ABSTRACT

Title of dissertation: UNDERSTANDING INFORMATION USE IN MULTIATTRIBUTE DECISION MAKING

Jeffrey S. Chrabaszcz, Doctor of Philosophy, 2016

Dissertation directed by: Professor Michael R. Dougherty
Department of Psychology

An inference task in one in which some known set of information is used to produce an estimate about an unknown quantity. Existing theories of how humans make inferences include specialized heuristics that allow people to make these inferences in familiar environments quickly and without unnecessarily complex computation. Specialized heuristic processing may be unnecessary, however; other research suggests that the same patterns in judgment can be explained by existing patterns in encoding and retrieving memories. This dissertation compares and attempts to reconcile three alternate explanations of human inference. After justifying three hierarchical Bayesian version of existing inference models, the three models are compared on simulated, observed, and experimental data. The results suggest that the three models capture different patterns in human behavior but, based on posterior prediction using laboratory data, potentially ignore important determinants of the decision process.

UNDERSTANDING INFORMATION USE IN MULTIATTRIBUTE DECISION MAKING

by

Jeffrey Stephen Chrabaszcz

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2016

Advisory Committee:
Professor Michael Dougherty, Chair/Advisor
Professor D.J. Bolger, Dean's Representative
Professor L. Robert Slevc
Professor Tracy Riggins
Professor Alexander Shackman

Dedication

For and despite my son, Nicolas James Chrabaszcz.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

A_C	Threshold for subsetting episodic memory traces
β	In HyGene, activation of a cue
C_C	Conditional echo content vector
Δ	In ΔI, size of credible cue value difference
ΔI	Delta Inference
γ	Probability of choosing counter to TTB
$\mathcal{L}$	Learning rate
S	Similarity between a memory trace and probe
S_A	Semantic activation
w	Relative weight parameter for combining CV and DR
CV	Cue Validity
DIC	Deviance Information Criterion
DR	Discrimination Rate
GCT	German cities task
HyGene	Hypothesis Generation
log $\mathcal{L}$	Logarithm of the likelihood
MCMC	Markov chain Monte Carlo
SOC	Set of leading contenders
SSL	Strategy Selection Learning
TTB	Take-the-Best
WADD	Weighted Adding

Chapter 1: Introduction

An inference task in one in which some known set of information is used to produce an estimate about an unknown quantity. People make inferences all the time: Any judgment based on indirect information about an outcome requires an inference. Inferences guide important decisions. Which stock is more likely to increase? Which applicant is more likely to improve a business? Better understanding of inferences would allow us to influence and improve such decisions. Psychologists have been interested in this problem for many years. The most recent resurgence in interest, motivated by Gigerenzer and Goldstein (1996), had garnered over 2,400 citations at the time of writing. One important concept from Gigerenzer and Goldstein (1996) is that people use only a subset of the available information when making an inference. Despite the volume of intervening research, the field is still in disagreement over the precise process used to select the information used in a given inference. This dissertation compares three hierarchical Bayesian implementations of existing models of inference. Comparing these three models gives insight into the problem of selecting information, both by demonstrating the effect this has on the ultimate inferences and by illuminating the differences between existing theories.

One of the first models designed to account for how people make inferences of

the type described above was the lens model proposed by Brunswik (1952, 1955). In the lens model, knowledge about a judgment is encoded as a set of discrete cues. To make an inference, relevant cues are weighted by importance and combined, akin predictions with a linear model. Figure 1.1 illustrates the lens model. The distal variable on the left is the quantity that a person intends to predict. The distal variable is unknown by the decision maker, who instead has access to information about the distal variable via proximal-peripheral cues. These cues are in turn related to the distal variable by ecological validities, correlations between cues values and the distal variable. While rational decision makers would combine the cues according to their ecological validities to form a central response (i.e., make an inference), individuals are not always perfectly calibrated to a given environment. People combine the cues according to utilization weights, simultaneously bringing all available information to bear on a particular inference. Correspondence between inferences and the environment is indexed by the functional validity. Distal variables are not necessarily determined by cues, so even perfect cue utilization could lead to decision errors.

As an example, imagine a person is asked to choose which of two doctoral students is more likely to graduate. Probability of graduation is an unknown quantity, but is likely related to some observable, intermediate information like number of publications, grade point average, and the advisor's number of previously graduated students. Assuming the person in question uses the lens model, he or she would combine these pieces of information about each student, weight them according to their utilization weights, and claim the higher weighted sum (Student B) has

a higher probability of graduating (Table 1.1).

	Cue			
Student	Publications	GPA	Advisor	Weighted Sum
A	5	3.2	5	4.5
B	4	4.0	10	5.2
Weight	.5	.3	.2	

Table 1.1: Two fictional students and predictors of their probability of graduation.

One limitation of the lens model is that it does not accommodate information processing constraints imposed on the decision maker. For example, in many real world decision tasks, the decision maker is required to retrieve decision relevant information from memory. The output of this memory retrieval process, and the inherent limitations of working memory, therefore constrain what information a decision maker brings to bear on any particular decision.

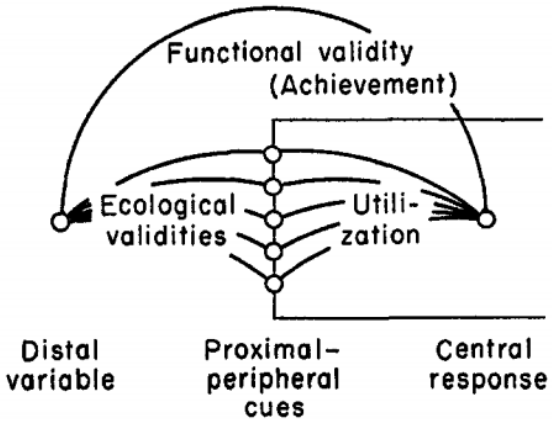


Figure 1.1: The lens model, from Brunswik (1952).

The lens model is just one of many weighted adding (WADD) models of inference (Anderson, 1990; Hammond, 1990). The class of weighted adding models all assume that cue values are multiplied by some set of values and summed to produce

an estimate of the judged outcome but differ primarily in the way cue weights are calculated. This view of cognition is prevalent even outside of psychology, (Yoon and Hwang, 1995; McCloskey, 1998). Meehl (1954) showed that perfect application of the weighted sums of cues, calculated by regressing a distal variable on relevant predictors, out-performs human clinical judgments. Many follow-up studies demonstrate that a variety of factors, including the type of environment, the availability of feedback, and the time spent learning, can affect the application of accurate cue weights (Karelaia and Hogarth, 2008).

One major limitation with this class of models is the computational demand placed on the individual. Cognitive demands for WADD models include: the needs to retrieve information about the decision environment from memory with very high accuracy (Gigerenzer et al., 1991), the need to compute cue validities (Dougherty et al., 2008), and the need to aggregate information across multiple cues (Newell, 2005). People also operate under scarcity, both of the time they have to gather information and of the resources available to remember and process information relevant to a decision. Though regression weights are the informationally optimal way to aggregate information in WADD, it may require substantial cognitive processing to calculate regression-equivalent weights (though see Chater et al., 2003). Alternative models reduce the computational burden of calculating and combining cue weights. Dawes (1979) showed that even improperly specified linear models, which preserve the valence of cue weights but mis-specify the exact weight, perform better than human judges. While improper linear models reduce the computational burden of cue weight generation, they still entail exhaustive search of relevant information.

The exhaustive search process on its own may exceed human cognitive capacity.

Herbert Simon argued that rationality for an actor with limited resources is bounded rather than absolute (Simon, 1955). A decision maker will always have competing goals. At some point the cost of increasing predictive accuracy relative to one goal will interfere with a competing goal. Imagine buying a car, which is a large investment and important to many people for a variety of reasons. Though finding an optimal car is important, finite time can be spent researching different makes and models of car, to say nothing of locating a specific, desired car that is available for purchase. Instead, Simon's bounded rationality suggests that a buyer will find a car that meets his or her needs well enough while also limiting the time spent researching and locating a car. Since then, a plethora of research shows that people do fail to maximize the single goal of accuracy or utility. In the laboratory, participants satisfice rather than maximize expected returns when buying information in an incentive-compatible economic decision task, (Bowen and Qiu, 1992; Fellner et al., 2009). These effects are reflected in affective responses to decisions. Self-reported maximizers, compared with satisficers, have more regret and are less satisfied with economic game outcomes (Schwartz et al., 2002); maximizers also report worse life outcomes than satisficers, (Parker et al., 2007). People seem to make social judgment based on heuristics, (Rand et al., 2014), which are satisficed equivalents of moral rules, (Gigerenzer, 2010). Computer simulations show that satisficing can lead to good performance in economics games, (Stirling and Goodrich, 1999).

The concept of bounded rationality has been applied to models of inference.

Gigerenzer and Goldstein (1996) proposed that people have a toolbox of fast and frugal heuristics that are adapted to different decision environments. Their theory reconciles the concepts of bounded rationality, (Simon, 1955), and ecological validity, (Brunswik, 1955), by assuming that people match their computationally efficient heuristics to the current decision environment, (Gigerenzer et al., 1991; Payne et al., 1988, 1992). According to this theory, individuals select among available decision heuristics and apply one that fits the inferential environment while minimizing cognitive demands when confronted with the need to make an inference. Perhaps the most widely studied among these fast and frugal heuristics is take-the-best (TTB).

