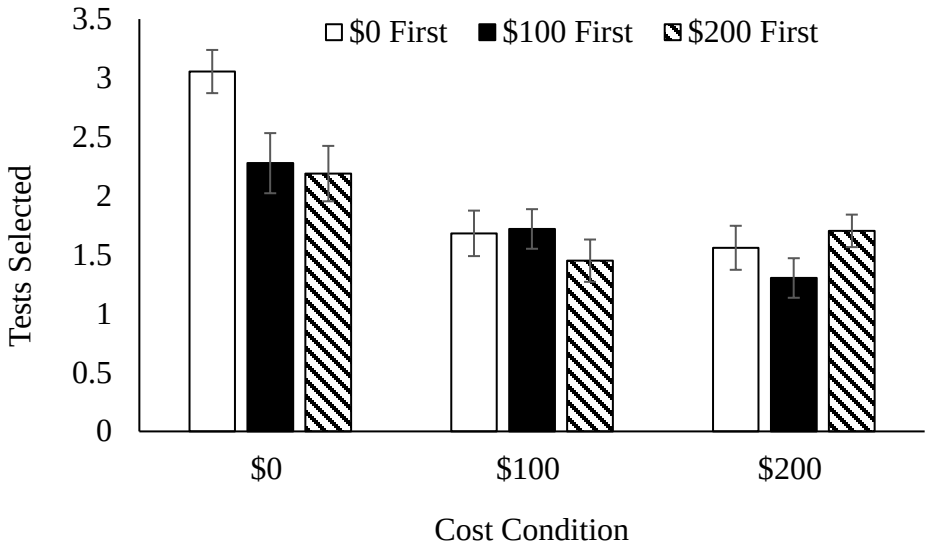


Figure 18. Experiment 4 total testing. The figure illustrates mean total tests selected by cost condition and first condition experienced. A contrast effect emerged where perceptions of cost—driven by experience—are a likely influence of observed termination strategies. Error bars represent standard errors.

Neither first cost condition (χ^2 (n=112,1) = 0.41, p = 0.52) nor frame (χ^2 (n=112,1) = 0.41, p = 0.52) were found to predict total tests select. Moreover, no interaction term that included frame reached statistical significance. This include the two-way interactions between frame and costs (χ^2 (n=112,2) = 0.61, p = 0.74) and between frame and first cost (χ^2 (n=112,2) = 4.74, p = 0.09), as well as the three-way interaction between frame, costs, and first cost (χ^2 (n=112,4) = 3.92, p = 0.42).

Test selection. Four binomial logistic regressions were run to evaluate the relation between selection of each available test and frame, costs, and first cost. Test 1 selection was predicted by costs (χ^2 (n=112,2) = 31.22, p < 0.0001). However, the influence of costs appears to be dependent on which cost condition is experienced first, as the interaction between costs and first cost was a significant predictor (χ^2 (n=112,4) = 18.56, p = 0.001). The nature of the cost effect was such that participants selected fewer tests as costs increased. Participants were 0.46 times as likely to select Test 1 in the $100 condition

relative to the $0 condition ($\beta$ = -0.77, *SE* = 0.15, $p < 0.0001$) and .41 times as likely to select Test 1 in the $200 condition relative to the $0 condition ($\beta$ = -0.90, *SE* = 0.14, $p < 0.0001$).

The general pattern of this interaction appears to map well to the interaction effect that predicted total tests selected. Selection of Test 1 is less likely as costs increase, but the pattern depends on the cost condition experienced first. For example, while Test 1 was generally selected more often when costs were $0, the proportion of trials for which Test 1 was selected differed substantially with respect to the cost condition experienced first. Participants were 0.40 times as likely to select Test 1 in the $0 cost condition after seeing $100 costs first relative to when the $0 condition was first ($\beta$ = -0.92, *SE* = 0.42, $p = 0.029$), and 0.42 times as likely to select Test 1 in the $0 cost condition after seeing $200 costs first compared to seeing the $0 condition first ($\beta$ = -0.77, *SE* = 0.155, $p < 0.0001$). See Figure 19 for an illustration of this pattern of results.

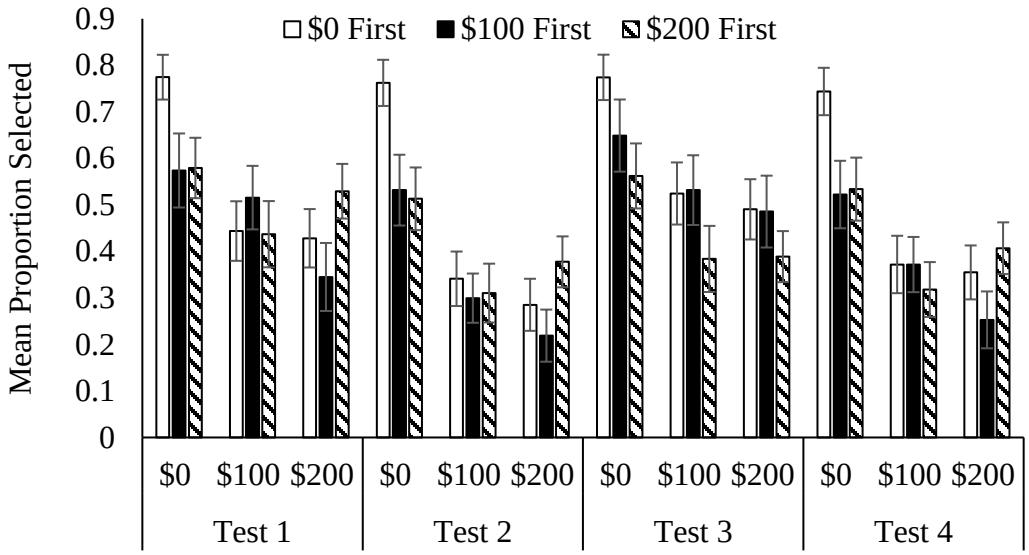


Figure 19. Experiment 4 test selection. Tests selected by cost condition and first condition experienced. This figure illustrates a contrast effect where perceptions of cost—driven by experience—are a likely influence of observed termination strategies. Error bars represent standard errors.

Cost condition also predicted Test 2 selection (χ^2 (n=112,2) = 41.56, $p < 0.0001$). Test 2 selection diminished when costs increased, where participants were 0.45 times as likely to select Test 2 as costs increased to $100 ($\beta$ = -0.79, SE = 0.16, $p < 0.0001$) and 0.35 times as likely to select Test 3 as costs increased to $200 ($\beta$ = -0.92, SE = 0.16, $p < 0.0001$). That influence, however, is modified by first cost condition as the interactive term including these two variables also reached statistical significance as a predictor of Test 2 selection (χ^2 (n=112,4) = 16.19, p = 0.0028). Once again, Test 2 selection in the $0 cost condition showcased differences that resulted from manipulating the order in which costs were experienced. Test 2 was 0.49 times as likely to be selected in the $0 condition after seeing the $100 condition first relative to the $0 first condition ($\beta$ = -0.70, SE = 0.27, p = 0.0094), and 0.37 times as likely to be selected in the $0 condition after seeing the $200 condition first relative to the $0 first condition ($\beta$ = -0.98, SE = 0.28, p = 0.0005).

Test 3 selection was predicted by costs (χ^2 (n=112,2) = 32.30, $p < 0.0001$). The cost pattern remained consistent for Test 3 such that there was a reduction in the rate at which it was selected as costs increased. Participants were 2.12 times as likely to select test 3 in the $0 condition relative to increased cost conditions (β = 0.74, SE = 0.18, $p < 0.0001$).

Cost condition predicted Test 4 selection (χ^2 (n=112,2) = 36.18, $p < 0.0001$). Again, Test 4 selection was reduced as costs increased. Participants were 0.35 times as likely to select Test 4 as costs increased from $0 to $100 ($\beta$ = -0.62, SE = 0.15, $p < 0.0001$) and 0.32 times as likely to select Test 4 as costs increased from $0 to $200 ($\beta$ = -0.70, SE = 0.16, $p < 0.0001$). The interaction between cost and first condition was also a significant predictor of Test 4 selection (χ^2 (n=112,2) = 32.30, $p < 0.0001$). Focusing on the $0 cost condition

to illustrate the different patterns of Test 4 selection resulting for the order manipulating, Test 4 was selected at a far lower rate when $100 and $200 cost conditions were experienced first. Specifically. Participants were 0.56 times as likely to select Test 4 in the $0 cost condition when the $100 cost condition was experienced first instead of the $0 condition ($\beta$ = -0.59, *SE* = 0.28, p = 0.032) and 0.53 times as likely to select Test 4 in the $0 cost condition when the $200 cost condition was experienced first instead of the $0 condition ($\beta$ = -0.63, *SE* = 0.26, p = 0.014).

