COST SENSITIVITY IN EXTERNAL INFORMATION SEARCH AND HYPOTHESIS TESTING

A Dissertation
Presented to
The Academic Faculty

by

Vikram Sahgal

In Partial Fulfillment
of the Requirements for the Degree
Masters in Psychology in the
Georgia Institute of Technology

Georgia Institute of Technology
May 2025

COST SENSITIVITY IN EXTERNAL INFORMATION SEARCH AND HYPOTHESIS TESTING

Approved by:

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

Dr. Christopher Stanzione
School of Psychology
Georgia Institute of Technology

Date Approved: April 17th, 2025

ACKNOWLEDGEMENTS

I would like to thank my mother, father, and sister for always believing in me.

I would also like to thank Rick Thomas for his unconditional support through this project.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

MDG: Medical Diagnosis Game

Chapter 1: Introduction

Human nature often leads people to search for information that aligns with the hypotheses they are testing (Bhatteracharjee & Machuga, 2004), through a process known as hypothesis-guided search (Thomas et al., 2008; Thomas et al., 2014). This cognitive mechanism highlights the value individuals place on certain information. Hypothesis guided search is also key to understanding the conditions under which people engage in confirmatory versus diagnostic information-seeking behaviors (Dougherty et al., 2010; Illingworth & Thomas, 2015; Lange et al., 2014; Thomas et al., 2014).

This body of research along with other literature published in the field of decision science outlines the fact that human decision-making is shaped by cognitive limitations and situational resource constraints, such as time or motivation. These aspects play a large role in explaining cognitive biases seen in day to day and professional decision-making contexts (Tversky & Kahneman, 1973; Dougherty et al., 1999; Thomas et al., 2008).

For example, healthcare professionals regularly make diagnostic decisions that directly impact patients and their plan of care (Elstein & Schwarz, 2002; Weber, Bockenholt, Hilton, & Wallace, 1993). Cognitive limitations in attention and memory along with task constraints can impact people's diagnostic ability by hindering the generation of a comprehensive set of hypotheses. This shortcoming often leads to errors in decision-making, overconfidence, and biased approaches to seeking information (Thomas et al., 2008; Thomas et al., 2014; Dougherty et al., 2010; Tidwell et al., 2016).

Hypothesis generation is key to understanding how individuals test the validity of the assumptions they hold. By understanding hypothesis generation, researchers gain insight into how people think and make critical decisions. Cued recall initiates hypothesis generation as data from the environment can be retrieval cues that prompt the retrieval or generation of hypotheses (Thomas et al., 2008). Using observed data allows for one to generate possible explanations based on prior knowledge and experiences stored in long term memory. In an optimal setting, participants will consider the complete hypothesis space, which includes all possible hypotheses in the sample space. The generated hypotheses then guide an information search to assess the validity of the selected hypotheses. This process repeats with new data presentation and new hypothesis generation. However, hypothesis testing in real world scenarios operates in a much different manner. As described by Thomas et al. (2008), the human brain has to deal with cognitive constraints which shape hypothesis generation and information search strategies. The working memory system can only actively consider a limited subset of hypotheses at a time. As a result, working memory constraints further direct information search by assessing the validity of new information and prompting further hypothesis generation based upon this information.

Hypothesis guided search highlights the value of a medical test in its ability to influence a physician's revised beliefs (posterior probabilities) regarding the diagnostic hypotheses under consideration. Hypotheses also play a crucial role in enabling iterative reevaluation of beliefs as incoming data from the environment is interpreted and evaluated. This process helps to progressively narrow the range of possible diagnoses, clarifying ambiguous data and identifying the most likely conditions.

An effective decision-making system must determine the appropriate point to cease its internal search for hypotheses and its external search for additional information. This decision occurs when further investigation or testing is unlikely to provide meaningful insights (Dougherty & Harbison, 2007; Harbison et al., 2009, 2012; Dougherty, Thomas, & Lange, 2010; Illingworth & Thomas, 2015; Thomas et al., 2014). Research suggests that search efforts typically end when the anticipated benefit of gathering additional information no longer outweighs the benefit of acting on the current information. Information search costs play a large role in the direction of hypothesis testing and search termination. The decision to stop is influenced by a threshold parameter that reflects individual differences or contextual factors, such as time pressure, which can significantly impact when search termination occurs (Dougherty & Harbison, 2007; Harbison et al., 2009, 2012).

Goal directed cognition has been documented in animal foraging behavior and can be extrapolated to humans searching for information (Hills 2006). The Optimal Foraging Theory states that animals adopt strategies that maximize energy intake based upon the effort expended (Stephens & Krebs, 1986). Hypothesis testing in humans follows this logic as humans will expend cognitive effort on the particular set of hypotheses they deem to hold the most value. Another foraging model with similarities to human hypothesis testing is the Marginal Value Theory (Charnov 1976) which predicts that animals stay in a patch of resources until they fall below the average rate of return for the environment. Human decision-making mirrors this cost benefit analysis as data gathering and hypothesis testing can be terminated when the expected value of the additional information falls below the associated cost of search (Harbinson et al,. 2009).

Previous literature has investigated foraging behavior as it provides a useful framework for understanding the role of costs in hypothesis testing. The Marginal Value Theorem (Charnov, 1976) details that resource acquisition is determined by a tradeoff between benefits and costs, a concept that applies to hypothesis testing, where individuals must balance belief confirmation and diagnostic information-seeking behaviors against constraints and limitations. Optimal foraging theory (Stephens & Krebs, 1986) and risk-sensitive foraging models (Lima & Dill, 1990; Nonacs, 2001) further emphasize how decision-making is shaped by external pressures such as opportunity costs, predation risk, and cognitive effort. Additionally, models of information foraging (Pirolli & Card, 1999) and memory search termination (Davelaar, 2014; Dougherty & Harbison, 2007; Harbison et al., 2009) suggest that individuals weigh the expected value of continued search against the costs of additional retrieval attempts. The aforementioned perspectives provide a more comprehensive understanding of hypothesis-guided search in environments where information acquisition carries significant costs.

Thomas, Dougherty, Harbison, and Spregler's work in *Psychological Review* (2008) was a key piece of literature in the field, providing a basis for subsequent research into the cognitive mechanisms of hypothesis testing and memory retrieval. Leveraging this theoretical framework, Dougherty, Thomas, and Lange (2010) introduced HyGene, a computational model of hypothesis generation that connects probability judgments with hypothesis testing. The paper also investigates the role of memory retrieval in pre-decisional processes. HyGene offers a theoretical basis for understanding how judgment and choice are influenced by memory processes prior to decision-making.

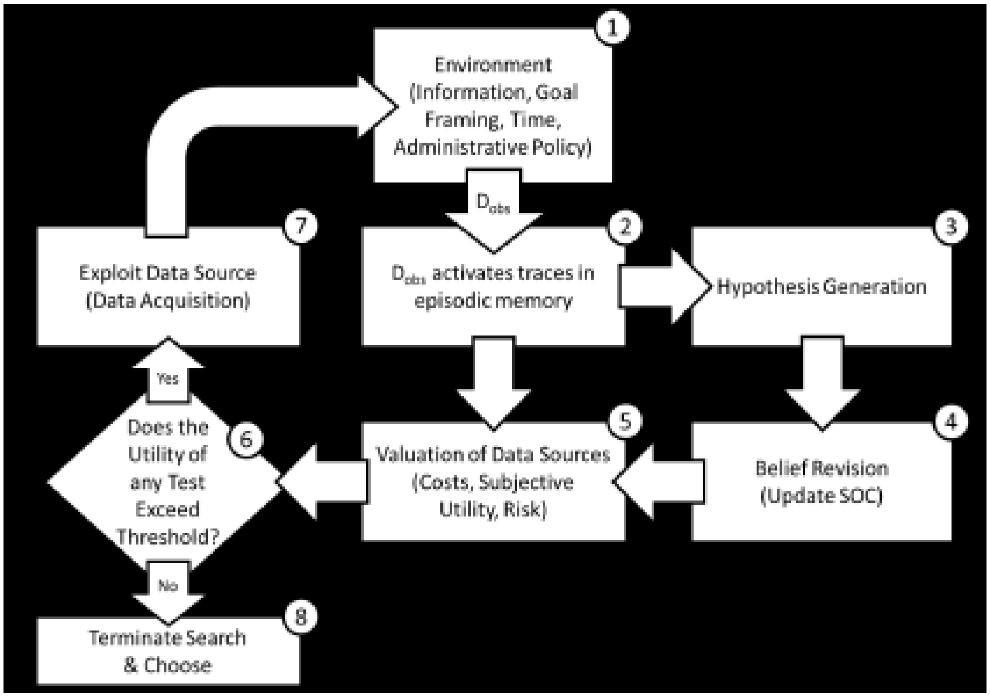

Figure 1- HyGene Seek

The diagram illustrates a hypothesis-driven information search model within a task. The elements represent cognitive system processes, while elements in Panel 1 correspond to the external environment.