Take-the-best begins with a structure similar to the lens model. Information relevant to an inference is organized into discrete cues. TTB differs from earlier rational models at this point, however; the cues are then ordered by cue validity (CV) rather than combined to produce a central response. Instead of bringing all information to bear on a given problem, TTB searches sequentially through cues and makes an inference on the first discriminating cue. For example, the earlier inference about which student has a higher chance of graduating can be made with TTB. Assuming the weights are CVs, TTB starts with the most valid cue (number of publications) and compares the values for each alternative. In this example, Student A has published more papers, so TTB would infer that Student A is more likely to graduate than Student B. TTB is frugal relative to WADD because it can potentially ignore most of the relevant cues. Only when alternatives are tied on the first cue will TTB utilize additional information. If both students had published the same number of papers, TTB would look to see whether one had a higher GPA.

While other cue quality metrics could be used to order cues, the most commonly assumed method of cue ordering is to search through cues based on CV. CV is the probability that an alternative has a higher criterion value given that it also has a higher cue value (Gigerenzer and Goldstein, 1996), and contrasted with discrimination rate (DR), the probability that two objects will have unequal criterion values given that they have unequal cue values.

$$CV = p(A > B | cue_A > cue_B) \tag{1.1}$$

$$DR = p(A \neq B | cue_A \neq cue_B) \tag{1.2}$$

CV is practically bound between 0.5 and 1, since cues with CV of less than 0.5 are reverse-coded. In our graduating student example, if advisor's number of previous students had a CV of greater than 0.5, the reciprocal would have a CV of less than 0.5, indicating that among pairs of students, the one with an advisor who advised fewer students is more likely to graduate. DR gives the probability that a cue could be used to discriminate between two alternatives. DR does not specify that an alternative will have a higher criterion value contingent on cue value, only the probability that a cue/criterion pair are not tied. For the dichotomous cue values assumed in TTB, DR is bound between 0 and 0.5 and is negatively correlated with CV.

Many factors influence the application of TTB and other one-reason decision making algorithms. Martignon and Hoffrage (1999) show that a non-compensatory environmental structure allows TTB to approximate the accuracy of WADD despite purportedly requiring fewer computational resources. Ordered search produces mod-

els that are non-compensatory; no combination of later cues can compensate for the decision implied by an earlier discriminating cue. Consequently, ordered search models work particularly well in environments where potential cues are highly correlated or where one strong cue overwhelms the explanatory power of other cues (Lee and Zhang, 2012; Todd and Dieckmann, 2004). Thus, TTB performs best in environments where the first cue is most important, either because it explains more variance in the criterion than any combination of the following cues or because it partially encodes the same information as other cues. Empirical evidence shows additional constraints on decision heuristic use. Participants in a laboratory task are more likely to use TTB when additional information is costly, when cue validities are explicitly given rather than learned by trial-and-error, and when the environment is deterministic (Newell and Shanks, 2003). According to evidence from eye-tracking, participants use an ordered search heuristic with easily-accessible cues but use a compensatory strategy when cue information is more difficult to retrieve, (Platzer et al., 2014).

Even experiments assuming a weighted-adding model for information aggregation can suggest satisficing. Newell et al. (2009) tested participants in a four-cue decision environment with the goal of predicting whether share price for an unknown company would increase based on the values of those four cues. After each decision, participants got feedback on the trial based on their assigned conditions. In one condition, participants saw the probability of the share price increasing based on the observed cue pattern. The other condition simply saw a message stating whether the share price did or did not increase for that trial. The long-run fre-

quencies of the dichotomous messages from the latter condition matched the stated probabilities seen in the former condition so that information was held constant between conditions; only the format of the information differed. While this manipulation was sufficient to cause better performance for the probability-feedback group, the advantage disappeared in a second study. Rather than providing feedback after each trial, participants saw either probabilistic or dichotomous information for each cue at each trial. That is, for each cue, participants saw either the probability of share price increasing for the current value or the cue, or saw a dichotomous increases/decreases paired with each cue value. Though the actual probabilities contained more information than the dichotomies that would enable higher performance is used rationally, both conditions performed at approximately the same level. This pattern of results suggests that telling participants the unit value of a cue prior to learning was sufficient to remove any effect of metric feedback on cue utilization. The environment for this series of studies could be predicted with relatively high accuracy (80%) using only unit weights for each cue, so further specification of cue weights may not have justified the additional effort to discover relative cue weights. The use of unit weights is rather simple, requiring only tallying of cues in favor of each alternative. Use of specific cue weights, on the other hand, required adding multiple rational numbers together, a process that may be difficult for individuals even without the added, implicit time pressure of the task. Participants may have been inclined to achieve satisfactory performance by using the simpler rule, since the more complex rule yields only slightly higher accuracy and is much more taxing to implement. This study reveals either a preference for a simple decision rule that

involves tallying unit weights for cues (Dawes and Corrigan, 1974; Dawes, 1979), or shows that participants satisfice by ignoring additional information having achieved sufficient performance by their individual standards.

Some evidence suggests that TTB is a viable model of human inference. Hogarth and Karelaia (2007) investigated a variety of decision models (including TTB and WADD) across a range of decision environments. They found that predictive success of different decision models depends on the structure of the underlying ecology. Assuming a toolbox of decision rules, accuracy is maximized by applying tools that match the ecology. For example, they find that TTB is more accurate than alternate models in non-compensatory environments. Another study showed that TTB was among the multiple decision tools necessary to capture variability in participant responses in a decision task. Cognitive toolbox models have been criticized for allowing unlimited flexibility – a researcher can always add another tool to account for unexplained variance or patterns in decisions (Glöckner et al., 2010). Scheibehenne et al. (2013) fit hierarchical Bayesian models to data from a number of different experiments to validate the idea of a cognitive toolbox and to demonstrate a method of comparing toolbox models while accounting for flexibility in inferences. They find that, across many of these experiments, the combination of TTB and WADD provides the most parsimonious account of participant choices.

Despite the normative success of fast and frugal heuristics, they may not describe the actual decision process that people use. Hilbig et al. (2010) proposed the nonsensical alphabet heuristic and showed that its use in a decision task comparing city populations produced results comparable to the recognition and fluency heuris-

tics. He argued that similar performance between a heuristic and participants on a task is insufficient to claim that participants use that heuristic. Instead, other aspects of the decision process must be considered. In TTB, for example, participants must demonstrably search cues in the same order as TTB and also make similar inferences in order to claim that TTB is applied. Despite TTB's simplicity, Dougherty et al. (2008) criticize the calculation of cue validity as implausible. They argue that the an automatic event-counter is unsupported by existing evidence in frequency encoding and that cue validity requires people to remember the absence of information, conflicting with logic and memory research.

Some researchers argue that TTB is part of an ecologically rational toolbox of decision algorithms that are selected and used as a function of fit to the environment. While the contents of the toolbox are debated, most of these proposals include some mixture of simultaneous and sequential search tools that vary the use of cue weights. For example, people could combine all information (simultaneous search) either by weighting cues by their CV (WADD) or simply by aggregating cues with unit weights (Dawes, 1979). Though TTB requires search of cues ordered by their decreasing CV, search could also be in random cue order to limit the computational burden of calculating cue orders (Gigerenzer and Goldstein, 1996; Gigerenzer and Todd, 1999).

A number of studies alter the proposed WADD and TTB models originally compared in Gigerenzer and Goldstein (1996). One proposed change is allowing a probability of guessing rather than applying the specified model (Bergert and Nosofsky, 2007; Lee and Newell, 2011). This change accounts for the high variability in

subject responses to most behavioral tasks. No single decision algorithm captures the responses made by participants in observed decision environments, in part because participants appear to be inconsistent in applying a given decision rule.

Another change alters the function used to weight or order cues. Newell et al. (2004) show superior prediction in many environments using success, a cue value metric defined as CV times DR, added to the product of one minus the discrimination rate times the probability of a correct choice when guessing.

$$Success = CV \cdot DR + \frac{1}{2}(1 - DR) \tag{1.3}$$

This method combines the probability that a cue will discriminate between alternatives with the probability of guessing and the probability of choosing the correct alternative given that a cue discriminates, producing an aggregate measure of single-cue usefulness.

Others suggest meta-heuristics to choose among heuristics in the toolbox for application in a given environment. Strategy Selection Learning theory (SSL) is one framework for comparing the use of simultaneous and sequential information search (Rieskamp and Otto, 2006). SSL claims that individuals learn which decision heuristic is best fit to an environment based on repeated feedback. One strength of SSL is that it encodes the heuristic toolbox, fully accounting for the flexibility of allowing many possible decision heuristics. In SSL, each decision heuristic is fully encoded as a model. SSL assumes that, over a number of learning trials, participants compare the predictive accuracy for each model and selectively reinforce models that provide higher accuracy within an environment. When a simulated learner

is allowed to learn to apply either WADD or TTB over repeated feedback trials, SSL yields higher accuracy (measured by percent correct) than WADD, TTB, or a memory-like categorization model. This evidence has been expanded to suggest that certain environmental characteristics occupy a cognitive niche that predispose decision makers to use one of a number of decision heuristics (Marewski and Schooler, 2011).

Based on the evidence reviewed above, it is obvious that there are competing models of how cues are generated and ordered in the context of inference tasks. In what follows, I outline three contemporary models of this process, which will then serve as the basis for the remainder of this dissertation.