All predictors that included frame fell short of predicting the selection of any test. This includes the variable acting alone and in any two-way or three-way interaction (all *p*s >.05).

Test preference. To evaluate the nature of sequential test selection behavior, a transformation was performed on selection data to reflect the order in which tests were selected. The order recorded during completion of the task was reverse scored such that a test selected first was scored a 4 and a test selected fourth was scored a 1. Any test that was not exploited during a trial received a score of 0. Four multinomial logistic regressions evaluated whether or not frame, costs, or first cost condition predicted test preference. Cost predicted Test 1 preference (χ^2 (n=112,2) = 17.98, p = 0.0001), such that Test 1 was 0.21 times as likely to be preferred first as costs increased to $100 ($\beta$ = -1.55, *SE* = 0.21, p < 0.0001) and 0.18 times as likely to be preferred first as costs increased to $200 ($\beta$ = -1.71, *SE* = 0.22, p < 0.0001).

The influence of cost was modified by first cost condition experienced (χ^2 (n=112,4) = 18.01, p = 0.0012). Participants were 0.24 times as likely to prefer Test 1 first in the $0 cost condition after experiencing the $100 condition first 100 (β = -1.40, *SE* =

0.32, $p < 0.0001$) and 0.25 times as likely to prefer Test 1 first in the $0 cost condition after experiencing the $200 condition first 100 (β = -1.37, SE = 0.26, $p < 0.0001$). Figure 20 illustrates test preference for all four medical tests, the patterns for which reflect test selection behavior depicted in Figure 19.

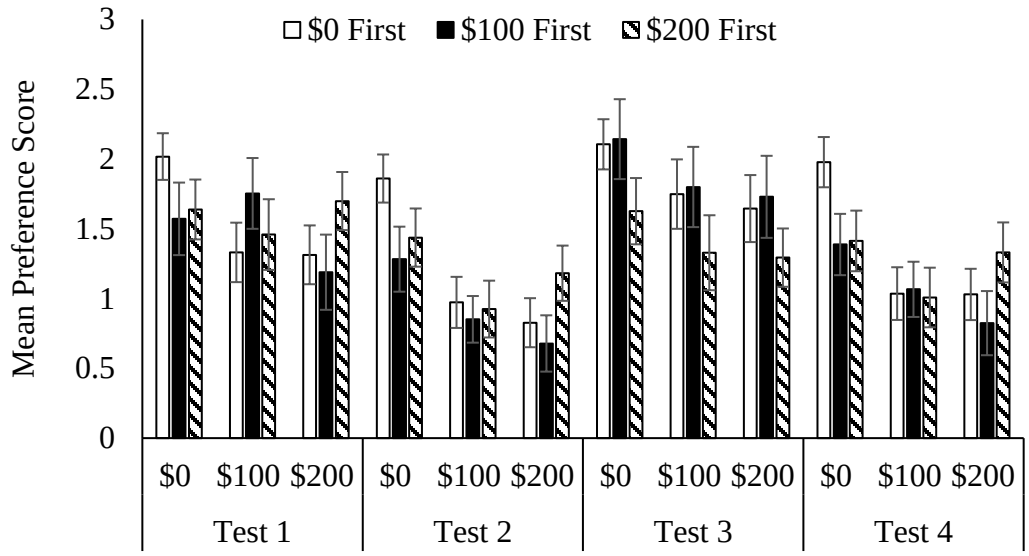


Figure 20. Experiment 4 test preference. The figure illustrates mean test preference scores broken out by cost condition and first condition experienced. Error bars represent standard errors.

Test 2 preference was predicted by cost condition (χ^2 (n=112,2) = 33.29, $p <$ 0.0001), where Test 2 was 0.08 times as likely to be preferred first when costs rose to $100 ($\beta$ = -2.47, SE = 0.23, $p < 0.0001$) and 0.08 times as likely to be preferred first when costs rose to $200 ($\beta$ = -2.54, SE = 0.25, $p < 0.0001$). However, the influence of costs was superceded by a cost by first cost condition interaction (χ^2 (n=112,4) = 11.76, p = 0.019). Within the context of the $0 cost condition, participants were 0.09 times as likely to prefer Test 2 first when the $100 cost condition was experienced first (β = -2.44, SE = 0.32, $p <$ 0.0001) and 0.13 times as likely to prefer Test 2 first when the $200 cost condition was

experienced first (β = -2.05, *SE* = 0.27, $p < 0.0001$). This illustrates the impact of different experience early in the test trials, and the effect of that experience on perceptions of cost.

Cost condition also predicted Test 3 preference (χ^2 (n=112,2) = 19.94, $p < 0.0001$), where participants were less likely to prefer Test 3 as costs increased. For example, 0.33 times as likely to prefer Test 3 first as costs rose to \$100 (β = -1.10, *SE* = 0.21, $p < 0.0001$) and 0.31 times as likely to prefer Test 3 first as costs rose to \$200 (β = -1.19, *SE* = 0.21, $p < 0.0001$).

Test 4 preference was predicted by cost condition (χ^2 (n=112,2) = 28.80, $p < 0.0001$), such that Test 4 was 0.12 times as likely to be preferred first as costs increased to \$100 (β = -2.10, *SE* = 0.24, $p < 0.0001$), and 0.12 times as likely to be preferred first as costs increased to \$200 (β = -2.11, *SE* = 0.26, $p < 0.0001$). Test 4 preference was also predicted by the cost by first cost condition interaction (χ^2 (n=112,4) = 15.61, $p = 0.0036$). Focusing—once more—one the \$0 cost condition, Test 4 was 0.21 times as likely to be preferred first when the \$100 condition was experienced first (β = -1.56, *SE* = 0.30, $p < 0.0001$), and 0.22 times as likely to be preferred first when the \$200 condition was experienced first (β = -1.51, *SE* = 0.30, $p < 0.0001$).

Overall, scoring testing behavior with respect to order reveals nothing beyond what was reporting for selection. Moreover, all predictors that included frame fell short of predicting the preference for any test. This includes the variable acting alone and in any two-way or three-way interaction (all *p*s >.07).

3.5.3 Discussion.

By manipulating risk, costs, and frame Experiment 4 took a first step in elucidating the role that environmental factors play in decisions to terminate search in sequential

hypothesis testing. To my knowledge, this is the first experiment to implement a frame manipulation within the context of hypothesis testing and information search. Traditionally, such paradigms limit choices to one risky option and one safe option, forcing participants to select one or the other. My paradigm differed from the traditional risky-choice manipulation because it permitted access to multiple sources of information during each trial, affording the selection of both risky and safe options.

However, participants failed to exhibit sensitivity to frame in two ways critical to my predictions: Participants showed no preference for riskier sources of information in a loss frame as compared to the gain frame and there was no proclivity for additional testing in the loss frame that would suggest loss aversion. This result differs considerably from those of typical framing manipulations (Kahneman & Tversky, 1979; Tversky & Kahneman, 1992), but extends the finding to scenarios where choice is not constrained to two options.

One possible account that may explain the absence of a framing effect is that the frequency manipulation of risk in this experiment (Dror, Busemeyer, & Basola, 1999) was not effective for conveying risk to participants, necessitating additional work to examine the role of risk in testing behavior more effectively. The best operational definition of risk is a topic of rich debate in the literature (Weber, 1988; Weber & Milliman, 1997). Two alternatives to the frequency risk manipulation are possible candidates to replace it to explore what if any effect can come about by modifying the operational definition of risk. A variance-based manipulation of risk in the MDG paradigm would see larger variability in test outcome diagnosticity for risky tests (e.g., Sharpe, 1968).

For example, exploitation of such a test could equally result in highly informative results or entirely uninformative results. A less risky test would have outcomes of near equal diagnosticity. Weber and her colleagues, however, argue that variance manipulations alone may not capture what it means to perceive risk inherent to some choice. Weber suggests that the history of outcomes that manifest over the course past decisions define for an individual what constitutes a risky choice. Thus, risk is defined not by objective control over outcome variability, but by the subjective variance experienced by an individual.