The experiment will employ the computational architecture known as HyGene as it is used to predict and explain human behavior in tasks related to hypothesis generation and information search, providing a framework for data analysis and hypothesis testing. HyGene is based on three principles: data from the environment are used as retrieval cues for hypotheses from long-term memory, working memory constrains the number of hypotheses actively considered, and the hypotheses generated in working memory are compared for probability judgments and hypothesis testing.

The process begins with external data (Step 1), which triggers cued recall from memory, activating episodic memory based on the observed data (Step 2). This leads to generating an initial hypothesis (Step 3) (Dougherty, Thomas, & Lange, 2010). Subsequently, belief revision occurs, where the *Set of Leading Contenders* is updated based on which hypotheses within this set are most supported (Step 4). The *Set of Leading Contenders* represents the dynamic workspace of a participant's working memory during the task (Thomas et al., 2008).

Participants then evaluate the utility of available information sources (Step 5) to decide if the utility of a source, such as a test selection, exceeds a predefined threshold (Step 6). If the threshold is surpassed, participants exploit the data source (Step 7), selecting a test and restarting the cycle at Step 1 with new data from the updated environment. If the threshold is not met in Step 6, the search behavior terminates, and participants make a final answer selection (Step 8) (Illingworth & Thomas, 2022).

This process is modeled by the HyGene architecture, which tracks decisions to terminate hypothesis testing. The termination rule used in the model is depicted below. This rule illustrates how the decision to stop hypothesis testing depends on the sensitivity to changes in the expected value of knowledge in potential future states compared to current knowledge (Δev), relative to a required threshold (θ), and most importantly for this experiment it captures the sensitivity of the decision maker to the costs of information (Harbison, Dougherty, Davelaar, & Fayyad, 2009). P(T) is the probability of termination. The model predicts that termination behavior decreases as the expected value change grows relative to the threshold.

$$P(T) = \frac{1}{1 + e^{g(\Delta ev - \theta)}}$$

Thomas, Buttacio, and Dougherty (2014) further discussed this work by highlighting connections between hypothesis generation, probability judgment, and information search in decision-making tasks. This collection of previous literature demonstrates how factors such as belief strength influence information valuation and foraging behaviors (Illingworth & Thomas, 2022). Illingworth and Thomas' 2016 study looks at another aspect of the information search process by looking at the perceived value of information sources and their role in search termination.

The Medical Diagnoses Game is a data collection paradigm that simulates medical diagnostic tasks (MDG, Illingworth & Thomas, 2015; 2022), capturing test selections and diagnostic decisions made by participants. This study uses the MDG task to examine how participants adjust their search behaviors in response to search cost and information value. Previous research has also investigated how cognitive processes, as well as belief strength, impact information search (Illingworth & Thomas, 2022).

The publication of Noise: A Flaw in Human Judgment by Daniel Kahneman, Olivier Sibony, and Cass Sunstein has led to the concept of decision variability receiving growing attention from both the scientific community as well as funding agencies. The book discusses the topic of noise, which is defined as the unwanted variability in human targets. Decision bias is conceptually defined as less variable decisions while bias refers to the systematic deviations from the average behavior.

1.1 Present Study

The experiment detailed in this paper investigates the impact of cost structures on diagnostic decision-making and information search. The MDG presents participants with simulated medical scenarios where they must diagnose patients based on their symptoms and the results of diagnostic tests. The game is designed to simulate the real-world challenges of medical diagnosis, where clinicians must generate hypotheses, order tests, and make treatment decisions based on limited information and resources available to them. The study builds upon previous research leveraging the MDG paradigm (Illingworth & Thomas, 2022), as the central aim of this study was to determine how the introduction of a cost structure for diagnostic tests influences participants' decision-making strategies when evaluating presenting cues. The experiment presents participants with one of two presenting cues. The first presenting cue is associated with an advantageous test that has diagnostic utility while the second presenting cue has two test selections associated that have equal diagnosticity and differing costs.

Researcher's general predictions for the experiment revolved around participants belief and cost sensitivity and can be seen in the hypothesis below.

Hypothesis 1 (Belief Sensitivity): We expect participants to be sensitive to the diagnosticity of tests where they prefer tests that provide more information to those that provide less information in general.

Hypothesis 2 (Cost Sensitivity): Under conditions where the 2 most informative tests have the same diagnosticity and differing costs, we expect participants to select the cheaper test.

Hypothesis 3 (Search Termination): We expect participants to terminate search earlier when they as they become more confident in a single hypothesis (Diagnosis).

Hypothesis 4 (Data Consistency): Participants are more likely to terminate search when a test outcome confirms their current beliefs.

Hypothesis 5 (Decision Variability): By presenting participants with identical presenting cues during the test phase, researchers can measure how much individuals differ in their test-selection behaviors. Hypotheses are exploratory but researchers generally expected that participants who were most consistent in their search behavior under the same context will tend to select better cues and be more accurate in their diagnosis.

Chapter 2. Methods

2.1 Participants

Participants were undergraduate students from the Georgia Institute of Technology and were recruited via the Georgia Tech subject pool (SONA). Georgia Tech students were compensated with course credit for their participation. Any participant who encountered technical errors or had an incomplete data set was not included in the study. As a result, a total of 119 participants completed the study session and had their results evaluated in the study.

2.2 Procedure

In the first part of experiment 1, participants underwent a diagnostic learning phase. During the phase, participants learned to identify presenting medical cues and use diagnostic tests to diagnose a fictitious disease. There was no cost associated with the medical tests, allowing participants to focus solely on learning the relationships between cues and diagnoses.