1.1 Search

Cue metrics like CV and DR, or aggregation methods like TTB and WADD, may occupy the ends of two spectra used in decision making heuristics. While the success metric gives an optimal method for combining CV and DR, participants show variability in preferred search order that may be related to preference for valid or discriminating cues. Similarly, people may apply WADD or TTB in the same decision environment. Lee and Newell (2011) developed a pair of hierarchical Bayesian to describe individual differences in cue ordering and search termination called Search and Stop. Search and Stop are a complimentary pair of models that determine the order and number of searched cues based on compromises between CV and DR and between all-reason and one-reason decision making. The Search

model assumes that participants order cues and make inferences similar to TTB. At the participant level, Search includes two parameters that distinguish it from TTB: γ and w. The γ parameter indexes the probability of choosing counter to the model prediction. TTB is normally a deterministic model, participants are assumed to choose whichever alternative has an earlier discriminating cue. Allowing for errors in applying the TTB model can reduce the penalty of incorrect estimates for participants that choose inconsistently or are otherwise poorly fit by the TTB model. The w parameter allows participants to weight CV and DR:

$$\text{weight} = CV \cdot w + (1 - w) \cdot DR.$$

Participants then search cues by ranking weight from largest to smallest. The Search model also includes hyperparameters for the mean and standard deviation of w. Partially pooling estimates of w in this case simultaneously improves individual estimates of w and summarizes the individual differences in search order. The Search model in isolation assumes that participants apply an error-prone TTB to make decisions but allows for individual differences in search order.

The Stop model captures differences between application of TTB and WADD by participant while assuming a fixed search order for cues. The model assumes that participants either apply TTB with probability θ or WADD with probability $1 - \theta$. Though more complicated than either WADD or TTB, a comparison of a modified Stop model with SSL shows that the complexity of SSL is almost never justified. Based on minimum description length, an information-theoretic measure of model complexity that balances fit and parsimony, Stop is preferred to stochastic

WADD and TTB across a range of plausible error rates (Newell and Lee, 2011). A similar comparison has not been performed for the Search model with alternative models of stochastic search order.

One way Search and Stop prevail is by allowing individual differences in decision strategy. By including hierarchical structure allowing cue ordering and model selection to vary by participant, the models account for variation that could otherwise be attributed to inconsistency in decision heuristic application. Compared with SSL, Stop is able to model individuals varying propensity to choose randomly or to prefer additional information. Stop also avoids the problem of fully encoding both TTB and WADD separately by adding a parameter to generalize between the alternative heuristics. Search and Stop are only tested on a single environment at a time, however, and can only account for individual differences in weighting of CV and DR or WADD and TTB. If decision making varies along any other dimensions, the Search and Stop models will be insufficient.

1.2 Delta Inference

The original conception of TTB operates on dichotomous cues, so all cue comparisons are either equal or differ by one. Luan et al. (2014) proposed Delta Inference (ΔI) as an elaboration on TTB that allows for continuous cue values. This slight change alters the potential flexibility of TTB. TTB operates by choosing an alternative based on a single discriminating cue, regardless of the discrepancy between cue values for the alternative choices. In DI, the stopping rule in TTB is amended to

stop cue search only when cue values for alternatives differ by more than a certain amount, Δ. This potentially allows search to continue despite a discriminating cue, when the difference between cue values is smaller that Δ, allowing some flexibility to accommodate compensatory ecological structures. Note that this is still one-reason decision making. While mere difference may not be sufficient to motivate a decision in ΔI, that cue has no bearing on the decision process during later cue consideration. This is different from a change like that in the Stop model, which weights the number of cues to aggregate in a decision (Lee and Newell, 2011) or another model which assumes that both TTB and WADD are accessible tools (Newell and Lee, 2011; Scheibehenne et al., 2013; Rieskamp and Otto, 2006).

Though (Luan et al., 2014) show that a Δ of 0 is best on average, they do not explore the fitting of Δ for subjects, environments, or cues. ΔI is also not compared to human performance, so its predictive validity for real decisions with non-zero Δ parameters, (where $\Delta = 0$ is equivalent to TTB), is unknown.

1.3 Hypothesis Generation

Another decision modeling framework comes the Hypothesis Generation (HyGene) model (Thomas et al., 2008). While not a decision making model *per se*, HyGene is a model of memory search based on MINERVA-2 (Hintzman, 1984). In addition to being consistent with memory research, HyGene has accurately modeled other psychological phenomena, including subadditivity (Dougherty et al., 1999) and visual search (Buttaccio et al., 2015). HyGene requires little substantial alteration

to produce decisions on a paired comparison task. Modifying HyGene to make inferences provides an opportunity to evaluate existing decision rules in the context of a plausible theory of memory, (which is missing in Search and violated by most fast and frugal heuristics).

In the HyGene model, memory is divided into episodic memory, semantic memory, and working memory. Episodic memory contains a memory trace, a vector of features taking the values 0, -1, or 1, for each event or experience. The traces in episodic memory are subject to degradation through forgetting and interference, governed by a learning rate, $\mathcal{L}$. Traces in episodic memory also encode frequency information in the environment: Events that occur and are encoded more frequently appear proportionally more often than less frequent events. In contrast, semantic memory encodes each potential outcome only once, regardless of the frequency of any individual event. For example, an emergency room doctor is likely to have diagnosed influenza much more often than smallpox. The doctor's episodic memory would contain a large number of feature vectors corresponding to influenza and few if any for smallpox, but each would appear a single time in semantic memory. In HyGene, working memory is a constraint on the number of semantic memory traces that can be considered as hypotheses at one time.

Hypothesis generation is motivated by a probe, or an event about which a hypothesis is necessary. In this recent example, the symptoms of a patient would act as a probe. The HyGene model assumes that people decide among hypotheses by first probing episodic memory to determine similarity between each event and the probe. Mathematically, this is accomplished by calculating the dot product between

the probe and each memory trace:

$$S_i = \frac{\sum_{j=1}^{N} P_j T_{ij}}{N_i}, \tag{1.4}$$

where P and T are a probe and trace of length N and i indexes the number of traces in episodic memory. The cube of similarity (S_i^3) is then compared to A_C, the latter being a free parameter in the model. A_C acts as a cutoff similarity value to limit search of memory to relevant traces. All traces with cubed similarities higher than A_C are used to generate a hypothesis, while all traces with lower similarities are ignored for subsequent calculations.

After identifying the relevant subset of memory, HyGene creates a conditional echo content vector (C_C) using the following formula:

$$C_C = \sum_{i=1}^{K} S_i^3 T_{ij}. \tag{1.5}$$

Each trace in episodic memory is multiplied by its cubed similarity and the resulting vectors are summed element-wise. The vector result of this process, normalized by dividing all values by the absolute value of the largest value in the vector, is an "unspecified probe" which combines the diagnostic information in the probe with base rate information from episodic memory. The dot product of this unspecified probe and each entry in semantic memory, yielding a semantic activation (S_A) for each semantic trace. Traces with S_A higher than zero then enter the set of leading contenders (SOC), a capacity-limited proxy for working memory. The entire search process: activate a subset of memory, create an unspecified probe, generate semantic activations, and populate the SOC, repeats until a pre-determined number of

iterations. The end result is a short-term store (the SOC) filled with hypotheses about the probe and their associated activations from semantic memory.

The HyGene process requires little change to search for cue orders based on the contents of memory. A probe of each cue value could be used to search memory, returning the short-term buffer's worth of cues that predict the highest values on the criterion value for a given decision environment along with their activations. Activation for each cue should condense DR and CV into a single measure and mimic success or Search model results. Thus, HyGene potentially gives a psychologically plausible method for calculating cue orders. This dissertation includes modeling studies that test this intuition and evaluate the necessity of complicated decision rules and metric derivation beyond memory processes. For example, calculation of CVs and cue ordering in general may be obviated by instead relying on emergent properties of episodic memory search. One might also reduce the stochasticity of WADD and TTB in Search by allowing cue orders to be determined by memory search, which is already a stochastic process.

1.3.1 Summary

Search, Delta Inference, and HyGene are models that seek to explain decision making behavior at different levels of analysis. Search captures abstract, individual differences in search order and error of strategy application. Though initial work with ΔI focused on average values of Δ across environments, ΔI could be adapted to investigate whether individual differences in decision making are limited to differ-

ences in Δ, the difference between cue values necessary to motivate a decision. The HyGene framework contains different restrictions, only containing free (and potentially varying) parameters that are involved in memory processes. These levels of explanation coincide with David Marr's levels of analysis (Marr, 1982). Search exists at the computational level of analysis, focusing primarily the types of information that are required to accurately capture patterns in cue search and judgment. ΔI parameterizes specific components of the algorithm used to combine information, while HyGene focuses on details of the implementation of a decision algorithm in the context of a subordinate memory system. Unfortunately, these models are all evaluated in different ways. Search exists as a hierarchical Bayesian model with parameters that are partially pooled by individual, while ΔI and HyGene are both fit with maximum likelihood that average over individual differences. This dissertation formulates all three models as hierarchical Bayesian models to allow direct comparison across these models and corresponding levels of analysis.

Chapter 2: Bayesian implementations of inference models

Comparing Search, ΔI, and HyGene requires both common implementation methods and shared data. The present chapter includes descriptions of hierarchical Bayesian implementations for models of inference based on Search, Δ Inference, and HyGene. Hierarchical Bayesian modeling allows natural extension to include individual differences, summarizes these differences with fixed parameters, and provides a very general method for relating cognitive models to observed data (Lee, 2010). There is a recurring structure present in all three models under consideration. The w for Search and ΔI, Δ for ΔI, and β for HyGene all vary by participant but are drawn from a distribution with parameters that are fixed across participants. Drawing participant parameters from shared distributions allows the models to bring the maximum information available to bear on estimating each parameter (Lee, 2008), and represents a compromise between fully pooled estimates, which assume that participants are identical, and unpooled estimates, which assume that participants are entirely unique (Gelman and Hill, 2006).