The findings demonstrated that strategies for acquiring information are sensitive to the costs attached to acquisition, as the participants selected fewer medical tests in cost conditions regardless of the manner in which the task was framed (Illingworth & Thomas, 2015)—a finding consistent with optimal foraging theory (Stephens & Krebs, 1986). We can further conclude that the valuation of information involves some mechanism by which costs impact perceived utility. Moreover, rules governing the termination of data acquisition must be sensitive to that valuation given that search stopped after significantly more foraging when information was free.

Experiment 4 provided further evidence in support of memory-based accounts of utility estimation. Specifically, I observed that prior experience exploiting information depositories in particular environments influenced the nature in which those resources were exploited when the environment changed. In other words, how people perceive the cost of information depends, at least in part, on experience of those costs. In this experiment, participants were willing to increase foraging behavior when they were accustomed to getting less for their money. Alternatively, foraging behavior was truncated when

participants acquired fewer resources in more expensive environments than the conditions to which they had become accustomed.

Process accounts of valuation judgments are lacking in the literature; however, many have explored the importance of memory and memory dynamics for estimating the value of a resource (Gallimore, 1994; Johnson et al., 2007; Weber et al., 2007). Further work is required to parse out the nature of the relation between memory and valuation. For instance, Johnson, Weber, and colleagues have shown that generated uses for an object have a positive relation with perceived value. This work, however, suggests that past costs associated with acquiring a resource directly influence the interpretation of future costs. It appears there are several traces of information available within memory that are active in the valuation process. These findings are important for continued development of valuation research and process accounts for testing and valuation behavior.

In conclusion, Experiment 4 ties information depository preference and search termination to ecological factors such as the cost of data acquisition and individual differences such as experience with cost. These findings demonstrate that hypothesis testing behavior is subject to manipulations commonly deployed in the broader context of foraging research, and can be explicated within the context of general cognitive search. More importantly, this study illustrates a strong role for memory processing in judgments of value. These results illustrate the need to investigate hypothesis testing with respect to cognitive processes known to govern both decision-making and information foraging behavior.

CHAPTER 4. SUMMARY AND IMPLICATIONS

4.1 Summary

Four experiments and three simulations were carried out to evaluate an extension of the HyGene architecture (Thomas, Dougherty, Sprenger, & Harbison, 2008): The hypothesis-guided valuation model. The singular premise of the model was that hypotheses generated to explain a decision-maker's observations also served as the foundation for estimating the value of information depositories with uncertain outcomes. An intelligent agent must also determine when continued information search is unjustified in light of the costs necessary to achieve it relative to the gains expected to be acquired. The studies reported here afforded me observations with which to explore the environmental and cognitive signals that contribute to decisions to terminate search. Specifically, the stopping rule implemented in the hypothesis-guided valuation model was fit to participant data, evaluating the degree to which participants were sensitive to the expected value of information depositories and how belief was related to termination decisions.

The experiments tested the propositions of the model by addressing three research questions intended to promote three distinctive lines of inquiry into the cognitive mechanisms underlying hypothesis testing behavior: the role for generation in test selection, the role for belief if termination, and how search operates in costly environments. The first queried the relation between hypothesis generation and patterns of test selection. This motivated two experiments that cast hypothesis testing within an empirical context typically formulated to study the hypothesis generation process. Experiment 1 was

designed to expand upon previous research that had limited diagnostic testing behavior to comparisons of single- and multiple-hypothesis scenarios, showing that the number of hypotheses considered by a decision-maker determines whether or not diagnostic tests are preferred (Lange, Thomas, & Dougherty, 2010). Specifically, Experiment 1 tested the prediction that the specific set of hypotheses under consideration by a decision-maker should account for test preferences beyond the mere number of hypotheses believed to be in contention—hypothesis-guided testing. The data were examined in search of evidence to support a hypothesis-guided testing heuristic, where the value or preference exhibited for a test changes as a consequence of a decision-maker's beliefs. Experiment 2 was designed to explore hypothesis testing within the context of time constraints in an attempt to expand upon Dougherty and Hunter's (2003a) demonstration that time pressure truncates hypothesis generation that, in turn, causes increased subadditivity.

The second research question motivated an exploration of the role belief states played in decisions to terminate testing behavior. This surprisingly sparse area of research has seldom seen researchers attempt to account for termination decisions within the context of information acquisition (c.f., Ficic & Buckman, 2015). Experiment 3 was designed to initiate inquiry regarding the role metacognitive self-assessment in search termination with the hope of detecting a relation between belief states and termination behavior. The third question gave rise to Experiment 4—an inquiry into the varying environmental factors that may influence test preference and termination. This experiment explored avenues through which hypothesis testing could be linked to the broader fields of choice and decision-making.

4.1.1 General discussion.

At least one element of the experimental design implemented for each of the four experiments was intended to capture behavior that would test the predictions of the hypothesis-guided valuation model. The first three experiments introduced a presenting symptom to the MDG paradigm that was structured to vary the set of hypotheses most strongly considered by decision-makers. The HyGene architecture predicts that people would exhibit a preference for those sources of information perceived to be best suited to disambiguate the uncertain mental state that manifests when more than one hypothesis is considered a candidate explanation for observed data. The statistical relation between disease states and medical tests was such that the diagnostic value of each test would change in response to the presenting symptom and each subsequent test selection. These shifting values are what was hypothesized to provide signals to decision-makers who would determine what test was desired next. More specifically, past experience interacting with information in these environments was expected to provide differential memory strength signals with respect to considered hypotheses—a memory-based mechanism for determining value in an information resource.

The behavior captured in Experiments 1 through 3 showed minimal signs of sensitivity to the statistical environments participants operated within. A number of exploratory analyses were conducted to analyze what if any difference considering learning phase behavior could make in elucidating test phase behavior. These analyses did not yield any learning-dependent relation between presenting symptom and medical test selection. As a whole, the collection of experiments conducted for this dissertation provide no substantial support to the hypothesis-guided valuation model.

The most important contribution of the hypothesis-guided valuation model was its solution to the limitation of current hypothesis testing theory: The inability for any contemporary model to account for the manner in which hypotheses under the consideration of decision-makers govern observable hypothesis testing behavior. As a natural extension of HyGene, this model of hypothesis testing carried with it the primary assumptions and hypothesized mechanisms of previous instantiations of the model. At face value, the findings of this dissertation cast doubt on HyGene's capacity to explain how people choose information depositories in sequential hypothesis testing tasks. Note, however, that participants exhibited exceptionally poor performance throughout the reported studies, and may not have been fully engaged in the experimental tasks.

A number of reasons may account for the poor performance exhibited by participants during the learning phase of these experiments. The easiest to address is the diagnostic value of the sources of information available to decision-makers in the empirical tasks. Across all studies, these values ranged between 2.09 and 3.21. While this confirms that the medical tests were informative, they may have been insufficient to support adequate learning to perform a challenging task well. Alternative, the effort put forth by participants was discussed throughout this document, focusing on the degree to which they exhibited the focal behavior (selection of tests) and the amount of time they spent on task. It was often observed that participants spent less than a second selecting tests and submitting a diagnosis—a feat unlikely to indicate careful consideration of their task.

Previously, the predictions of the HyGene architecture regarding hypothesis testing were supported. Lange, Thomas, and Dougherty (2010) found that pseudo-diagnostic search was a direct consequence of the hypothesis generation process, such that people

engaged in positive search only their search behavior was guided by a single hypothesis. Alternatively, people exhibited sensitivity to diagnostic tests when they generated more than on hypothesis. It was concluded that a memory strength heuristic was operating such that experience with more diagnostic tests was only useful when more than one hypothesis was probing the traces in memory that would bring that value to consciousness. In other words, people became aware of the value of a test when the association in memory between its results and one hypothesis under consideration were strong for but weak for another. Though more complicated than Lange et al.'s (2010) environmental manipulations, the experiments reported here emulated some of the critical features of that earlier work. Additionally, the results of this dissertation conflict with those of previous studies that found people to be sensitive to the diagnostic value of tests (Illingworth & Thomas, 2015; Nelson, 2005; Nelson et al., 2010).