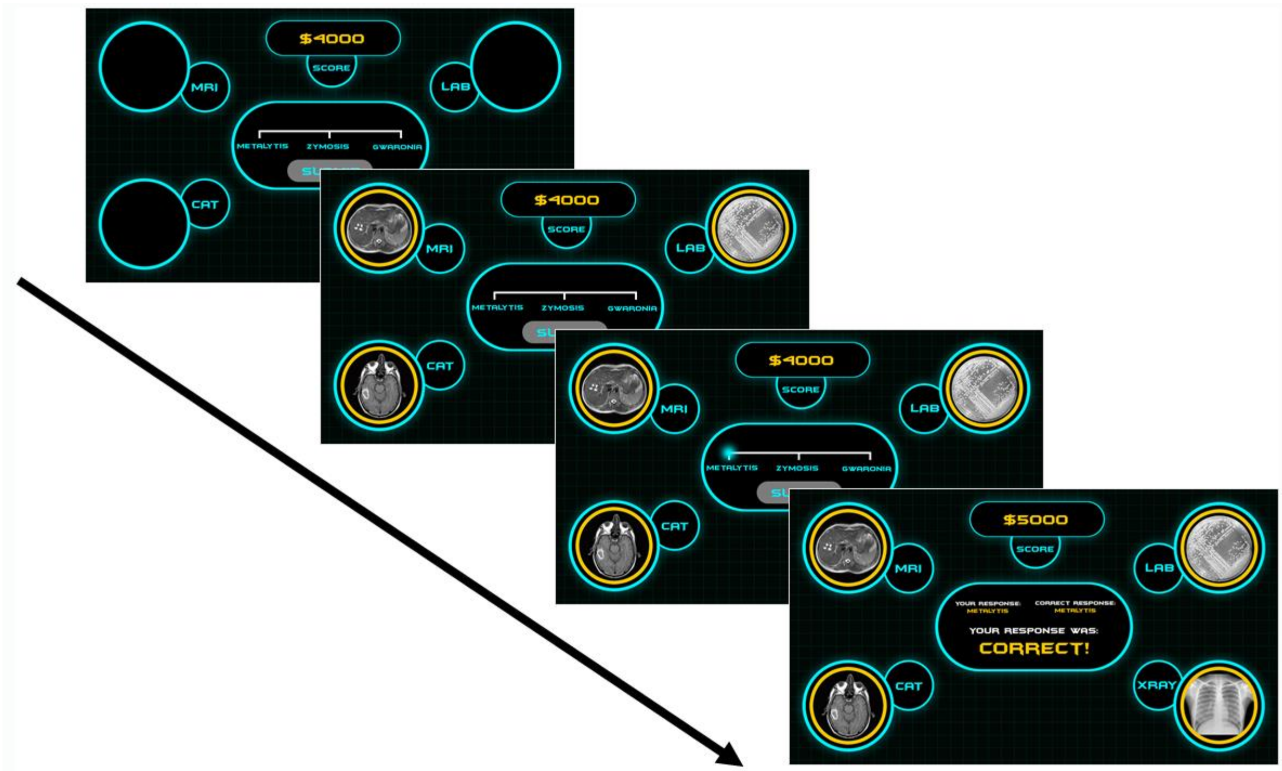

Figure 2- Training Trail Example Screen

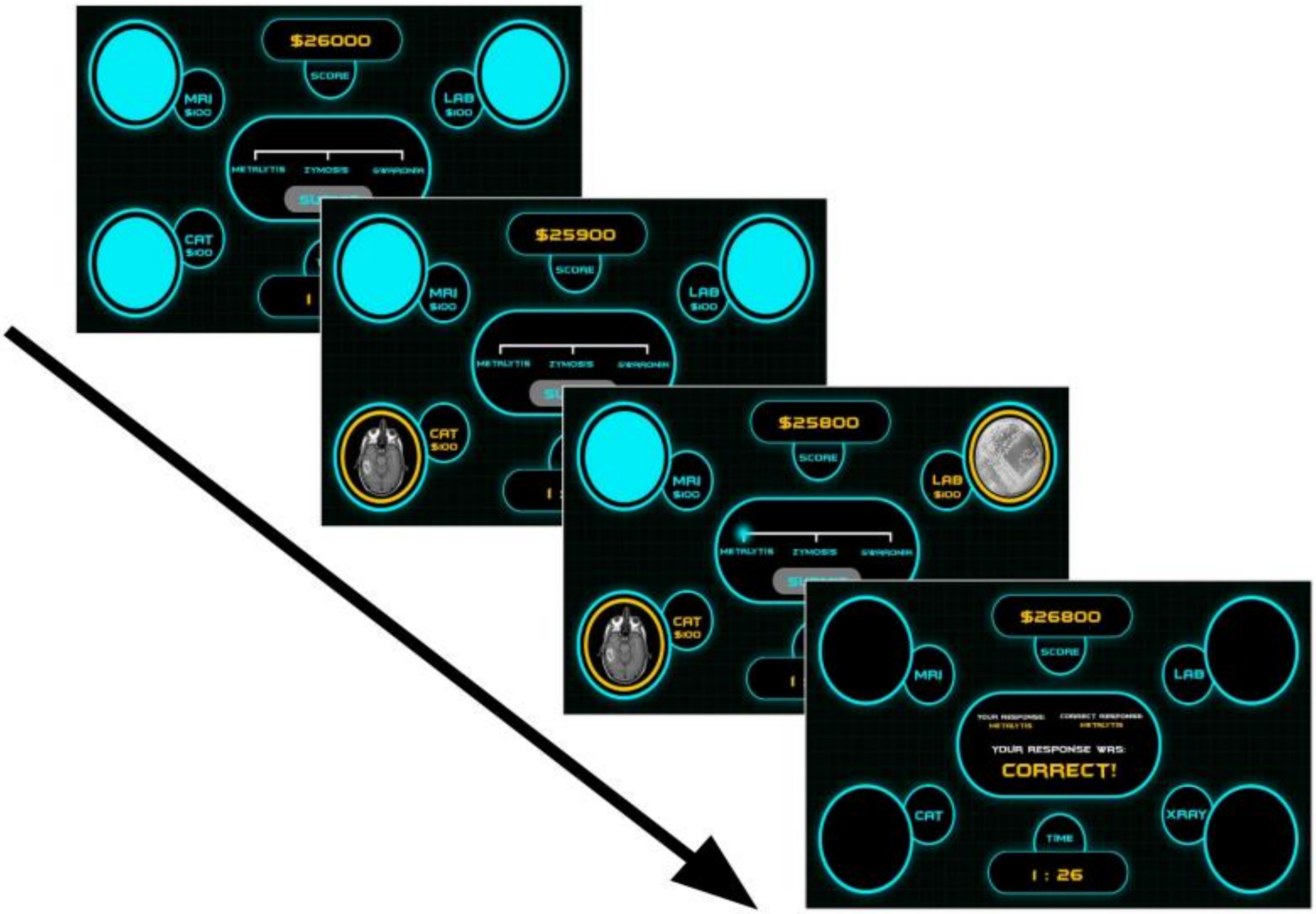

Figure 3- Test Trail Example

Figure 2 (Panel A) illustrates a learning trial in the MDG paradigm. During learning trials, all test results appear automatically after 2000 ms. In contrast, during test trials (panel B), test results appear only after participants selects their preferred tests. Additionally, during the test phase of the experiment, the cost of each test (in game dollars) is displayed alongside it's label. During training, there is no penalty for an incorrect diagnosis; however, during the testing phase, an incorrect diagnosis incurs a $500 penalty. In both phases, a correct diagnosis earns $1000. Participants choose when to make a diagnosis, after which feedback is provided.

Participants earned points for correct diagnoses, and all medical test results were provided to them, ensuring they had the maximum possible information. This phase aimed to teach participants how to recognize cues before the second part of the experiment and to measure their learning progress. The learning phase consisted of hundreds of trials, organized into twenty-three learning blocks.

The experiment introduced a cost structure for the diagnostic tests as participants had to consider the cost associated with each test when deciding which ones to select in the test phase. Participants were required to predict which tests would be the most informative based on the presenting cue before making a diagnosis. Selecting more tests resulted in a greater loss of in-game money, reducing the potential earnings from a correct diagnosis. Participants gain 1,000 points for a correct diagnosis and are penalized 500 points for an incorrect diagnosis. This setup was designed to sensitize participants to the cost of each test and encourage them to be more intentional in their selection of informative tests based on the presenting cue. Participants had to weigh the diagnosticity (how informative each test is) of each test against its cost. Consequently, the training phase should heighten participants' awareness of test costs, encouraging them to be more deliberate in selecting the most informative tests based on the presenting cue.

2.3 Experiment Ecology

Table 1 displays the presenting cues, hypotheses, and diagnosticity associated with the experiment. In the learning phase, participants are presented with 2 different presenting cues. Each of the cues is associated with given diseases (hypothesis) in order to help participants learn the associations as they see more trials in the learning phase of the experiment. This will also allow participants to learn the diagnosticity of each test relative to the presenting cue. In the testing phase of the experiment, participants will have to balance informativeness of each test with its associated cost in their hypothesis guided search. As seen in Table 1, presenting cue 1 is associated with hypothesis 1 and 2. With hypothesis 1 and 2, the most informative test is the first diagnostic test which is also the most expensive. Researchers hypothesize that when presented with cue 1, participants first test selected will be test one prior to diagnosis. Presenting cue 2 is associated with hypothesis 3 and 4. In hypothesis 3 and 4, the first and second tests have equal diagnosticity but test 1 has a higher associated cost than test 2. As a result, researchers hypothesize that cost sensitivity will lead participants to select test 2 more frequently than test 1 despite the tests have equal informativeness. The study manipulated the presenting cues to observe how participants responded to the informational value and cost of each test. In other words, the experiment was constructed to shed light on how the participants traded changes in the informational value of the medical tests, resulting from belief updated via the manipulation of presenting cues, with the monetary costs (points in the MDG game) of the tests.