2.1 Search

The Search model, already a hierarchical Bayesian implementation, is drawn almost directly from Lee and Newell (2011). Given the focus on cue ordering, which is the major difference between Search and the comparison models, I ignore the Stop model entirely. The stopping rule in Stop is modeled independent of search order; Search determines search order by individual after which Stop subsequently determines the number of cues searched. These studies focus on how the models order cues differently and whether this influences judgments, though later work could examine the interaction with varied stopping rules.

The Search model is completely described in Figure 2.1. The order of cue search is governed by a weighted combination of CV and DR. The individually-varying relative weight for CV is drawn from a bounded normal distribution with both mean and variance varying as beta distributions with $\alpha = \beta = 1$ and bounds at .01 and .99. The DR weight is 1 minus the CV weight. Based on this balance of CV and DR, the cues are searched sequentially until any difference between the cues allows for the model to stop and choose one of the two alternatives. The model then selects the TTB-chosen alternative with probability of γ or the alternative with probability $1 - \gamma$. This allows the model to account for the fact that human participants very rarely apply a single decision rule consistently. In the event that two alternatives are exactly tied on all cue values, the model chooses between the alternative with an equal probability for each outcome.

In a small departure from earlier work, I have implemented the Search model

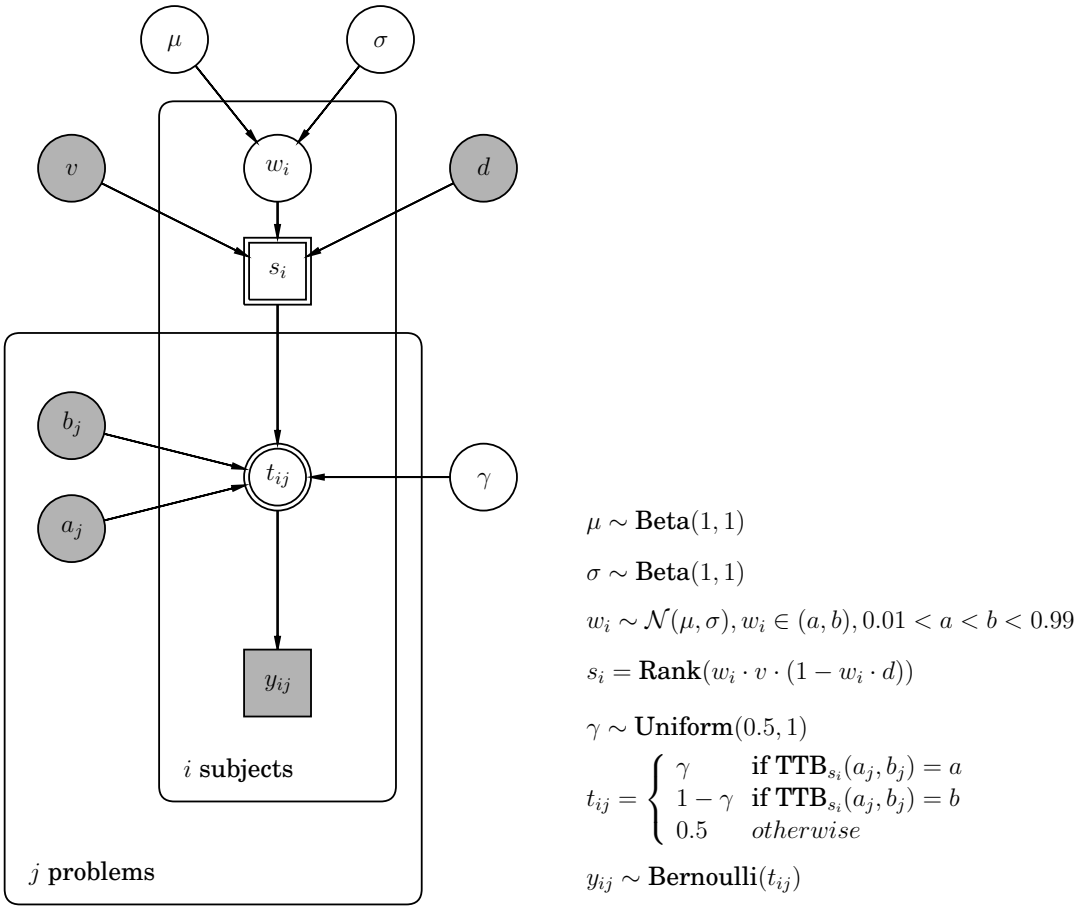


Figure 2.1: Graphical model of Search.

with continuous, rather than dichotomous, cue values. While changing the cue support should not pose any problems for the Search framework, one goal of this dissertation is to assure that changing the support of these cue values does not interfere with the model.

2.2 Delta Inference

For the purpose of this study, the Δ Inference model is a modified version of the Search model including an elaboration that potentially allows the TTB search algorithm to continue past a discriminating cue (Luan et al., 2014). In some cases, a small difference in cue values may not be sufficiently informative to terminate search. The Δ parameter is allowed to vary by both cue and participant, so the model converges on Δ parameters most consistent with the data. While ΔI modifies the stopping rule for TTB, it does so in a way that preserves one-reason decision making. When ΔI makes a decision, it is based only on the value of a single cue; previous cues are treated as ties and ignored. Unlike the Stop model, the stopping rule from ΔI can interact with CV and DR weighting, influencing search order.

Earlier research showed that, on average, the best value of Δ is zero (Luan et al., 2014). This research kept a consistent value of Δ across all cues and individuals in the studies, though, precluding the possibility that individuals or cues might differ in their values of Δ. Varying Δ by person amounts to the suggestion that individuals may differ in the amount of information they require before making a decision; varying by cue allows that cues can be differentially informative. Instead of a consistent value of Δ for all cues and participants, I allow Δ to vary across both of these dimensions. Though the Δs for each cue are independent, I define a hyperparameter for each cue's Δ and allow subject-varying deviations from this average value (Figure 2.2).

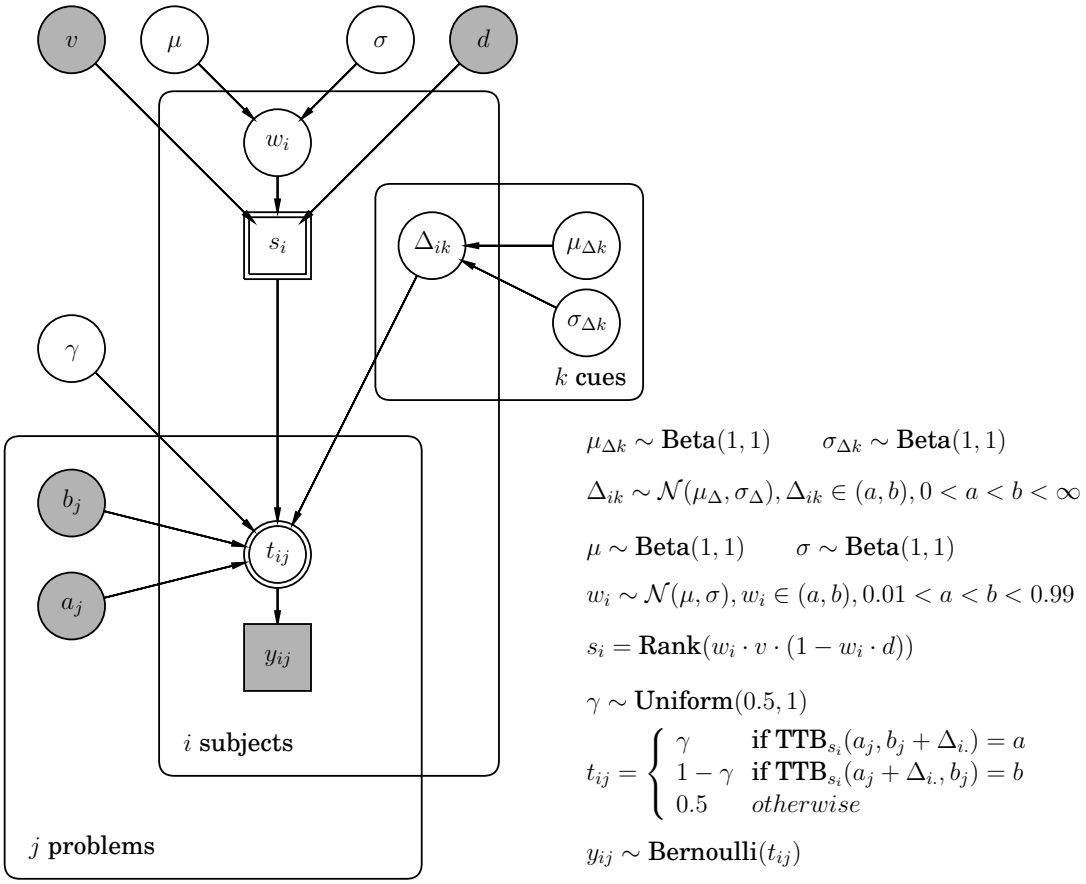


Figure 2.2: Graphical model of Delta Inference.

2.3 Hypothesis Generation

The version of HyGene used in this dissertation is a Bayesian model inspired by Thomas et al. (2008). This HyGene implementation decides between pairs of choices by searching sequentially through cues (as in TTB) and using a minimal difference in cues to terminate search and select an alternative (Figure 2.3). Search order is determined by using weighted logistic regression on scaled cue values for a training set, deemed episodic memory or M in the graphical model, to generate normalized regression coefficients. Cues are searched in descending order of coeffi-

cient magnitude, serving as a proxy for CV and DR as used directly in both Search and ΔI. After determining search order based on the normalized regression weights, HyGene makes decisions like the Search model, complete with a γ parameter for error in application and TTB-like sequential cue use.