As previously mentioned, hypothesis generation models appear to provide a sound foundation for hypothesis testing theory. These theories, however, have been poorly equipped to address more nuanced behavior, which was the goal of this dissertation. A notable deficit repeatedly discussed in this document is the fact that no theory before the hypothesis-driven valuation model considered how it was that people terminated search behavior. The absence of a stopping rule points to a lack of consideration of value on a broader scale. If an information source's value will inevitably fail to exceed what is deemed worthwhile for exploitation, how was its value determined in the first place? What, if anything, contributes to perceptions of access cost? How do these concepts tradeoff when considering decisions to exploit an information depository or terminate search?

Greater consideration should be afforded to valuation judgments of information depositories, as their features beyond those related to information value may disassociate core cognitive functioning (e.g., hypothesis generation) from information search behavior. For instance, the cost of an information depository may be perceived as a cue to its worth (Illingworth & Thomas, 2015). It may be the case that memories associated with an information depository (e.g., its cost) contribute in unexpected ways to interfere with the signals postulated by HyGene. Additional study of the environments in which hypothesis testing takes place will help to parse out the varied potential signals combined to determine depository value.

A limitation of this dissertation worth noting is the abandonment of full-scale HyGene modeling in light of the performance exhibited by participants in all four studies. HyGene is best equipped to showcase a process account for behaviors linked to hypothesis generation. Within the context of the studies reported here, this behavior would manifest in response to evidence that participants' testing patterns were a byproduct of the presenting cue manipulation. In the place of HyGene simulations, components of the general process hypothesized to operate in the experimental tasks were used to evaluate specific behaviors recorded in the participants. Simulations demonstrated participants were responding to the information building into the tasks of Experiment 1 and 2. This result suggests HyGene cannot be discounted as a possible account of these data, as some consideration of the hypotheses could explain sensitivity to test diagnosticity.

Experiment 3 was the first study to my knowledge to both elicit metacognitive self-assessment in an information acquisition task and report a relation between that self-assessment and the amount of information foraging recorded in the experiment.

Specifically, participants generally increased the number of tests they exploited as their initial self-assessments decreased—a finding generally supportive of the probabilistic stopping rule posited in the hypothesis-driven valuation model. This result suggests that participants experienced some sensitivity to the information available in the environment despite the poor accuracy performance reported in the results section. The general rise in JOK over the course of trials wherein multiple tests were selected also suggested that participants were reacting to the newly acquired information. That increased confidence judgments followed data acquisition is evidence that participants recognized improvement in their knowledge about a case as its details were revealed—a sign that meaningful data was gleaned from test results. Moreover, the modeling endeavor for Experiment 3 suggested that these judgments mapped consistently to predicted posterior belief distributions.

Two results peripheral to the primary goal of this dissertation detected factors not previously associated with search behavior. Experiment 3 provided evidence suggesting that decisions to terminate search were related to beliefs. The initial confidence reported by participants predicted the amount of testing behavior they would exhibit in the task. Additionally, JOK elicitations revealed increasing confidence as more information was consumed, culminating in a plateau around 50% prior to search termination. The pattern points to changes in belief brought about by newly acquired data and the subsequent rise in confidence as important contributors to decisions to terminate search. That a link was detected between belief and termination afforded examination of the probabilistic termination rule posited in the hypothesis-driven valuation model. Simulation of the stopping rule found evidence in support of the model, as participants fit parameters

indicative of a high threshold for increases in expected value. Participants exhibited an expectation that the costs expended to acquire information would substantially increase the possibility of a correct diagnosis.

Experiment 4 provided insight regarding the role that environmental factors play in decisions to terminate search in sequential hypothesis testing. The findings demonstrated that decisions to exploit information depositories are sensitive to the costs attached to exploitation. I can also conclude that the valuation of information involves some mechanism by which costs impact perceived utility. The rules governing search termination must be sensitive to that valuation because search stopped after more testing behavior when information was free.

The role of memory in utility estimation was also enlightened by the results of Experiment 4. Specifically, behavior recorded in Experiment 4 suggests that experience plays an important role in how people perceive the cost of information. People accustomed to getting a lot of information for few costs are more sensitive to increased cost compared to those accustomed to moderate costs. Memory-based accounts for valuation show promise in their capacity to further explain testing behavior, especially as it applies to test preference and decisions to terminate search.

The HyGene predictions evaluated in these experiments have not been supported by the observed patterns of behavior. Specifically, the data cast doubt on the arrows feeding into Step 5 of the hypothesis-driven valuation model of hypothesis testing (Figure 1). Those arrows represent how episodic events activated by observed data and generated hypotheses contribute to valuation judgments prior to test exploitation. Not one experiment in this document found an effect of presenting cue on test selection or preference—which

suggests no influence of generated hypotheses. However, the findings reported in this document are insufficient for claiming a falsification of the cognitive process posited in Figure 1. In a number of ways discussed previously in this report, participants exhibited poor understanding of the task environments. This lack of understanding accounts for the observed behavior and is consistent with the theory. Had participants exhibited learning in Phase 1 of these experiments prior to exhibiting unpredicted patterns of testing behavior, the theoretical premise of these studies would be considerably jeopardized.

4.2 Implications

Hypothesis testing is a ubiquitous behavior, exhibited by people in a number of diverse settings. Poletiek (2001), for example, lists numerous behaviors—such as glancing at one's surroundings to evaluate expectations of surrounding objects, uttering sounds while learning a language to assess one's mastery of novel phonemes, and solving problems in novel ways so as to observe their impact on the world—that are, essentially, different forms of hypothesis testing behavior. Acquisition of information in such carefully crafted situations is important for how people behave within and understand the world. Thus, a comprehensive hypothesis testing theory can inform our understanding of a wide scope of human behavior.

The studies reported here showcase the complexity inherent in foraging tasks, and the care necessary to construct environments that can give rise to informative search behavior. Manipulations of information in sequential hypothesis testing tasks must be carefully planned to elicit specific predictions. Although careful consideration of numerous factors preceded design of the statistical structures implemented in the reported experiments, available tests could have been designed to be more informative. This would

shift the demands of the task further towards forming more nuanced connections between hypotheses and information depositories.

It is critical to understand that hypothesis testing does not occur in a vacuum. The parallels observed spanning myriad literatures strongly suggest that a number of cognitive processes share common mechanisms with hypothesis testing, precede and, thus, influence hypothesis testing, or are antecedents of hypothesis testing and are influenced by its byproducts. An abundance of environmental factors constrains or enhances these processes in ways that map to broader conceptualizations of human foraging behavior. This dissertation has taken an initial step towards establishing the first and only research program to account for individual differences in cognitive resources, environmental constraints, and the hypotheses considered by the decision-maker in hypothesis testing investigations. Although data to support the predictions of the HyGene architecture have proven to be elusive in this instance, the results of the work reported here clearly demonstrate a role for metacognitive self-assessment in decisions to terminate search and highlight the interaction of access costs with experience of costs when people perceive the value of engaging in testing behavior. These findings represent a starting point from which future investigations of hypothesis testing will be forged.

APPENDIX A. FIT STATISTICS FOR EXPERIMENT 1

BIC fit to each model for each participant.

		Tau (τ)			
Participant	Ideal	0.2	0.8	1.4	Random
1	-179.307	-70.5415	-20.6832	-13.8655	-179.307
2	-91.8218	-64.3031	-17.3763	-11.783	-91.8218
3	-1396.24	-235.925	-48.5642	-28.7361	-1396.24
4	-99.7972	-69.8917	-20.0456	-13.3902	-99.7972
5	-811.515	-161.05	-35.1511	-21.4655	-811.515
6	-17.4471	-17.5126	-7.60596	-6.62424	-17.4471
7	-1587.56	-263.166	-53.0861	-31.1241	-1587.56
8	-153.039	-121.158	-31.1817	-19.3017	-153.039
9	-17.6111	-23.7569	-10.1127	-8.00993	-17.6111
10	-932.091	-155.792	-30.5151	-18.6632	-932.091
11	-1551.21	-259.209	-52.5127	-30.8239	-1551.21
12	-87.8105	-74.7404	-19.7351	-12.9865	-87.8105
13	-175.644	-126.173	-32.5043	-20.0776	-175.644
14	-776.273	-154.242	-36.2502	-22.3694	-776.273
15	-224.18	-81.1814	-22.0218	-14.5511	-224.18
16	-134.968	-109.198	-29.6823	-18.627	-134.968
17	-415.731	-86.4187	-20.326	-13.4722	-415.731
18	-823.682	-168.444	-35.9661	-21.8059	-823.682
19	-15.4489	-16.3641	-6.88928	-6.27823	-15.4489
20	-1512.09	-254.316	-51.5014	-30.2641	-1512.09
21	-963.424	-180.486	-39.2195	-23.7156	-963.424
22	-154.815	-125.181	-32.3178	-20.014	-154.815
23	-210.476	-60.7055	-16.3916	-11.374	-210.476
24	-7.37721	-6.64443	-1.12265	-2.75686	-7.37721
25	-47.8818	-38.3234	-6.05915	-4.89262	-47.8818
26	-13.3214	-19.4583	-9.71004	-7.84457	-13.3214
27	-230.211	-22.057	-4.79308	-4.95095	-230.211
28	-356.461	-96.0871	-25.9657	-16.841	-356.461
29	-131.086	-58.7087	-18.4541	-12.7125	-131.086
30	-166.501	-21.7283	1.217918	-0.68634	-166.501
31	-120.985	-41.2147	-12.9049	-9.45061	-120.985