	Presenting Symptoms			Diagnostic Tests			
	Cue 1	Cue 2		Test 1	Test 2	Test 3	Test 4
Hypothesis 1	.9	.1	Outcome 1	.8	.7	.4	.5
			Outcome 2	.2	.3	.6	.5
Hypothesis 2	.9	.1	Outcome 1	.2	.3	.6	.5
			Outcome 2	.8	.7	.4	.5
Hypothesis 3	.1	.9	Outcome 1	.8	.8	.5	.5
			Outcome 2	.2	.2	.5	.5
Hypothesis 4	.1	.9	Outcome 1	.2	.2	.5	.5
			Outcome 2	.8	.8	.5	.5
Test Costs				$320	$280	$160	$120

Table 1: Conditional Probabilities based upon Presenting Cues

Chapter 3. Results

The results of Experiment 1, which used the Medical Diagnosis Game (MDG), indicated that participants demonstrated learning and were sensitive to both the informational value and cost of diagnostic tests.

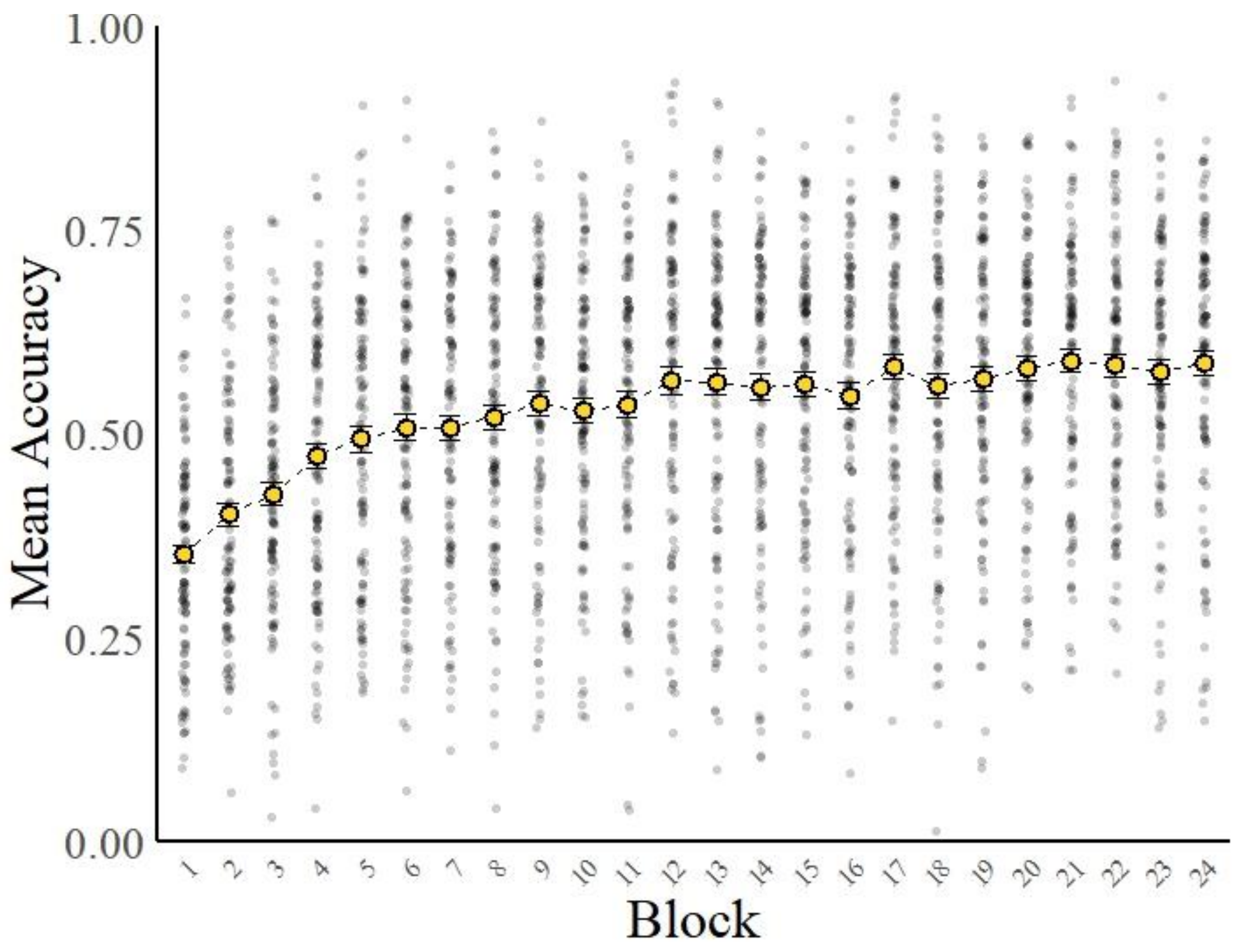

Figure 4 - Mean Diagnostic Accuracy in each learning block of the study

Figure 4 illustrates the results of the first part of the experiment, in which participants learn to pick a diagnosis based on a presenting cue. A generalized linear mixed model evaluated how well accuracy was predicted by trial block. Accuracy was coded 1 for correct responses and 0 for incorrect responses. Trial block was found to be predictive of learning phase accuracy, $\chi2(23, N = 119) = 846.80$, $p < .001$, suggesting that performance improved over the course of the 24 blocks of learning trials. Participants were more likely to submit a correct diagnosis in the final block than they were in the first block (odds ratio [OR] = 2.74, $p < .001$, 95% confidence interval [CI] = [2.43, 3.09]). The results of this generalized linear mixed model supports the hypothesis that confirm that participants demonstrated learning through the experiment.

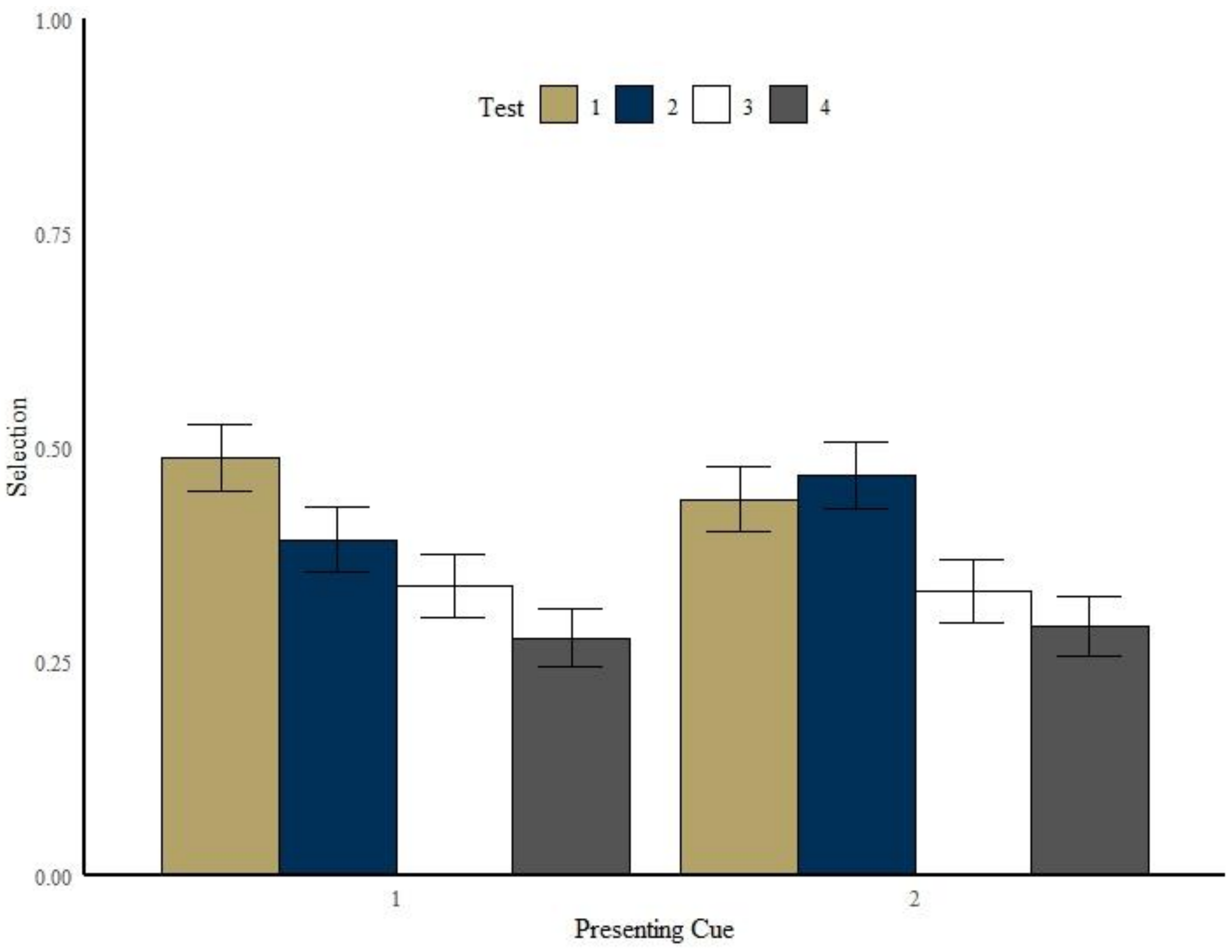

Figure 5- Test Selection Results based on Presenting Cue Displayed

To examine the effects of the presenting cue, test, and their interaction on overall test selection, a generalized linear mixed model was constructed with participant as a random effect. The results revealed significant main effects of presenting cue, ($\chi^2(1, N = 119) = 4.22$, $p = .039$, and test, ($\chi^2(3, N = 119) = 1124.41$, $p < .001$), indicating that participants' test selection preferences were influenced by both factors. However, these main effects were accompanied by a significant interaction between presenting cue and test, ($\chi^2(3, N = 119) = 111.85$, $p < .001$), as shown in Figure 5, suggesting that participants' test preference patterns varied depending on the presenting cue.