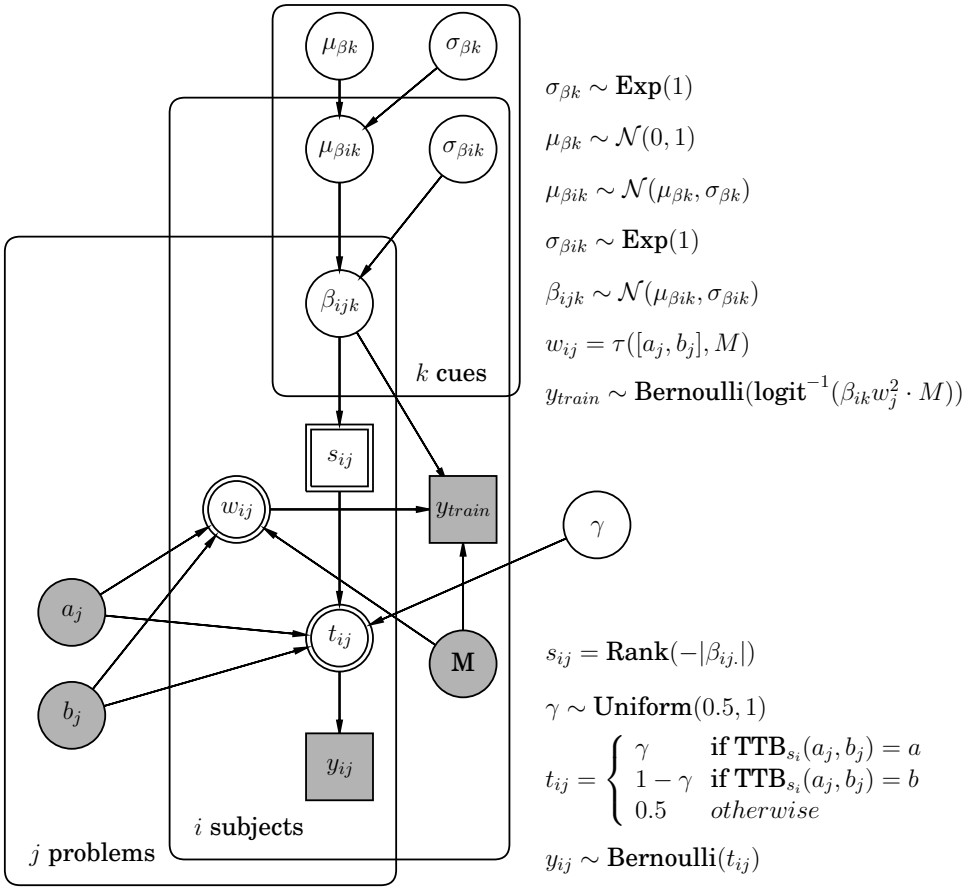


Figure 2.3: Graphical model of HyGene.

There are two major differences between the current instantiation of HyGene and the model specified by Thomas et al. (2008). The first is that the current instantiation uses continuous weighting for episodic memory traces. Rather than using a threshold (A_C) and ignoring memory traces below a modeled or assumed threshold, the current HyGene implementation accomplishes a similar goal by weighting

observations in episodic memory more heavily as a function of the magnitude of ordinal correlation with the cue values for a given observation. The current method of weighted regression should have similar performance to selecting relevant memory traces based on A_C. Assuming a true non-linearity in trace selection, the weighted regression technique in the current HyGene model provides a first-order, smoothed approximation of the discrete, underlying function (Shalizi, 2015). The A_C parameter is is a context-free value without a comparable parameter in Search or ΔI. Inferences from the hierarchical Bayesian are based on fixed parameters and search orders, so removing A_C in favor of continuous weighting of episodic memory improves interpretability of comparisons with Search and ΔI.

The second difference is that this HyGene instantiation is allowed to search all cues in the ecology. The original HyGene model includes a finite SOC, which limits the number of simultaneous activated semantic traces, in this case cues, that an individual can consider. The limited SOC could easily be included in any of the models currently under consideration and could have a variety of difficult-to-predict effects on search order and judgment accuracy. Similarly, nothing about HyGene requires sequential search, the model could aggregate over a subset of cues to produce a response. Given the focus on search order, HyGene is implemented with only the weighting mechanic and uses a TTB-like stopping and decision rule on HyGene-ordered cues to minimize differences from the comparison models.

These decisions about HyGene modeling produce a version of the model that is maximally comparable to Search and ΔI, allowing for a direct inspection of the cue ordering mechanism without interference from other model dissimilarities. The

γ parameter, for example, increases model flexibility by allowing some proportion of responses that violate the direct predictions from sequential search of the cues. Exclusion of this modification, which is present only in Search and not in ΔI or HyGene as originally formulated, potentially conflates the cue ordering mechanism and other, empirically-motivated modeling decisions.

Chapter 3: Cross-validation of models to simulated data

One aspect of comparing computationally-specified theories of decision making is understanding the relative flexibility of these theories. There are at least three ways to validate new modeling methods:

1. analytic proof of behavior in the limit;
2. validation on a standard dataset; or,
3. validation on simulated data.

The first method is intractable in the case of most cognitive process models, the current models included. Closed-form solutions to these models would be difficult both because of the diverse prior specifications and because of the unconventional likelihood statement. The second method will be useful later when an external reference helps explore the usefulness of these cognitive models in naturalistic conditions. These data lack a defined generating process, however; no model can be identified as correct in these circumstances. The only alternative is model comparison, but the best method to compare cognitive models is under debate. Therefore, simulated data will fuel an initial attempt to understand how Search, ΔI, and HyGene relate to one another.

This chapter will focus on two questions. The first is: How do Search, ΔI, and HyGene account for decision processes in simulated environments with known structure? This question is answered with two ecologies that vary in predictive difficulty. The simpler of these ecologies has orthogonal, non-compensatory cues. This means that the cues are uncorrelated with one another and that predictions based on the strongest cue cannot be reversed by any combination of subsequent cues. This first ecology is contrasted with a second ecology that has compensatory cue structure and positive cue intercorrelations.

The second question for this chapter is: How related are the predictions from Search, Δ Inference, and HyGene? Though this question will return regarding data generated by human decision makers, using simulated data allows for direct examination of how structure in the environment is represented in the fixed parameters of each model. Fitting the three models to the same environments also allows for understanding of interactions between the shared components among the models (e.g., γ) and their unique components.

3.1 Methods

The questions in this chapter require both generating structured ecologies and fitting of relevant cognitive models (Search, ΔI, HyGene).

	Ecology 1				Ecology 2			
	Outcome	1	2	3	Outcome	1	2	3
Outcome	1.0	0.5	0.3	0.1	1.0	0.3	0.2	0.1
1	0.5	1.0	0.0	0.0	0.3	1.0	0.2	0.2
2	0.3	0.0	1.0	0.0	0.2	0.2	1.0	0.2
3	0.1	0.0	0.0	1.0	0.1	0.2	0.2	1.0

Table 3.1: Covariance Matrices for both simulated ecologies.

Ecologies

Both ecologies are generated from multivariate normal distributions with all means equal to zero, the defining ecologies differ only in their covariances. The first ecology consists of 20 samples from a covariance matrix with three orthogonal cues and decreasing correlations with the outcome (Table 3.1). Though dissimilar from the empirical data in later chapters, this ecology will provide a reference for all models. The cues in the first ecology are non-compensatory, so both sequential and simultaneous cue use both reach the same conclusions on these stimuli. The strict orthogonality of the cues also makes cue weighting easier, allowing inspection of the relative influence the priors in each model have on cue ordering. This ecology also permits assessment of the influence of individual differences in cue order drawn directly from the priors in each model. Direct fits to any fixed ecology should lead to a consistent search order. Differing search orders in this ecology reflect the influence of prior information in each model.

The second ecology is intended to be closer to empirical data than the first. The cues are poorer indicators of the outcome on average and have non-zero covariances. The set of objects in the ecology are also more numerous, with 100 unique

objects instead of 20 as in the simple ecology. While the simple ecology is intended as a reference distribution to help assess prior influence, the complex ecology serves to foreshadow the success of these models when fitting messier, empirical data. The difference is that this complex ecology still has a known structure. While empirical samples are useful for different reasons, we have no way of knowing the true population structure from which they are drawn.

Each ecology is used to generate three distinct sets of data, one corresponding to each Search, ΔI, and HyGene. The priors from each model are used to generate parameters for 20 imaginary participants. These parameters are then used to produce responses to all paired comparisons of the objects in each of the two ecologies. For each set of generated data, shared parameter values are consistent across models. For example, γ is consistent across all three models and w for each "participant" is the same for both Search and ΔI within an ecology.

Simulations

I produced simulations from each model with each of the two ecologies. For Ecology 1, the training set consisted of simulated model predictions for all 190 unique pairs of stimuli in the 20-object ecology. The training set for Ecology 2 consisted of only a subset of the possible stimulus pairs. Though the second ecology includes 100 objects, the training set included only 100 random pairs of these objects rather than the exhaustive 4,950 pairs. I generated hypothetical participants' responses for each ecology to allow for a representative range of the individual differences for

each model. After simulating these responses, I fit each of the three models using to the generated responses and the same training set.