APPENDIX B. FIT STATISTICS FOR EXPERIMENT 2

BIC fit to each model for each participant under no time pressure

		Tau (τ)			
Participant	Ideal	0.2	0.8	1.4	Random
1	-795.97	-143.36	-31.23	-20.05	0
2	-36.32	-32.64	-14.24	-11.17	0
3	-110.13	-17.22	-7.16	-7.06	0
4	-795.97	-143.36	-31.23	-20.05	0
5	-777.10	-140.69	-30.83	-19.85	0
6	-382.32	-73.26	-18.55	-13.24	0
7	-180.36	-36.07	-11.59	-9.51	0
8	-418.43	-77.18	-18.78	-13.31	0
9	-313.20	-67.84	-18.78	-13.48	0
10	-735.21	-132.05	-29.06	-18.89	0
11	-795.97	-143.36	-31.23	-20.05	0
12	-6.09	3.72	-3.10	-4.89	0
13	57.14	45.70	14.21	5.73	0
14	14.16	7.50	-2.38	-4.50	0
15	-626.17	-119.35	-27.69	-18.22	0
16	-291.02	-50.18	-13.53	-10.52	0
17	-229.96	-63.33	-19.44	-13.94	0
18	14.16	7.50	-2.38	-4.50	0
19	-118.82	-10.16	-2.58	-4.12	0
20	-716.92	-129.78	-28.72	-18.70	0
21	-795.97	-143.36	-31.23	-20.05	0
22	41.04	45.49	14.88	6.18	0
23	-795.97	-143.36	-31.23	-20.05	0
24	-112.09	-18.22	-6.74	-6.70	0
25	-404.77	-78.97	-20.09	-14.13	0
26	-652.81	-115.73	-25.65	-17.01	0
27	14.16	7.50	-2.38	-4.50	0
28	-114.83	-33.06	-11.95	-9.72	0
29	14.16	7.50	-2.38	-4.50	0
30	-736.60	-133.15	-29.39	-19.07	0
31	-531.29	-92.60	-20.52	-14.11	0
32	-301.84	-71.09	-20.07	-14.21	0

33	-639.49	-117.48	-26.57	-17.54	0
34	-663.90	-124.69	-28.48	-18.63	0
35	-41.29	-36.66	-15.51	-11.90	0
36	-775.72	-139.59	-30.50	-19.66	0
37	-33.34	-17.73	-9.20	-8.28	0

BIC fit to each model for each participant under time pressure

		Tau (τ)			
Participant	Ideal	0.2	0.8	1.4	Random
1	-674.45	-120.48	-26.50	-17.44	0
2	-31.35	-28.37	-12.58	-10.17	0
3	-239.97	-46.59	-13.66	-10.65	0
4	-395.10	-57.23	-11.02	-8.64	0
5	-112.97	-31.71	-11.75	-9.65	0
6	-422.25	-80.53	-20.15	-14.15	0
7	-221.68	-44.20	-13.11	-10.32	0
8	-481.62	-90.74	-21.99	-15.13	0
9	-201.96	-53.33	-16.80	-12.45	0
10	-717.73	-130.48	-29.00	-18.87	0
11	-795.97	-143.36	-31.23	-20.05	0
12	-109.82	-41.81	-15.44	-11.79	0
13	-57.79	-11.93	-5.94	-6.29	0
14	-13.02	-1.79	-4.74	-5.81	0
15	-494.10	-100.68	-24.94	-16.79	0
16	-513.81	-91.66	-21.46	-14.79	0
17	-171.97	-54.23	-17.93	-13.14	0
18	-37.13	-33.34	-14.53	-11.34	0
19	-155.95	-28.49	-9.08	-8.00	0
20	4.14	13.59	1.84	-1.81	0
21	-701.63	-130.02	-29.26	-19.03	0
22	-1.30	36.24	12.52	4.80	0
23	-777.10	-140.69	-30.83	-19.85	0
24	-19.11	-19.00	-10.26	-8.94	0
25	-437.50	-92.68	-23.76	-16.18	0
26	-557.09	-101.41	-23.56	-15.94	0
27	-72.40	-12.01	-6.58	-6.79	0
28	-11.13	-12.46	-8.10	-7.69	0

29	4.46	-0.23	-4.68	-5.79	0
30	-588.43	-114.02	-26.90	-17.81	0
31	-774.33	-138.36	-29.98	-19.34	0
32	-468.30	-92.49	-22.91	-15.66	0
33	-581.50	-108.50	-25.26	-16.89	0
34	-720.50	-132.69	-29.65	-19.24	0
35	-41.29	-36.66	-15.51	-11.90	0
36	-795.97	-143.36	-31.23	-20.05	0
37	-48.29	-17.40	-8.84	-8.08	0

REFERENCES

Anderson, J. C. (1993). Problem solving and learning. *American Psychologist, 48*(1), 35.

Audley, R. J., & Pike, A. R. (1965). Some alternative stochastic models of choice. *British Journal of Mathematical and Statistical Psychology, 18*(2), 207-225.

Axt-Adams, P., van der Wouden, J. C., & van der Does, E. (1993). Influencing behavior of physicians ordering laboratory tests: A literature study. *Medical Care, 31*(9), 784-794.

Bainton, R. J., Tsai, L. T., Singh, C. M., Moore, M. S., Neckameyer, W. S., & Heberlein, U. (2000). Dopamine modulates acute responses to cocaine, nicotine and ethanol in Drosophila. *Current Biology, 10*(4), 187-194.

Barrett, S. L., Bell, R., Watson, D., & King, D. J. (2004). Effects of amisulpride, risperidone and chlorpromaze on auditory and visual latent inhibition, prepulse inhibition, executive function and eye movements in healthy volunteers. *Journal of Psychopharmacology, 18*(2), 156-172.

Baron, J. (1981). *Rationality and intelligence.* Cambridge: Cambridge University Press.

Bates, D. W., Evans, R. S., Murff, H., Stetson, P. D., Pizziferri, L., & Hripcsak, G. (2003). Detecting adverse events using information technology. *Journal of the American Medical Informatics Association, 10*(2), 115-128.

Beck, V. M., Hollingworth, A., & Luck. S. J. (2012). Simultaneous control of attention by multiple working memory representations. *Psychological Science, 23*(8), 887-898.

Beyth-Marom, R., & Fischhoff, B. (1983). Diagnosticity and pseudodiagnosticity. *Journal of Personality and Social Psychology, 45*(6), 1185.

Bower, G. H. (1961). A contrast effect in differential conditioning. *Journal of Experimental Psychology, 62*(2), 196-199

Busemeyer, J. R., & Rapoport, A. (1988). Psychological models of deferred decision-making. *Journal of Mathematical Psychology, 32*(2), 91-134.

Buttaccio, D. R., Lange, N. D., Hahn, S. & Thomas, R. P. (2014). Explicit awareness supports conditional visual search in the retrieval guidance paradigm. *Acta Psychologica, 145,* 44-53.

Buttaccio, D. R., Lange, N. D., Thomas, R. P., & Dougherty, M. R. (2015). Using a model of hypothesis generation to predict eye movements in a visual search task. *Memory & Cognition, 43*(2), 247-265.

Casscells, W., Schoenberger, A., & Grayboys, T. (2010). Interpretation by physicians of clinical laboratory results. *New England Journal of Medicine, 299*, 999-1000.

Cassini, M. H., Kacelnik, A., & Segura, E. T. (1990). The tale of the screaming hairy armadillo, the guinea pig and the marginal value theorem. *Animal Behavior, 39*, 1030-1050.