To further investigate this relationship, researchers analyzed each test individually using a mixed model, which treated presenting cue as a fixed factor and participant as a random effect. The overall selection patterns for Test 1 and Test 2 demonstrated that participants learned to select advantageously and were sensitive to the diagnostic ability of the tests. Specifically, Test 1 was the most diagnostic given Cue 1, while Test 2 becomes equally diagnostic when presented with Cue 2. This sensitivity to diagnostic ability is reflected in the significant changes in test preferences across cues.

Researchers argue hypothesis that cost based sensitivity will be seen in the participants where selecting more tests results in greater in-game monetary loss and reduces the reward for a correct diagnosis.

As predicted, the selection of Test 1 was significantly lower when presented with Cue 2 compared to Cue 1, ($\chi^2(1, N = 119) = 57.67$, $p < .001$. Specifically, the odds of selecting Test 1 were 0.609 times lower when Cue 2 was presented ($OR = 0.609$, $Z = -7.59$, $p < .001$, 95% CI [0.54, 0.69]). This decrease in preference for Test 1 under Cue 2 suggests that participants adjusted their choices based on the diagnostic ability of the test, as Test 1 is less diagnostic under

Cue 2 compared to Cue 1. The intercept for this model was not statistically significant (OR = 0.85, Z = -0.33, $p = 0.739$, 95% CI [0.33, 2.18]), indicating no baseline preference for Test 1 when controlling for the cue.

For Test 2, the selection was significantly influenced by the presenting cue, ($\chi^2(1, N = 119) = 140.81$, $p < .001$. The odds of selecting Test 2 were 2.22 times higher when Cue 2 was presented compared to Cue 1 (OR = 2.22, Z = 11.87, $p < .001$, 95% CI [1.94, 2.53]). This increase in preference for Test 2 under Cue 2 reflects participants' sensitivity to the diagnostic ability of the test as well as cost sensitivity, as Test 2 becomes more advantageous under Cue 2 since it carries a lower cost than Test 1. The intercept for this model was significant (OR = 0.17, Z = -3.72, $p < .001$, 95% CI [0.07, 0.43]), indicating a low baseline preference for Test 2 when controlling for the cue.

The selection of Test 3 was strongly predicted by the presenting cue, ($\chi^2(1, N = 119) = 78942.84$, $p < .001$. The odds of selecting Test 3 were slightly lower when Cue 1 was presented compared to Cue 2 (OR = 0.92, Z = -280.92, $p < .001$, 95% CI [0.92, 0.92]). The intercept for this model was highly significant (OR = 0.10, Z = -8031.26, $p < .001$, 95% CI [0.10, 0.10]), reflecting an extremely low baseline preference for Test 3 when controlling for the cue. This suggests that Test 3 was generally less preferred, regardless of the presenting cue.

Finally, for Test 4, the selection was significantly influenced by the presenting cue, ($\chi^2(1, N = 119) = 5.43$, $p = .020$. The odds of selecting Test 4 were 1.18 times higher when Cue 2 was presented compared to Cue 1 (OR = 1.18, Z = 2.33, $p = .020$, 95% CI [1.03, 1.35]). The intercept for this model was significant (OR = 0.05, Z = -6.06, $p < .001$, 95% CI [0.02, 0.13]), indicating a very low baseline preference for Test 4 when controlling for the cue. This suggests that Test 4 was also less preferred overall, though its selection increased slightly under Cue 2.

The observed decrease in selection for Test 1 and the corresponding change in performance for Test 2 under different cues imply that participants were not only sensitive to the diagnostic ability of the tests but also to the associated costs. This cost sensitivity is consistent with the idea that participants optimized their choices by balancing the diagnostic ability of the tests against the effort or resources required to complete them. Together, these findings highlight the nuanced decision-making strategies employed by participants, reflecting both diagnostic accuracy and cost sensitivity in their test selections.

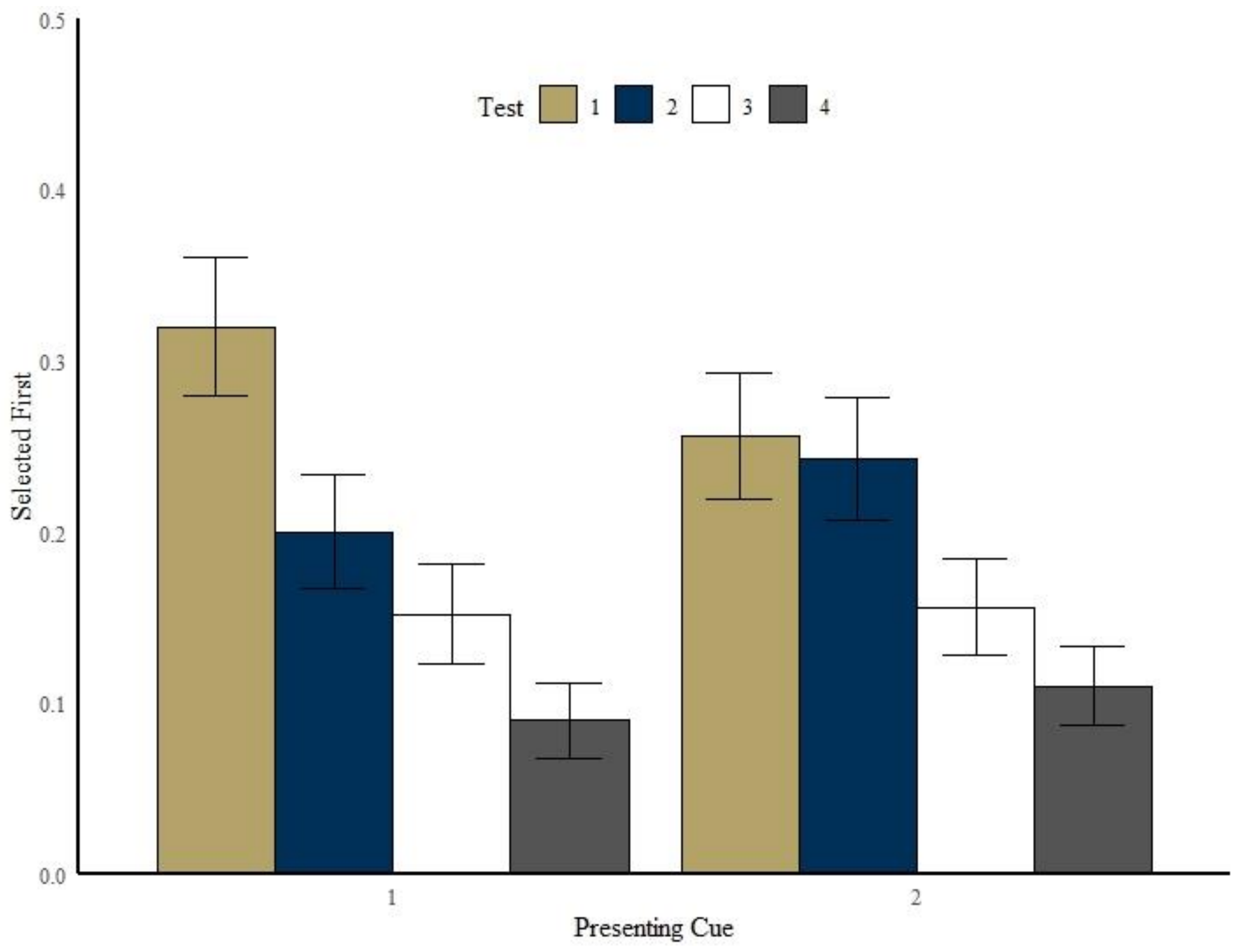

Figure 6- First Test Selected Based Upon Presenting Cue

To examine the effects of the presenting cue, test, and their interaction on the first test selected, a generalized linear mixed model was constructed with participant as a random effect. The results revealed a significant main effect of test, $\chi^2(3, N = 119) = 1242.07$, $p < .001$, and a significant interaction between presenting cue and test, $\chi^2(3, N = 119) = 92.06$, $p < .001$. The main effect of presenting cue was not significant, $\chi^2(1, N = 119) = 0.06$, $p = .813$. These findings indicate that participants' first test selection was strongly influenced by test identity and that the pattern of first test selection depended upon which cue was presented.