Analyses of the results both assess model performance and examine how different sources of variability are represented in each joint posterior distribution. Model performance can be examined in a variety of ways. I first present both the likelihood and the penalized likelihood using the Deviance Information Criterion (DIC; Spiegelhalter et al., 2002), which communicates average model effectiveness when accounting for complexity:

$$\text{DIC} = -2 \log \mathcal{L} + \widehat{var}(-2 \log \mathcal{L}). \tag{3.1}$$

For the DIC, complexity is a measure of the range in data that could be fit by the model and is measured as the observed variance in deviances observed in a convergent MCMC model[1]. Model comparison using the raw likelihood may make sense in this context. Human decision making is a complex, and perhaps stochastic, process, so any preference for simpler models, especially in a dataset of such limited size relative to the variation in the decision making system, may be unjustified. One caveat is that these densely-parameterized models allow for individual differences in different parameters for each model. Model comparison using penalized likelihood is especially unstable in these types of models because the likelihood function is very flat and the appropriate penalty is contentious (Weng and Gelman, 2014). In addition to fit statistics (log likelihood and DIC), posterior distributions of the fixed effects are reported for each model. The summarized fixed effects include γ for all

[1]The complexity term, $\widehat{var}(-2 \log \mathcal{L})$, is commonly referred to as the penalty.

models and the means and standard deviations of w for Search and ΔI, Δ in ΔI, and β in HyGene.

A major goal in this chapter is to assess overall model flexibility. Though both Δ Inference and HyGene are instantiated as elaborations on the Search model in this series of studies, it is possible that the additional parameters for these models dramatically change model flexibility. The following analyses take data generated from each of the three models using the two ecologies specified in the methods and then fit each of the three models to this generated data.

All inferences are based on models fit using JAGS (Plummer et al., 2003) and called from the rjags package in R (Plummer, 2015). JAGS is a C++ implementation of a Gibbs sampler. Parameters in each model met minimum $\hat{R}$ and *effective n* diagnostic criteria prior to further analysis. $\hat{R}$ is ratio of between- and within-chain variability, with values larger than 1 indicating poor mixing. Effective n is a measure of MCMC sample size that accounts for autocorrelation between successive samples. While no strict cutoff exists for effective n, a few hundred independent samples is considered sufficient to support inferences from the posterior distribution (Gelman et al., 2014).

	Parameter	Value
Ecology 1	γ	0.9
	CV	0.663, 0.668, 0.511
	DR	1, 1, 1
Ecology 2	γ	0.9
	CV	0.61, 0.69, 0.59
	DR	1, 1, 1

Table 3.2: Fixed parameters for data generation.

3.2 Results

First Ecology

Samples from multivariate normal distributions with the aforementioned parameters yielded CV and DR rates seen in Table 3.2[2]. For both ecologies, the second cue gives the highest CV, followed by the first cue and then the third cue. There are no tied cue values, so all cues discriminate between all paired comparisons and $DR_k = 1$.

The results of fitting Search, ΔI, and HyGene to data generated using Ecology 1 show nearly equivalent performance for Search and HyGene (Table 3.3). ΔI is consistently, if only slightly, better able to account for variability in simulated participant responses as reflected by higher average likelihoods. While counter to earlier research, this suggests that non-zero Δ parameters may be useful in some decision environments. Particularly in ecologies like the one generated, where many cue values are near the mean and explain relatively little variance in the outcome, ignoring small cue differences and continuing search may be more important than a

[2]Values throughout this dissertation are rounded to an appropriate number of significant digits.

simpler model (Search) or a more flexible method of ordering cues (HyGene). Preferring ΔI to Search and HyGene for these data grants that the small differences between likelihoods and DICs are credible and favor ΔI, which may be unjustified given the small differences between model fits.

Data	HyGene		
Model	HyGene	Search	ΔI
log $\mathcal{L}$	−1386	−1385	−1378
Penalty	1.276	0.7953	7.074
DIC	2773	2771	2763
Data	Search		
Model	HyGene	Search	ΔI
log $\mathcal{L}$	−1386	−1387	−1361
Penalty	1.019	0.2386	28.71
DIC	2774	2774	2750
Data	Delta Inference		
Model	HyGene	Search	ΔI
log $\mathcal{L}$	−1387	−1384	−1377
Penalty	0.48	0.9434	8.579
DIC	2774	2769	2762

Table 3.3: Summaries of models fit to data generated from each model using first ecology.

While the patterns in fit quality for Ecology 1 make sense given the continuous cue values, posterior distributions for the parameters that summarize each model suggest poor calibration, (Table 3.2). Though the γ parameter for all data generation processes was very close to 1 (Table 3.2), all models returned a posterior distribution on γ with a median closer to $\frac{1}{2}$. ΔI gives higher values for γ, but nothing close to the true, fixed value of $\frac{9}{10}$. All three models are nearly guessing at the outcome, given that the probability of a model-inconsistent response is nearly as high as a model-consistent response.

Ecology 1				
		Generating Model		
Model		HyGene	Search	ΔI
HyGene	γ	0.52	0.51	0.51
	μ_β	−0.04, −0.06, −0.26	−0.03, −0.04, −0.3	−0.04, −0.04, −0.25
	σ_β	0.27, 0.42, 1.07	0.28, 0.38, 1.13	0.24, 0.39, 1.07
Search	γ	0.52	0.5	0.53
	μ_w	0.53	0.53	0.48
	σ_w	0.7	0.71	0.71
ΔI	γ	0.55	0.6	0.56
	μ_Δ	0.66, 0.2, 0.59	0.69, 0.14, 0.91	0.16, 0.77, 0.2
	σ_Δ	0.2, 0.4, 1.74	0.89, 0.24, 0.13	0.37, 1.24, 0.48
	μ_w	0.51	0.5	0.51
	σ_w	0.7	0.71	0.71
Ecology 2				
		Generating Model		
Model		HyGene	Search	ΔI
HyGene	γ	0.55	0.53	0.55
	β	≈ 0, 0.01, −0.19	≈ 0, ≈ 0, −0.2	≈ 0, 0.01, −0.22
	σ_β	0.22, 0.31, 0.76	0.21, 0.27, 0.8	0.23, 0.29, 0.73
Search	γ	0.53	0.53	0.53
	μ_w	0.51	0.51	0.5
	σ_w	0.71	0.69	0.71
ΔI	γ	0.64	0.64	0.65
	μ_Δ	0.11, 0.72, 0.87	0.14, 0.97, 0.81	0.12, 0.55, 0.8
	σ_Δ	0.24, 0.23, 0.11	0.2, 0.08, 0.06	0.23, 0.08, 0.06
	μ_w	0.48	0.5	0.52
	σ_w	0.71	0.71	0.71

Table 3.4: Median fixed effects for all models fit to simulated data by generating model. γ is the probability of choosing consistent with the terminating model prediction (i.e., not failing to apply the model), μ_w is the average relative weight of CV and DR, and σ_w is the standard deviation of relative weight parameters. μ_Δ and σ_Δ give the mean and standard deviation of the delta parameters for each cue, indicating the distribution of differences in cue values needed to terminate search. μ_β and σ_β give the distributions for the weights in HyGene, indicating average search order.

Search order for HyGene follows the covariance between each cue and the outcome, on average, searching the first, then second, and finally third cue. The

standard deviation on the βs for HyGene is quite small for the first two μ_β parameters, but large for the third cue, suggesting that this cue is sometimes searched first. HyGene has separate search orders for each item. The high participant-wise variability seen in $\sigma_{\beta 3}$ is a reflection of the fact that, for some items, β_3 is larger than β_1. Both μ_w and σ_w for Search and ΔI follow the prior distributions for these parameters. No value of w will change the search order for these data because of the invariant DR values, so both models search in descending order of CV. The Δ parameters for Ecology 1 are larger than zero with a relatively small σ_Δ. This means that Search claimed all participants searched only the second cue and then made a decision, while ΔI participants searched the second cue and occasionally continued on to the first and then third cues, guessing only if all three cues differed by less than the applicable value of Δ.

Second Ecology

The results of fitting each model to data generated from Ecology 2 are similar to Ecology 1, with larger differences between the models (Table 3.5). On average, ΔI is preferred based on likelihood and DIC. This is a product of the noisy environment in which small differences in cue values are even less likely to correctly favor the larger outcome value. No ties exist for any cues in this environment and DR is always 1, causing Search to examine exactly one cue (the second cue) and make a decision based on these values. HyGene is allowed to search in different orders but also picks based on whatever cue is searched first. Only ΔI stops as a function of

Data	HyGene		
Model	HyGene	Search	ΔI
log $\mathcal{L}$	-1377	-1383	-1339
Penalty	3.373	1.033	6.811
DIC	2757	2767	2685
Data	Search		
Model	HyGene	Search	ΔI
log $\mathcal{L}$	-1384	-1383	-1332
Penalty	1.807	0.9384	7.152
DIC	2769	2767	2671
Data	Delta Inference		
Model	HyGene	Search	ΔI
log $\mathcal{L}$	-1376	-1384	-1340
Penalty	3.636	0.9114	3.746
DIC	2755	2769	2685

Table 3.5: Summaries of models fit to data generated from each model using second ecology.

informativeness for the first cue. When differences between the first-cue values are sufficient large, ΔI ceases search; otherwise, it continues through the other cues and either stops or guesses.

Table 3.2 gives the median values for all participant-varying (and cue-varying, as appropriate) parameters for the models fit to each set of data in each ecology. Compared with the posteriors for these parameters in Ecology 1, the models for Ecology 2 estimate approximately the same γ values despite the noisier ecology. This can only be attributed to bias in the models, not surprising given that decision models are designed to trade lower variance for higher bias (Gigerenzer and Brighton, 2009).

3.3 Discussion

The simulations presented above illuminate how Search, ΔI, and HyGene operate on both structured and noisy data. To explore this, I cross-fit each model to data generated from each of the three models from two separate underlying ecologies.