Charnov, E. L. (1976). Optimal foraging, the marginal value theorem. *Theoretical population biology, 9*(2), 129-136.

Cohen, J. D., McClure, S. M., & Yu, A. J. (2007). Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. *Philosophical Transactions of the Royal Society of London B: Biological Sciences, 362*(1481), 933-942.

Costermans, J., Lories, G., & Ansay, C. (1992). Confidence level and feeling of knowing in question answering: The weight of inferential processes. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18*(1), 142.

Cuthill, I. C., Haccou, P., & Kacelnik, A. (1994). Starlings (*Sturnus vulgaris*) exploiting patches: Response to long-term changes in travel time. *Behavioral Ecology, 5*, 81-90.

Daniels, M., & Schroeder, S. A. (1977). Variation among physicians in use of laboratory tests II. Relation to clinical productivity and outcomes of care. *Medical Care,* 482-487.

Davelaar, E. (2014). Semantic search in the remote associates test. *Topics in Cognitive Science, 7,* 494-512.

DeGroot, M. H. (1970). *Optimal statistical decisions.* New York, NY: McGraw-Hill.

Dougherty, M. R., Franco-Watkins, A. M., & Thomas, R. (2008). Psychological plausibility of the theory of probabilistic mental models and the fast and frugal heuristics. *Psychological Review, 115*(1), 199.

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

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

Dougherty, M. R., & Hunter, J. (2003b). Probability judgment and subadditivity: The role of working memory capacity and constraining retrieval. *Memory & Cognition, 31*(6), 968-982.

Dougherty, M. R., & Harbison, J. (2007). Motivated to retrieve: How often are you willing to go back to the well when the well is dry? *Journal of Experimental Psychology: Learning Memory, & Cognition, 33*(6), 1108-1117.

Dougherty, M. Thomas, R., & Lange, N. (2010). Toward an integrative theory of hypothesis generation, probability judgment, and hypothesis testing. *Psychology of Learning and Motivation, 52*, 299-342.

Downs, M., Turner, S., Bryans, M., Wilcock, J., Keady, J., Levin, E., O'Carroll, R., Howie, K., & Iliffe, S. (2006). Effectiveness of educational interventions in improving detection and management of dementia in primary care: Cluster randomized controlled study. *British Medical Journal, 332*(7543), 692-695.

Dror, I. E., Busemeyer, J. R., & Basola, B. (1999). Decision-making under time pressure: An independent test of sequential sampling models. *Memory & Cognition, 27* (4), 713-725.

Dursun, S. M., Wright, N. & Reveley, M. A. (1999). Effects of amphetamine on saccadic eye movements in man: Possible relevance to schizophrenia? *Journal of Psychopharmacology, 13*(3), 245-247.

Edwards, W. (1965). Optimal strategies for seeking information: Models for statistics, choice reaction times, and human information processing. *Journal of Mathematical Psychology, 2*, 312-329.

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

Estes, W. K. (1960). A random walk model for choice behavior. In K. J. Arrow, S. Karlin, & P. Suppes (Eds.), *Mathematical methods in the social sciences* (pp. 265-276). Stanford, CA: Stanford University Press.

Evans, J. S. B., & Over, D. E. (1996). Rationality in the selection task: Epistemic utility versus uncertainty reduction. *Psychological Review, 103*(2), 356-363.

Feldstein, A., Elmer, P. J., Smith, D. H., Herson M., Orwoll, E., Chen, C., Aickin, M., & Swain, M. C. (2006). Electronic medical record reminder improves osteoporosis management after a fracture: A randomized, controlled trial.

Ficic, M., & Buckman, M. (2013). Stopping rule selection (SRS) theory applied to deferred decision-making. In M. Knauff, M., Pauen, N., Sebanz, & I. Wachsmuth (Eds.) *Proceedings of the 35th Annual Conference of the Cognitive Science Societ* (pp. 2273-2278). Austin TX: Cognitive Science Society.

Fific, M., Little, D. R., & Nosofsky, R. M. (2010). Logical-rule models of classification response times: A synthesis of mental-architecture, random-walk, and decision-bound approaches. *Psychological Review, 117*, 309-348.

Flottorp, S., Oxman, A. D., Havelsrud, K., Treweek, S., & Herrin, J. (2002). Cluster randomized controlled trial of tailored interventions to improve the management of urinary tract infections in women and sore throat. *British Medical Journal, 325*(7360), 367-372.

Fu, W. T., & Anderson, J. R. (2006). From recurrent choice to skill learning: A reinforcement-learning model. *Journal of Experimental Psychology: General, 135*(2), 184.

Fu, W. T., & Pirolli, P. (2007). SNIF-ACT: A cognitive model of user navigation on the World Wide Web. *Human-Computer Interaction, 22*(4), 355-412.

Gallimore, P. (1994). Aspects of information processing in valuation judgment and choice. *Journal of Property Research, 11*(2), 97-110.

Gama, R., Harland, A. J., & Holland, M. R. (2001). Changing clinicians' laboratory test requesting behavior: Can the poacher turn gamekeeper? *Clinical Laboratory*, *47*(1-2), 57-66.

Gettys, C. F., & Fisher, S. D. (1979). Hypothesis plausibility and hypothesis generation. *Organizational behavior and human performance*, *24*(1), 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*(1), 23-51.

Gigerenzer, G. (2004). Fast and frugal heuristics: The tools of bounded rationality. *Blackwell handbook of judgment and decision-making*, 62-88.

Glucksberg, S., & McCloskey, M. (1981). Decisions about ignorance: Knowing that you don't know. *Journal of Experimental Psychology: Human Learning and Memory*, *7*(5), 311.

Gray, W. D., & Fu, W. T. (2004). Soft constraints in interactive behavior: The case of ignoring perfect knowledge in-the-world for imperfect knowledge in-the-head. *Cognitive Science*, *28*(3), 359-382.

Gray, W. D., Sims, C. R., Fu, W. T., & Schoelles, M. J. (2006). The soft constraints hypothesis: a rational analysis approach to resource allocation for interactive behavior. *Psychological Review*, *113*(3), 461.

Grether, G. F., Palombit, R. A., & Rodman, P. S. (1992). Gibbon foraging decisions and the marginal value model. *International Journal of Primatology*, *13*(1), 1-17. Gronlund, S. D., & Shiffrin, R. M. (1986). Retrieval strategies in recall of natural categories and categorized lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *12*(4), 550.

Gutzwiller, R. S., Wickens, C. D., & Clegg, B. A. (2014, September). Workload overload modeling An experiment with MATB II to inform a computational model of task management. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 58, No. 1, pp. 849-853). SAGE Publications.

Hanson, J. (1987). Tests of optimal foraging using an operant analogue. In A.C. Kamil, J.R. Krebs, & H. R. Pulliam (Eds.) Foraging Behavior. NY: Plenum Press

Harbison, J. I., Dougherty, M. R., Davelaar, E. J., & Fayyad, B. (2009). On the lawfulness of the decision to terminate memory search. *Cognition, 111*, 397-402.

Hendrickson, A. T., Navarro, D. J., & Perfors, A. (2016). Sensitivity to hypothesis size during information search. *Decision*, *3*(1), 62.

Herr, P. (1986). Consequences of priming: Judgment and behavior. *Journal of personality and Social Psychology, 51*, 1106-1115.

Hickner, J., Thompson, P. J., Wilkinson, T., Epner, P., Shaheen, M., Pollock, A. M., Lee, J., Duke, C. C., Jackson, B. R., & Taylor, J. R. (2014). Primary care physicians'

challenges in order clinical laboratory tests and interpreting results. *Journal of the Board of Family Medicine, 27*, 268-274.

Hills, T. T. (2006). Animal foraging and the evolution of goal-directed cognition. *Cognitive Science, 30*(1), 3-41.

Hills, T. T., Jones, M. N., & Todd, P. M. (2012). Optimal foraging in semantic memory. *Psychological Review, 119*(2), 431-440.

Hills, T. T., & Pachur, T. (2012). Dynamic search and working memory in social recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*(1), 218-228.

Hills, T. T., Todd, P. M., & Goldstone, R. L. (2008). Search in external and internal spaces evidence for generalized cognitive search processes. *Psychological Science, 19*(8), 802-808.