Specifically, the odds of selecting Test 2 first were 1.85 times higher under Cue 2 compared to Cue 1 (OR = 1.85, 95% CI [1.61, 2.12], $p < .001$). Similarly, the odds of selecting Test 3 first were 1.46 times higher under Cue 2 compared to Cue 1 (OR = 1.46, 95% CI [1.26, 1.69], $p < .001$), and the odds of selecting Test 4 first were 1.78 times higher under Cue 2 compared to Cue 1 (OR = 1.78, 95% CI [1.51, 2.10], $p < .001$). Conversely, Test 1 was less likely to be selected first when Cue 2 was presented compared to Cue 1. This pattern mirrors the results observed in the overall test selection analysis and further supports the hypothesis that participants adjusted their initial test selections based on the diagnostic value and associated costs of each test. These findings suggest participants engaged in search strategies that were hypothesis guided and sensitive to cost from the start of the information search process.

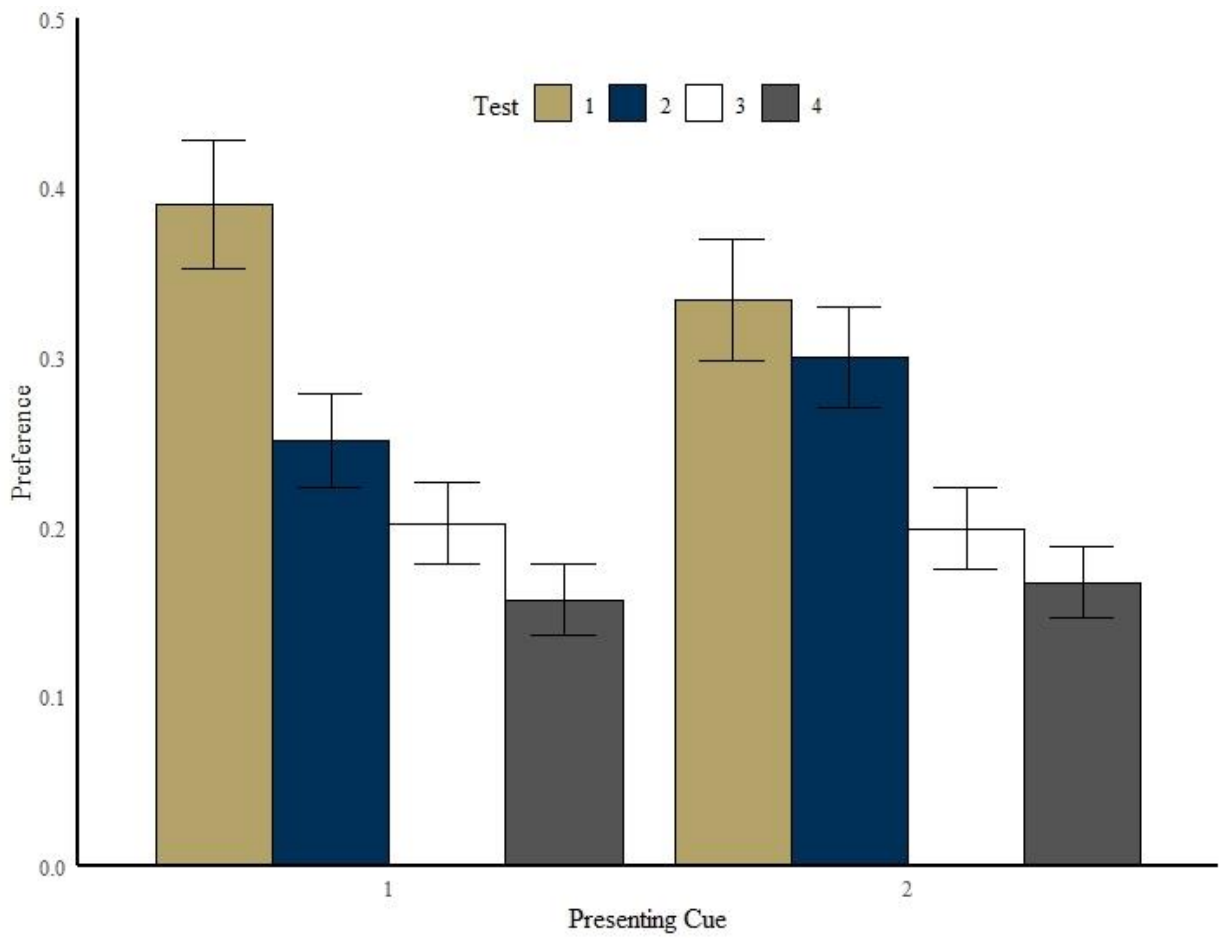

Figure 7- Participant Test Preference based upon Presenting Cue

A preference score was created to account for the order in which tests were selected in each trial. Order was reverse scored so that tests selected first were scored 4, tests selected second were scored 3, tests selected third were scored 2, tests selected fourth were scored 1, and unselected tests remained at 0. These values were normalized so that the total preference across the four tests summed to 1 (Rehder & Hoffman, 2005). A general linear model is used instead of generalized, because the preference score is assumed to be interval data. The model showed significant main effects of presenting cue, test, and their interaction, indicating that participants adjusted their test preference scores based on both the presenting cue and the identity of the test.

Bonferroni-adjusted post-hoc analyses further demonstrated that participants assigned higher preference scores to tests with greater diagnostic value and lower costs under each presenting cue, providing additional evidence of cost-sensitive, hypothesis-guided search behavior.

Researchers further investigated participants search behavior by analyzing search termination behavior and how those decisions related to task performance. Researchers calculated the objective maximum Bayesian posterior belief each participant's individual trials, based on the presenting cue and the test outcomes observed during their search behavior.

The calculated Kendall's τ highlighted the relationship between participants' mean accuracy, mean number of tests selected, and mean maximum posterior belief. Accuracy and the total number of tests selected ($r = 0.22$, $p = 0.00073$) as well as accuracy and maximum posterior belief ($r = 0.42$, $p < 0.00001$) had significant and positive correlations indicating that participants who had higher accuracy had an increased number of tests selected and a higher posterior belief.

Participants search termination behavior was also related to posterior beliefs. Posterior gain measures the change in participant belief after the results of the first test selected. Posterior gain was negatively correlated with continued search ($r = -0.35$, $p < 0.00001$) indicating that participants who had greater posterior gain (increased certainty) were more likely to terminate search following the first test selected, while the participants who were more uncertain continued search.

Researchers also created a simple within-subjects code for the trials where the participants first test selection was confirmatory as well as one for where the first test was non-confirmatory. The first test being confirmatory indicated the participant would likely have increased belief in the disease they associated with the presenting cue while a non-confirmatory first test indicated a lack of increased belief. A paired-samples t-test showed that participants had a significant increase in test selection when the first observed result was non-confirmatory ($M = 2.47$, $SE = 0.13$) than when it was confirmatory ($M = 1.90$, $SE = 0.14$), $t(52) = 4.61$, $p < 0.001$, $d = 0.55$. These analyses highlight that increased uncertainty leads to increased search behavior compared to participants who were more certain, confirming that search decisions were influenced by the perceived certainty of the data.