One major consideration, which in hindsight is obvious, is that the Search method of weighting CV and DR is only useful when there are unequal DRs for the cues. This is much more likely with discretely-valued cues, unlike the continuously-valued cues used in these investigations. As a result, Search effectively simplifies to TTB with probability $(1 - \gamma)$ of choosing counter to the TTB prediction. Thus, for these examples, Search is modeling no individual differences in search order. In addition to a single search order governed entirely by CV, the perfect discrimination rate of all cues means that Search was deciding based on a single cue and never searching beyond that.

This equivalence of DR across cues also leads the ΔI model to a consistent search order across participants. The difference between Search and ΔI is that, because of the Δ parameters, ΔI searches beyond the initial cue when the cue for the objects under consideration differ by less then Δ_1. While Search acts like an error-prone TTB, ΔI sometimes searches additional cues to reach a decision.

In these environments, HyGene is the only model that can model differing search orders. The relative weights of the cues for any given decision problem are influenced by the similarity between the cues for the choices and the cues for each of the choices in episodic memory, allowing cues to vary in importance between

choices. The model allows for weights to vary on average between participants as well, meaning search orders can vary both by individual and by item. This gives HyGene an advantage when search order should actually vary along these dimensions, assuming this variance is is sufficiently large relative to the error variance in the underlying ecology. HyGene's flexibility is unwarranted in noisy data or data simulated from Search or ΔI processes, however, giving the model a higher penalized likelihood relative to the other models on data generated with a consistent search order.

The consistent search order used in generating Search and ΔI models leave HyGene with unnecessary functionality on these simulated environments. This is especially true in the second, more complex environment. The positive covariances between the cues mean that, on average, the first cue is likely to partially encode the information present in later cues. Search, and to some extent ΔI, decide based on the first cue searched in this context, relying on the cue with highest CV. HyGene behaves in much the same way except that the single, deciding cue is less likely to be the highest-CV cue, since search order for HyGene can vary even with continuously-valued cues.

These results are useful for understanding human decision making outside of this modeling context. While the Search model is intended to improve understanding of individual differences in decision making (Lee and Newell, 2011), it applies only when there are sufficient ties in cue values to produce differences in DR. This is either a substantial limitation on the generality of the Search model or requires the discretization/dichotomization of available information is a necessary component

of the underlying representations people use to make decisions. The assumption that cues are learned or stored with discrete values is not necessarily unjustified or even novel (Gigerenzer and Goldstein, 1996), but is a strong assertion that should be explicitly considered. This assumption is even testable, providing qualitative falsifiability to the method of subject-varying cue ordering found in the Search model. Search is based on a fixed value of w. If individuals show different search orders with continuously-valued cues, then the proposed mechanism of cue ordering by combining weighted CV and DR cannot account for this pattern and the Search method of cue ordering must be altered or abandoned.

3.3.1 Summary

These simulated ecologies give important information about what to expect in future model fitting. The CV/DR weighting parameter in Search and the current version of ΔI will be more relevant with dichotomous cues but presents a potential avenue for empirical testing of Search cue ordering adequacy. HyGene is the only model under consideration that can account for search orders that differ by both individuals and test items, though this flexibility is unjustified in the current modeled environments.

Chapter 4: Prediction in a real-world data set

Though fitting models to simulated data can increase understanding of the models themselves, inferential models must also fit data without known parameters. Data generated from real environmental sources may differ in unpredictable ways from data generated with known properties. In this chapter, I use a well-known decision environment, the nine-cue ecology predicting population in German cities, to compare the resulting posterior distributions on the parameters of Search, ΔI, and HyGene.

Synthetic decision environments that come from fixed parameters and well-behaved probability distributions may not accurately reflect these natural environments. While it is trivial to design environments that are difficult or impossible to predict, people would simply guess in these circumstances. More interesting are environments that can be predicted despite uncertainty. Even difficult-but-predictable environments must be difficult in the same way that ecologies encountered by human decision makers are difficult. For these reasons, synthetic ecologies are of limited use.

Among existing decision ecologies, few are as thoroughly-explored as the German cities task (GCT). This task has been used in a large number of studies on

human memory and decision making (Gigerenzer et al., 1991; Gigerenzer, 1993; Gigerenzer and Goldstein, 1996), providing a reasonable baseline for performance of different models of decision making. The ecology for this task includes 9 dichotomous cues use to predict the population of the 83 German cities with populations larger than 100,000 (as of 1993, Figure 4.1).

Cue	CV	DR
Is the city is the national capital?	1	0.02
Was the city was an exposition site?	0.91	0.28
Does the city have a major-league soccer team?	0.87	0.3
Is the city on the Intercity line?	0.78	0.38
Is the city a state capital?	0.77	0.3
Is the license plate abbreviation more than one letter?	0.75	0.34
Is a university located in that city?	0.71	0.51
Is the city in the industrial belt?	0.56	0.3
Was the city in East Germany?	0.51	0.27

Table 4.1: Cues for the German Cities Task.

If the only goal was to predict city size, one could just as well check the CIA fact book and get the best predictors of population size. The strength of the GCT, in addition to being widely-used in the decision making literature, is that all of the cues are relatively easy to remember and use, since each can only take on one of two values. In addition to the supposed psychological plausibility of dichotomous cue values, the GCT will allow a better comparison between cue ordering due to w in Search and ΔI, Δ in ΔI, and β in HyGene. While the modeling in Chapter 3 compared these three models, it accomplished this comparison in a way that largely ignored the cue ordering aspects of Search and ΔI.

Delta Inference was designed to allow search of decision environments to continue past a marginal difference in cue values. Despite this, there is nothing in

principle to prevent ΔI from operating on dichotomous cues. Despite a consistent model structure, the interpretation of Δ changes with dichotomous input. The posterior distribution on a given participants Δ parameter for any cue only potentially changes decisions when it is ≥ 1. One can compare the density of the probability distribution for Δ above and below 1 to see how likely the model is to consider this cue, conditioned on the search order dictated by w.

These data reflect consistent subject-wise search orders. The original purposes of these data were to validate that the Search model can effectively model individual differences in search order. This puts HyGene at a disadvantage, but allows us to directly assess the influence of β priors when the data are generated with a single search order by subject.

4.1 Methods

The following simulations use the models specified in Chapter 2. The only slight difference is that cue values are dichotomous, rather than continuous, which would be reflected in the a_j and b_j nodes for each model.

The data for this chapter come from earlier work on the Search model (Lee and Newell, 2011). Twenty participants with 100 responses each are simulated using search orders of the GCT that differ by participants but are consistent within a participant across the 100 choices. These data are generated deterministically from the Search model, albeit with stronger relationship to the criterion than in Chapter 3 and using dichotomous cues.

4.2 Results

Metric	Model HyGene	Search	ΔI
log $\mathcal{L}$	−496.4	−482.7	−482.2
Penalty	46.47	5.904	47.71
DIC	1039	971.4	1012

Table 4.2: Model comparisons for HyGene, Search, and ΔI on the GCT.

Model fits using the GCT data are in Table 4.2. Search and ΔI are about equally effective at explaining variance in participant judgments, though the substantial complexity for ΔI is unjustified according to the DIC. HyGene yields only a slightly lower likelihood, though it is substantially more complicated than even ΔI and has a correspondingly higher penalized likelihood. This pattern of results expected, given that the data are effectively generated from the Search model.

Model summaries are in Table 4.2. In this environment, all three models converge on γ values near one. Only very rarely do the models assume that participants misapply the decision rule and choose counter to the predictions of the model. The median of the μ_β parameters for HyGene suggest that this method of cue ordering produces sightly different search behavior on average than the relative weight of CV and DR. This is likely to be relatively inconsistent across participants, given the accompanying σ_β parameter which are large relative to the sizes of the corresponding μ_{beta}s.

Though HyGene produces different cue ordering, Search and ΔI have the same average search order (Table 4.4). While search order differs by individual, the inclu-

Model	Parameter	Value
HyGene	γ	0.944
	μ_β	0.01, 0, 0.02, −0.05, −0.21, −0.1, 0, −0.12, −0.04
	σ_β	0.31, 0.12, 0.19, 0.25, 0.38, 0.49, 0.18, 0.83, 0.92
Search	γ	0.947
	μ_w	0.548
	σ_w	0.26
ΔI	γ	0.949
	μ_Δ	0.54, 0.59, 0.6, 0.64, 0.67, 0.64, 0.59, 0.72, 0.61
	σ_Δ	0.88, 2.31, 2.12, 1.86, 1.56, 1.68, 2.3, 1.38, 1.61
	μ_w	0.556
	σ_w	0.3

Table 4.3: Median fixed effects for all models fit to simulated participants with the GCT data. γ is the probability of choosing consistent with the terminating model prediction (i.e., not failing to apply the model), μ_w is the average relative weight of CV and DR, and σ_w is the standard deviation of relative weight parameters. μ_Δ and σ_Δ give the mean and standard deviation of the delta parameters for each cue, indicating the distribution of differences in cue values needed to terminate search. μ_β and σ_β give the distributions for the weights in HyGene, indicating average search order.

sion of Δ has little influence on search order: on average both Search and ΔI follow CV until the fourth cue. Though μ_w and σ_w are similar for Search and ΔI, the latter has non-zero probability density for each participant at $\Delta_k > 0$, resulting in probabilistic search of each cue. Assuming the HyGene model, the average person searches in a completely novel order, though the large standard deviations on the β parameters suggest substantial variability in HyGene search order.