Hills, T. T., Todd, P. M., Lazer, D., Redish, A. D., & Couzin, I. D. (2015). Exploration versus exploitation in space mind, and society. *Trends in Cognitive Sciences, 19*(1), 46-54.

Hintzman, D. L. (1984). MINERVA 2: A simulation model of human memory. *Behavior Research Methods, Instruments, & Computers, 16*(2), 96-101.

Hintzman, D. L. (1986). Schema abstraction in a multiple-trace memory model. *Psychological Review, 93*, 306-355.

Houben, P. H. H., Winkens, R. A. G., van der Weijden, T., & Grol, R. P. T. M. (2010). Towards better use of test results: current problems and a research agenda. *Interpretation of diagnostic test results: Pretest expectations, test interpretation and*, 19.

Illingworth, D. A., & Thomas, R. P. (2015). *Price as information: Incidental search costs affect decisions to terminate information search and valuations of information sources.* Paper presented at the annual meeting of the Human Factors and Ergonomics Society, Los Angeles, California.

Illingworth, D. A., & Thomas, R. P. (in prep). Loss aversion costs a pretty penny: Framing affects decisions to terminate information acquisition.

James, W. (1890). The principles of psychology (Vol. 1). New York: Holt.

Johnson, E. J., Haubl, G., & Keinan, A. (2007). Aspects of endowment: A query theory of value construction. *Journal of Experimental Psychology: Learning, memory and Cognition, 33*(3), 461-474.

Johnson-Laird, P. N. (2010). Mental models and human reasoning. *Proceedings of the National Academy of Sciences, 107*(43), 18243-18250.

Johnson-Laird, P. N., & Byrne, R. M. (1991). *Deduction.* Lawrence Erlbaum Associates, Inc.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47*(2), 263-292.

Kahneman, D. & Tversky, A. (1996). On the reality of cognitive illusions. *Psychological Review, 103*(3), 582-591.

Kahneman, D., & Tversky, A. (2000). *Choices, values, and frames.* Cambridge University Press.

Kerstholt, J. H. (1994). The effect of time pressure on decision-making behavior in a dynamic task environment. *Acta Psychologica, 86*(1), 89-104.

Kirby, K. N. (1994). False alarm: A reply to Over and Evans. *Cognition, 52*(3), 245-250.

Klin, C. M., Guzman, A. E., & Levine, W. H. (1997). Knowing that you don't know: Metamemory and discourse processing. *Journal of Experimental Psychology: Learning, Memory, Cognition, 23*(6), 1378-1393.

Klayman, J., & Ha, Y. W. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review, 94*(2), 211-228.

Kolers, P. A., & Palef, S. R. (1976). Knowing not. *Memory & Cognition, 4*(5), 553-558.

Kruglanski, A. W., & Webster, D. M. (1996). Motivated closing of the mind: Seizing and freezing. *Psychological Review, 103*(2), 263-283.

Lange, N. D., Buttaccio, D. R., Sprenger, A. M., Harbison, I., Dougherty, M. R., & Thomas, R. P. (in press). The essential nature of memory underlying judgment, decision-making, and visual search. In V. Thompson & A. Feeney (Eds.), *Reasoning as memory*.

Lange, N. D., Thomas, R. P., Buttaccio, D. R., Illingworth, D. A., & Davelaar, E. J. (2013). Working memory dynamics bias the generation of beliefs: The influence of data presentation rate on hypothesis generation. *Psychonomic Bulletin & Review, 20*(1), 171-176.

Lange, N. D., Thomas, R. P., & Davelaar, E. J. (2012). Temporal dynamics of hypothesis generation: The influences of data serial order, data consistency, and elicitation timing. *Frontiers in Cognitive Science, 3*.

Levy, B. A., & Baddeley, A. (1971). Recall of semantic clusters in primary memory. *The Quarterly Journal of Experimental Psychology, 23*(1), 8-13.

Lima, S. L., & Dill, L. M. (1990). Behavioral decisions made under the risk of predation: A review and prospectus. *Canadian Journal of Zoology, 68*(4), 619-640.

Lindley, D. V. (1956). On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics, 27*(4), 986-1005.

Lopes, L. L. (1984). Risk and distributional inequality. *Journal of Experimental Psychology: Human Perception and Performance, 10*(4), 465-485.

Manktelow, K. I., & Over, D. E. (1990). Deontic thought and the selection task. *Lines of Thinking, 1*, 153-164.

McDonald, C. J., Hui, S. L., Smith, D. M., Tierney, W. M., Cohen, S. J., Weinberger, M., & McCabe, G. P. (1984). Reminders to physicians from an introspective computer medical record: A two-year randomized trial. *Annals of Internal Medicine, 100*(1), 130-138.

McNamara, J. M., & Houston, A. I. (1992). Evolutionarily stable levels of vigilance as a function of group size. *Animal Behaviour, 43*(4), 641-658.

Mehlhorn, K., Newell, B. R., Todd, P. M., Lee, M. D., Morgan, K., Braithwaite, V. A., Hausmann, D., Fiedler, K., & Gonzalez, C. (2015). Unpacking the exploration-exploitation tradeoff: A synthesis of human and animal literatures. *Decision, 2*(3), 191-215.

Metcalfe, J., & Murdock, B. B. (1981). An encoding and retrieval model of single-trial free recall. *Journal of Verbal Learning and Verbal Behavior, 20*(2), 161-189.

Miller, J. F., Weidemann, C. T., & Kahana, M. J. (2012). Recall termination in free recall. *Memory & Cognition, 4*, 540-550.

Mischel, W., Grusec, J., & Masters, J. C. (1969). Effects of expected delay time on the subjective value of rewards and punishments. *Journal of Personality and Social Psychology, 11*(4), 363-373.

Morrow, D. G., & Rogers, W. A. (2008). Environmental support: An integrative framework. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *50*(4), 589-613.

Muir, B. M. (1987). Trust between humans and machines, and the design of decision aids. *International Journal of Man-Machine Studies*, *27*(5-6), 527-539.

Muir, B. M. (1994). Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems. *Ergonomics*, *37*(11), 1905-1922.

Murdock, B. B., & Okada, R. (1970). Interresponse times in single-trial free recall. *Journal of Experimental Psychology*, *86*(2), 263.

Mynatt, C. R., Doherty, M. E., & Tweney, R. D. (1977). Confirmation bias in a simulated research environment: An experimental study of scientific inference. *The quarterly journal of experimental psychology*, *29*(1), 85-95.

Nelson, J. D. (2005). Finding useful questions: On Bayesian diagnosticity, probability, impact and information gain. *Psychological Review, 112,* 979-999.

Nelson, J. D., McKenzie, C. R. M., Cotrell, G. W., & Sejnowski, T. J. (2010). Experience matters: Information acquisition optimizes probability gain. *Psychological Science, 21,* 960-969.

Nelson, T. O., & Narens, L. (1980). A new technique for investigating the feeling of knowing. *Acta Psychologica*, *46*(1), 69-80.

Newcombe, F. (1969). *Missile wounds of the brain: A study of psychological deficits.* Oxford, England: Oxford University Press.

Nonacs, P. (2001). State dependent behavior and the marginal value theorem. *Behavioral Ecology*, *12*(1), 71-83.

Oaksford, M., & Chater, N. (1996). Rational explanation of the selection task. *Psychological Review, 103*(2), 381-391.

Olivers, C. N. L., Peters, J., Houtkamp, R., & Roelfsema, P. R. (2011). Different states in visual working memory: When it guides attention and when it does not. *Trends in Cognitive Sciences, 15*(7), 327-334.

Over, D. E., & Evans, J. S. B. (1994). Hits and misses: Kirby on the selection task. *Cognition, 52*(3), 235-243.

Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 52*(3), 381-410.

Peabody, F. W. (1922). The physician and the laboratory. *The Boston Medical and Surgical Journal, 187*(9), 324-327.

Phillips, L. D., & Edwards, W. (1966). Conservatism in a simple probability inference task. *Journal of Experimental Psychology, 72*(3), 346.

Pirolli, P. (2005). Rational analyses of information foraging on the web. *Cognitive Science, 29,* 343-373.

Pirolli, P, & Card, S. (1999). Information foraging. *Psychological Review, 106,* 643-675.

Pitz, G. F. (1968). Information seeking when available information is limited. *Journal of Experimental Psychology, 76*(1p1), 25.

Pleskac, T. J. (2012). Decision and choice: Luce's choice axiom. *International Encyclopedia of the Social and Behavior Sciences, 5*, 895-900.