As mentioned in the discussion of noise, consistency is a measure of how reliably participants make the same choice under a specific context. To evaluate how consistent participants were in their test selections within specific decision contexts, we calculated entropy-based consistency scores at three stages of decision-making: first test, second test (conditioned on first test choices), and third test (conditioned on first and second test choices). Higher consistency corresponds to lower entropy (i.e., more predictable responses). These consistency scores were then correlated with diagnostic accuracy at corresponding stages and overall. The formula for entropy is calculated as $\text{Consistency}(x) = 1 - H(x) / \log_2(4)$. Maximum entropy occurs when posterior beliefs are equally distributed across all hypotheses (a flat distribution). Conversely, entropy decreases (inverse entropy increases) as beliefs become more concentrated on some hypotheses over others. The maximum inverse entropy occurs when all probability is assigned to a single diagnosis, reflecting complete certainty (confidence = 1). First-test selected consistency is calculated by grouping data by participant and the presenting cue, then determining the individual variations for the proportion of trials in which each test is selected. For the second test selected, the context is expanded to include the participant's first test selection and its outcome, so that the consistency reflects how consistently participants respond given that initial context. When calculating third-test consistency, the grouping context is further refined by including the second test selection and its respective outcome. This means that the third-test consistency is calculated by finding the maximum proportion of a particular response among trials that are grouped by participant, presenting symptom, first test selection and outcome, as well as the second test selection and outcome.

Entropy-based first-test consistency was significantly associated with diagnostic accuracy at multiple levels. A Pearson correlation revealed a moderate positive relationship between first-test consistency and first-choice accuracy ($r = .24$, $p < .001$), and with overall accuracy ($r = .27$, $p < .001$). The Kendall Tau coefficients supported these findings, ($\tau = .20$ and $\tau = .23$, respectively, both $p < .001$). These results suggest that participants who responded consistently on their initial test selections tended to perform more accurately overall. Second-test consistency, reflecting choice stability given the first decision context, was also positively related to overall accuracy ($r = .15$, $p < .001$) and aggregate consistency ($r = .68$, $p < .001$). However, the correlation with second-choice accuracy was small and only marginally significant ($r = .07$, $p = .13$). Kendall Tau analyses mirrored this trend, with a small but significant association with first-test consistency ($\tau = .20$, $p < .001$) and a weak correlation with second-choice accuracy ($\tau = .06$, $p = .004$). These findings indicate a modest relationship between consistency at the second decision stage and diagnostic accuracy. At the third stage, entropy-based consistency was not significantly related to diagnostic accuracy. Correlations between third-test consistency and accuracy at the third stage were weak and non-significant ($r = .06$, $p = .13$; $\tau = .07$, $p = .04$). Moreover, third-test consistency did not meaningfully correlate with consistency at earlier stages or with overall accuracy. Although this lack of association might indicate that consistency in later stages of the decision process has increased noise or individual variability rather than strategic behavior, the weak correlation is likely due to small sample sizes in this context. Specifically, calculating consistency for the third test requires sufficient sampling of unique contexts, which depends both on participants' consistent behavior and on the nature of test outcome sampling. Consequently, the limited sample size available for calculating third-test consistency results in reduced statistical power of the analyses.

Entropy-based analyses support the conclusion that early decision consistency, particularly at the first test stage, is a robust predictor of diagnostic accuracy. While consistency at the second stage offers some predictive value, consistency at the third stage is largely uninformative, likely due to increased task complexity and reduced trial numbers per unique context. These results reinforce the importance of early, stable decision-making in multi-stage diagnostic tasks.

Overall, the test results indicate that early-stage consistency in test selection is a meaningful predictor of diagnostic accuracy, underscoring the importance of initial decision stability in multi-step reasoning tasks. The decline in predictive power at later stages highlights potential challenges in maintaining consistent strategies as the decision context becomes more complex.

Chapter 4. Discussion

The findings indicate that participants adapted their information search strategies based on cost sensitivity and hypothesis-guided search. As seen in Figure 1, the presenting cue acts as the information provided from the environment and serves as the cue prompting hypothesis generation (cued recall). Participants were sensitive to both the informational value and the cost of diagnostic tests. Participants exhibited behaviors consistent with hypothesis-guided search termination, particularly in how they weighed the diagnosticity against costs.

In the learning phase, participants showed an increase in diagnostic accuracy over time, confirming they learned the relationships between presenting cues and symptoms. A logistic regression analysis supported that participants were more likely to perform with higher accuracy in later trial blocks compared to earlier ones. With each subsequent trial, participants were 0.034 times more likely to make a correct diagnosis, showing learning over time.

Testing phase data showed that participants were sensitive to how informative each test was and engaged in hypothesis-guided search to make correct diagnoses. Results of the experiment suggest that presenting cue reliably predicted test selection. This aligned with predicted probabilities associated with diagnosticity in the probability table. Participants' preference for test one changed in the predicted fashion. Test one was less likely to be selected after presenting cue two was observed rather than presenting cue one. The odds of selecting test two was lower after observing cue one as opposed to observing presenting cue two.

These analyses highlighted the relationship between participants' search behavior and their perceived certainty of the disease hypothesis, as reflected in the posterior belief. This relationship can be confirmed via the posterior beliefs as participants with more uncertainty and lower posterior belief gains persisted with search while those with more certainty will terminate search after the first test selection. These findings provide researchers with key insights into the data-driven nature of search behavior participants adaptively change search termination behavior based on information learned during the experiment.

Future analyses on this data could potentially investigate the relation between participant entropy and learning from Phase 1 of the experiment. Additional analysis could also explore different metrics of consistency to judge the operational validity of entropy as measured in this experiment as well as looking at the relationship between entropy and participants' likelihood of selecting the most informative test for the context.

4.1 Future Research and Limitations

Future research is currently underway at the Georgia Tech Decision Processing Lab to extend the findings of this study. The experiment currently being conducted highlights both changes in the pattern of test selection due to cost sensitivity as well as participant deviations in the number of cues chosen and search termination behavior. These additions address a gap in information foraging and hypothesis generation research as there is a lack of investigation into the role of cost sensitivity in search termination decisions during hypothesis testing. This novel experiment seeks to fill this gap in current literature, hypothesizing that participant sensitivity to diagnostic test costs significantly influences both their test selection preferences and their search termination behavior.

The predicted result of this experiment include the impact of cost sensitivity on search termination as participants facing higher test costs will likely terminate their search earlier. Additionally, test selection patterns should vary based on the cost condition presented to the participants. The findings from this experiment combined with the results of the study outlined in this paper will offer insights into the dynamics of cost sensitive decision making.
Limitations of these studies include potential variability in participants' familiarity with medical diagnostic scenarios, and constraints in generalizing findings beyond the undergraduate student population used.

Appendix

HyGene-Seek Overview

The HyGene model is based on 3 core tenants; first being cued recall in which the environment activates memory traces of associated cues from long-term memory (LTM). Second, task and working memory capacity constraints limit the active processing of cues and the number of hypotheses that can be held in working memory and guide future information search. In general, events and their memory traces are encoded as odd-length vectors with randomly assigned values: 0, -1, or +1, representing patterns of inhibition and excitation. An encoding parameter (L) determines the quality of encoding, where an L of 1 indicates perfect encoding and 0 indicates complete degradation.