Poletiek, F. H. (1995). Testing in a rule discovery task: Strategies of test choice and test result interpretation. *Contributions to decision research-I*, 335-350.

Poletiek, F. H. (1996). Paradoxes of falsification. *The Quarterly Journal of Experimental Psychology: Section A*, *49*(2), 447-462.

Poletiek, F. (2001). Hypothesis Testing Behavior (Essays in Cognitive Psychology).

Poletiek, F. H., & Berndsen, M. (2000). Hypothesis testing as risk behaviour with regard to beliefs. *Journal of Behavioral Decision-making*, *13*(1), 107.

Popper, K. R. (1959). The propensity interpretation of probability. *The British journal for the philosophy of science*, *10*(37), 25-42.

Popper, K. R. (1963). *Conjectures and Refutations. The Growth of Scientific Knowledge. (Essays and Lectures.)*. Routledge & Kegan Paul.

Pyke, G. H. (1978). Optimal foraging: movement patterns of bumblebees between inflorescences. *Theoretical population biology*, *13*(1), 72-98.

Raaijmakers, J. G., & Shiffrin, R. M. (1981). SAM: A theory of probabilistic search of associative memory. *The psychology of learning and motivation: Advances in research and theory*, *14*, 207-262.

Real, L., & Caraco, T. (1986). Risk and foraging in stochastic environments. *Annual Review of Ecology and Systematics*, 371-390.

Reder, L. M. (1987). Strategy selection in question answering. *Cognitive psychology*, *19*(1), 90-138.

Rieskamp, J., & Hoffrage, U. (2008). Inferences under time pressure: How opportunity costs affect strategy selection. *Acta Psychologica*, *127*(2), 258-276.

Rogers, S. (1941). *The anchoring of absolute judgments* (No. 261).

Roukema, J., Steyerberg, E. W., Van der Lei, J., & Moll, H. A. (2008). Randomized trial of a clinical decision support system: Impact on the management of children with

fever without apparent source. *Journal of the American Medical Informatics Association, 15*(1), 107-113.

Savage, L. J. (1954). *The foundations of statistics.* New York: Wiley.

Scaife, M., & Rogers, Y. (1996). External cognition: How do graphical representations work? *International Journal of Human-Computer Studies*, 45, 115–143.

Schroeder, S. A., Kenders, K., Cooper, J. K., & Piemme, T. E. (1974). Use of laboratory tests and pharmaceuticals: Variation among physicians and effect of cost audit on subsequent use. *JAMA, 225*, 969.

Sherif, M., Taub, D., & Hovland, C. I. (1958). Assimilation and contrast effects of anchoring stimuli on judgments. *Journal of Experimental Psychology, 55*(2), 150-155.

Singer, M. (1984). Toward a model of question answering: Yes-no questions. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 10*, 285-297.

Singer, M., & Tiede, H. L. (2008). Feeling of knowing and duration of unsuccessful memory search. *Memory & Cognition, 36*(3), 588-597.

Skinner, J. (2013). The costly paradox of health-care technology. *Technology Review, 116,* 69-70.

Smith, K. A., Huber, D. E., & Vul, E. (2013). Multiply-constrained semantic search in the Remote Associates Test. *Cognition, 128*, 64-75.

Solomon, D. H., Hashimoto, H., Daltroy, L., & Liang, M. H. (1998). Techniques to improve physicians' use of diagnostic tests: A new conceptual framework. *JAMA, 280*(3), 2020-2027.

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

Stephens, D. W. (1981). The logic of risk-sensitive foraging preferences. *Animal Behaviour, 29*(2), 628-629.

Stephens, D. W., & Krebs, J. R. (1986). *Foraging Theory*, Princeton: Princeton University Press.

Stewart, N., Chater, N., & Brown, G. D. A. (2006) Decision by sampling. *Cognitive Psychology, 53*(1), 1-26.

Stock, J. B., & Surette, M. G. (1996). Chemotaxis. In F. C. Neidhardt (Ed.), *Escherichia coli and salmonella: Cellular and molecular biology* (pp. 1103-1129). Washington, DC: ASM Press.

Sundaram, V., Lazzeroni, L. C., Douglass, L. R., Sanders, G. D., Tempio, P., & Owens, D. K. (2009). A randomized trial of computer-based reminders and audit and feedback to improve HIV screening in a primary care setting. *International journal of STD & AIDS, 20*(8), 527-533.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction (Vol. 1, No.1)*. Cambridge: MIT press.

Thomas, R., Dougherty, M., R., & Buttaccio, D. R. (2014). Memory constraints on hypothesis generation and decision-making. *Current Directions in Psychological Science, 23*(4), 264-270.

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

Tierney, W. M., McDonald, C. J., Martin, D. K., HUui, S. L., & Rogers, M. P. (1987). Computerized display of past test results: effect on outpatient testing. *Annals of internal medicine, 107*(4), 569-574.

Trope, Y., Bassok, M., & Alon, E. (1984). The questions lay interviewers ask. *Journal of Personality, 52*(1), 90-106.

Troyer, A. K., Moscovitch, M. & Winocur, G. (1997). Clustering and switching as two components of verbal fluency: Evidence from younger and older healthy adults. *Neuropsychology, 11*(1), 138-146.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science, 211,* 453-458.

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty, 5*, 297-323.

Tversky, A., & Koehler, D. J. (1994). Support theory: A nonextensional representation of subjective probability. *Psychological Review, 101*(4), 547-567.

van der Weijden, T., van Bokhoven, M. A., Dinant, G. J., van Hasselt, C. M., & Grol, R. P. (2002). Understanding laboratory testing in diagnostic uncertainty: A qualitative study in general practice. *British Journal of General Practice, 52*(485), 974-980.

van der Weijden, T., van Velsen, M., Dinant, G. J., van Hasselt, C. M., & Grol, R. (2003). Unexplained complaints in general practice: Prevalence, patients' expectations, and professionals' test-ordering behavior. *Medical Decision-making, 23*(3), 226-231.

Verdolin, J. L. (2006). Meta-analysis of foraging and predation risk trade-offs in terrestrial systems. *Behavioral Ecology and Sociobiology, 60*(4), 457-464.

Viviani, P. (1979). A diffusion model for discrimination of temporal numerosity. *Journal of Mathematical Psychology, 19*(2), 108-136.

Wald, A. (1947). *Sequential Analysis*. Wiley, New York.

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology, 12*(3), 129-140.

Wason, P. C. (1966). Reasoning. In B. M. Foss (Ed.), *New horizons in psychology I* (pp. 135-151). Harmondsworth, UK: Penguin.

Wason, P. C. (1968). Reasoning about a rule. *The Quarterly Journal of Experimental Psychology, 20*(3), 273-281.

Weber, E. U. (1988). A descriptive measure of risk. *Acta Psychological, 69*, 185-203.

Weber, E. U., Bockenholt, U., Hilton, D. J., & Wallace, B. (1993). Determinants of diagnostic hypothesis generation: Effects of information, base rates, and experience. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19*(5), 1151-1164.

Weber, E. U., & Johnson, E. J. (2006). Constructing preferences from memories. In S. Lichtenstein & P. Slovic (Eds.). *The construction of value* (pp. 397-410). New York: Cambridge University Press.

Weber, E. U., Johnson, E. J., Milch, K. F., Chang, H., Brodscholl, J. C., & Goldstein, D. G. (2007). Asymmetric discounting in intertemporal choice: A query-theory account. *Psychological Science, 18*(5), 516-523.

Weber, E. U., & Milliman, R. (1997). Perceived risk attitudes: Relating risk perception to risky choice. *Management Science, 43,* 122-143.

Wegwart, O., Schwartz, L. M., Woloshin, S., Gaissmaier, W., & Gigerenzer, G. (2012). Do physicians understand cancer screening statistics? A national survey of primary care physicians in the United States. *Annals of Internal Medicine, 156*(5), 340-349.

Wickens, C. D. (2014). Effort in human factors performance and decision-making. *Human Factors, 56*(8), 1329-1336.

Wickens, C. D., Santamaria, A., & Sebok, A. (2013, September). A computational model of task overload management and task switching. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 47, No. 1, pp. 763-767). SAGE Publications.

Winkens, R., & Dinant, G. J. (2002). Rational, cost effective use of investigations in clinical practice. *British Medical Journal, 324*(7340), 783.