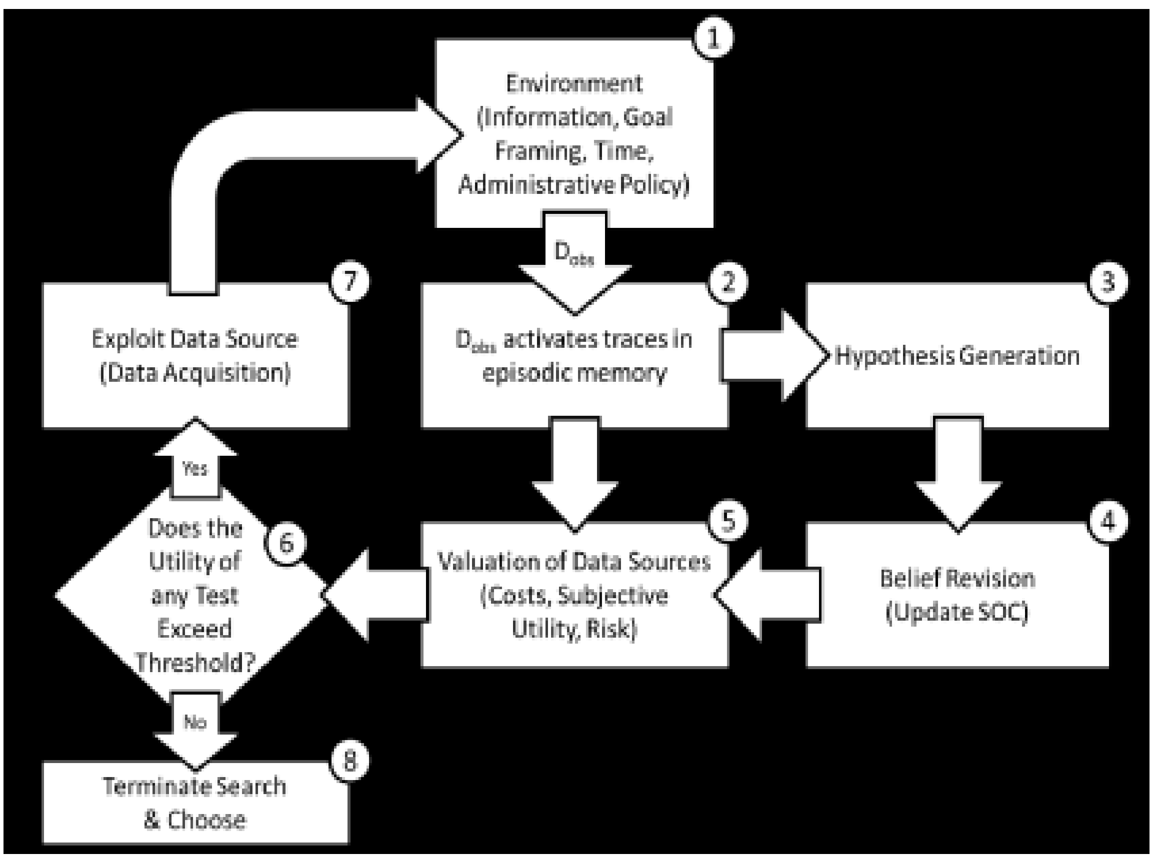

Figure A-1- HyGene's iterative process.

Figure A-1 displays HyGene's iterative process via an example of medical diagnosis. In Step 1, the process begins with external data, Dobs, sampled from the environment. For a physician, this might involve information gathered from a patient's history and symptoms. In Step 2, this data triggers cued recall, activating similar past experiences (memory t in schema or episodic memory). Activation between Dobs and a memory trace (Ti) is calculated as: Ai = (∑j=1Ni Pj × Tij)^3, where *Pj* is the j-th feature of the probe, *Tij* is the j-th feature of the i-th trace, and *Ni* is the number of non-zero overlapping features.Step 3. Memory traces activated above a threshold (Ac) are used to generate an unspecified probe, a composite vector representing the shared features of the activated traces: Cj = ∑i=1K Ai × Tij, where *K* is the number of traces with Ai ≥ Ac. The generation of the unspecified probe is akin to schema extraction where unknown information (e.g., unobserved symptoms, test results, and diagnoses) are extracted into the formation of the composite vector. Thus, experiences as inculcated in the memory system are used to predict the future, including the initiation of the generation of potential diagnoses in the next step. Step 4. The unspecified probe is matched to stored hypotheses in classification or semantic memory. Hypotheses sufficiently activated by the probe are added to the Set of Contenders (SOC) in working memory. These "leading contender hypotheses" forms the decision-maker's working explanation for the data. The probability of a focal hypothesis H_i given Dobs is estimated as: $P(H_i \mid Dobs) = \sum A_i(H_i) / \sum_{j=1_w} A_j(H_j)$, where activations are summed across memory traces for each hypothesis. Step 5a. Most important for the current thesis, the SOC guides hypothesis-driven search for new data. HyGene assumes that hypotheses in the SOC guide the selection of diagnostic cues. Cue diagnosticity is computed using Bayesian principles, considering both the base rate of a cue and its likelihood ratio under competing hypotheses (Bassok & Trope, 1984; Poletiek, 2001):

$D = p(C_A) \times (p(C_A|H_1)/p(C_A|H_2)) + p(C_B) \times (p(C_B|H_1)/p(C_B|H_2))$

To ensure interpretability, the numerator and denominator are flipped when necessary so the ratio always exceeds 1. HyGene approximates this using echo intensities from schema memory rather than objective conditional probabilities: $D = I_{CA} \times (I_{CA|H1} / I_{CA|H2}) + I_{CB} \times (I_{CB|H1} / I_{CB|H2})$. Here, IC values reflect echo intensity (memory activations) for cue values conditional on active hypotheses. The computation of ($I_{CA|H1}$) which corresponds to the numerator of the Bayesian formula is achieved by computing the echo intensity of Cue A (C_A) conditional on the subset of traces in exemplar memory that were activated above the A_c threshold by the H1 portion of the memory probe. All the Bayesian diagnosticity components can be estimated in this manner. Perceived cue diagnosticity can diverge from objective values due to limited SOC hypotheses or biased schema memory sampling (Fiedler, 2001).Step 6. Participants evaluate the utility of further information (e.g., test results) by comparing expected value improvements against a predefined threshold. If expected utility surpasses the threshold (θ), the participant selects a new cue or test and reinitiates the cycle with updated data. If not, search terminates and a final judgment is made. The probability of terminating the search is modeled as: $P(T) = 1 / (1 + g(\Delta ev - \theta))$, where Δev is the change in expected value and *g* governs sensitivity to this change.This iterative process continues as new evidence is encountered and SOC hypotheses are revised. HyGene assumes hypothesis generation ends when time is exhausted or after a set number of failed retrieval attempts (KMAX). This integrated framework connects cue-based hypothesis generation, probabilistic judgment, and search termination. Research by Thomas, Dougherty, Illingworth, and others has shown how belief strength and information costs shape foraging behavior and decision-making (Thomas et al., 2014; Illingworth & Thomas, 2016, 2022).

References

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

Davelaar, E. J. (2014). When the ignored gets bound: Sequential dependencies in ignored and recognized repeated words in a lexical decision task. Journal of Experimental Psychology: Learning, Memory, and Cognition, 40(2), 481–491.

Dougherty, M. R., & Harbison, J. I. (2007). Motivated to retrieve: How often and when do we accurately retrieve weakly encoded memories? Journal of Experimental Psychology: Learning, Memory, and Cognition, 33(5), 1108–1126.

Harbison, J. I., Dougherty, M. R., Davelaar, E. J., & Fayyad, B. (2009). On the lawfulness of the decision to terminate memory search. Cognition, 111(3), 397–402.

Kahneman, D., Sibony, O., & Sunstein, C. R. (2021). Noise: A flaw in human judgment. Little, Brown Spark.

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.

Manktelow, K. I., & Over, D. E. (1990). Deontic thought and the selection task. In W. Jung, H. Singh, & F. Schmalhofer (Eds.), Conditionals, Probability, and Paradoxes (pp. 155–177). Springer.

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

Pirolli, P., & Card, S. (1999). Information foraging. Psychological Review, 106(4), 643–675.

Poletiek, F. H. (1995). Hypothesis-testing behaviour. Psychological Bulletin, 118(2), 251–270.

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

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

Dougherty, M. R., Harbison, J. I., & Davelaar, E. J. (2014). Optional stopping and the termination of memory retrieval. Current Directions in Psychological Science, 23(5), 332-337.

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

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

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.

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

Harbison, J. I., Dougherty, M. R., Davelaar, E. J., & Fayyad, B. (2009). On the lawfulness of the decision to terminate memory search. Cognition, 111(3), 416–421.

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

Illingworth, D. A. (2020). Hypothesis-Guided Testing Behavior: The Role of Generation, Metacognition, and Search. [Doctoral dissertation, Georgia Institute of Technology]. ProQuest Dissertations Publishing.

Illingworth, D. A., & Thomas, R. P. (2015). Price as information: Incidental search costs affect decisions to terminate information search. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting (Vol. 59, No. 1, pp. 225–229). SAGE Publications.

Illingworth, D. A., & Thomas, R. P. (2022). Strength of belief guides information foraging. Psychological Science, 33(3), 453–463.

Lange, N. D., Davelaar, E. J., & Thomas, R. P. (2013). Data acquisition dynamics and hypothesis generation. Cognitive Systems Research, 24, 9–17.

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, 171–176.