Model	Median Search Order	Unique Orders
HyGene	2, 3, 1, 6, 9, 7, 4, 8, 5	1,379
Search	1, 2, 3, 6, 4, 5, 8, 7, 9	174
ΔI	1, 2, 3, 6, 4, 5, 8, 7, 9	269

Table 4.4: Search order information by model.

Recall that these data are generated with a consistent search order for each

participant that is the results from a weighted combination of CV and DR for each cue; Search is the true model for these data. We can see that, despite 20 true search orders (one for each participant), Search still allocates some probability (via the w parameter) to 174 unique search orders. Adding variability in Δ allows the ΔI model to identify 269 unique search orders. HyGene's βs give even more flexibility, exploring nearly 1,400 separate search orders. Each of these models explores far fewer than the total possible search orders, which for nine cues is 362,880.

The flexibility to identify additional search orders is not necessarily inappropriate, given the probabilities associated with the parameter values that give rise to these unique search orders. Figure 4.1 gives the distribution of τ order between the true search order and those proposed by each model for each of the 20 participants. τ order is the number of paired switches necessary to match the order between two vectors, it is related the Kendall's τ correlation coefficient. Search and ΔI overlap substantially for all participants, though the ΔI densities are more dispersed than those for Search. HyGene produces more varied results. The mean of the HyGene density varies in relation to the Search mean by participant, with HyGene showing higher numbers of disconcordances for some participants' fitted search orders and lower numbers for others.

The differences in search order between models must be interpreted with two caveats in mind. The average likelihoods are comparable for all three models. Despite the differences between the models in fitting the true search order for each participant, HyGene is only slightly worse at predicting participant responses. Also, the wider dispersion of HyGene τ orders is due to the variety of search orders, which

is a feature (and not a limitation) of the model. These τ distributions for HyGene are collapsing over the 100 search orders by stimulus pair for each participant, each of which is potentially unique. If participants truly search cues differently based on the stimulus in question, then HyGene is almost certain to fit better in expectation than any model that assumes homogeneous search orders for a given participant.

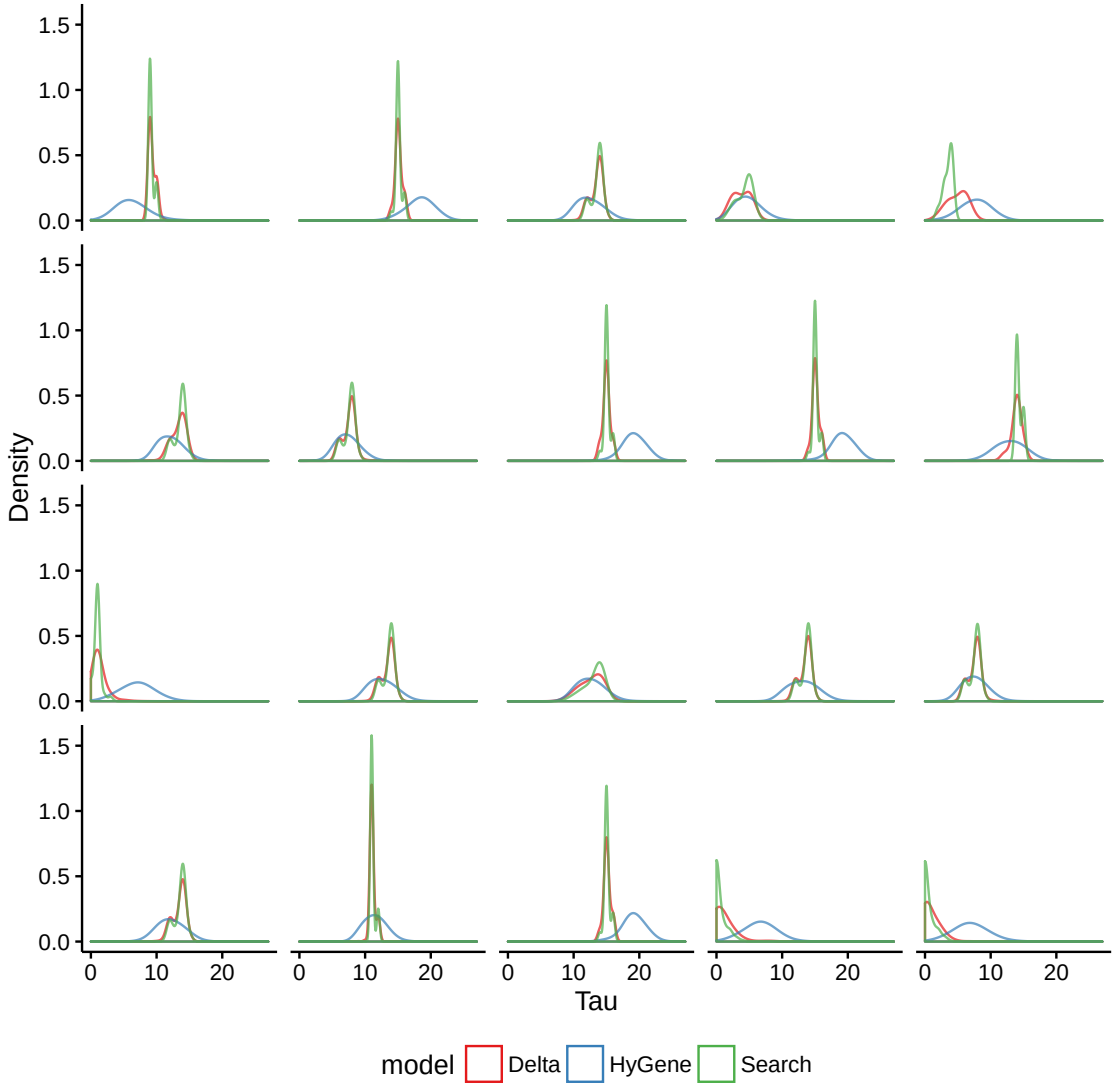


Figure 4.1: Density of tau distance between generating search orders and model search orders by participant, colored by model.

4.3 Discussion

This chapter focuses on fitting the three models to simulated participants from a widely-studied, naturally-occurring ecology. While this set of simulations does not directly inform on human decision making, one can learn more about how uncertainty is differently represented by each model. This establishes a baseline against which to compare these models when fit to data generated by human participants.

One important observation is that, in the GCT ecology, the additional parameters in ΔI and HyGene lead to likelihoods that are comparable to Search, even though Search is the true generating model. The current results contrast with the results in Chapter 3, which found higher likelihoods for ΔI across most generating models. This difference may be due to the relative predictability of the outcome based on the cues, which is much higher for the GCT relative to the ecologies in Chapter 1 (Table 4.5).

Ecology	R^2
1	0.35
2	0.333
GCT	0.868

Table 4.5: Comparison of variance in outcome explained by cues in the ecologies from chapters 1 and 2 using multiple regression.

At least one inference is consistent across all three models: The credible values of γ are very close (Table 4.2). Despite differing cue-ordering mechanisms, simulated participants make model-consistent responses at similar rates. Cue order matters very little for accurately predicting population in this ecology. The similarity across

models could also reflect that a certain subset of comparisons are difficult or impossible predict based on the observed cue values. This consistency across models is reassuring, though with a multiplicity of reasonable models, the only permissible inferences are those for which the models agree (Breiman, 2001). In this case, one can infer that participants are consistently choosing consistent with the result of the TTB process, but should remain agnostic about the method of ordering cues, since the models disagree on this and produce comparable likelihoods and DICs.

Interpretation of the ΔI model differs with dichotomous cues. This is because w and Δ will have an odd relationship on dichotomous cue values. If Δ is less than the discrete step size for a cue, the value of this parameter cannot influence search. On the other hand, if Δ is larger than this step size, then the model will always search the next cue. For these dichotomous cues, the only important question about Δ for a given cue is whether it is less than one or not. If Δ_1 is less than one, the model will only search past the first cue when it does not discriminate between the alternatives. If Δ_1 is greater than or equal to one, the model will never stop at the first cue. Since the decision mechanism is non-compensatory and only focuses on a single cue at a time, the latter case means that the first cue would not influence the decision process. This same reasoning applies to cues in any position and potentially limits the application of ΔI to dichotomous-cue environments. The hierarchical Bayesian models used in this chapter allow distributions on parameter values and Δ is continuously-valued with probability density both above and below zero. The observed values of Δ turn stopping into a stochastic process, the ΔI model will only stop at a discriminating with a probability that the applicable Δ parameter is less

than one. Stochastic stopping effectively adds another source of variability into the model, though the addition appears to cause almost no change in search order.

The GCT is one environment where working memory constraints might have played a role. Conditional selection can cause problems for later cues. For example, the ninth cue in the cities environment is whether or not a given city was in East Germany. Cities in East Germany tend to have larger populations than those that weer in West Germany. This is not necessarily the case when the prior either cues have already been searched. Given that the previous eight cues are all tied, cue nine might even have a negative cue validity. This would cause the model to make the incorrect choice once it has search to the ninth discriminating cue. Earlier work has explored the application of greedy algorithms that account for this conditional dependency with TTB, but find that decisions based on this strategy are worse in cross-validation than unconditioned decision rules, though this is not compared to any decision models that explicitly cease search after a set number of cues (Martignon and Hoffrage, 2002). In the case of truncated search as in the original implementation of HyGene, a limited working memory would cause the model to exit search and guess after searching a set number of cues. If conditional cue validity is negative for later cues, ignoring those later cues potentially leads to fewer incorrect choices, though it does not guarantee better concordance with human decision processes.

4.3.1 Summary

This chapter uses Search, ΔI, and HyGene to explore an ecology with dichotomous cue values and with a strong relationship between the cues and the outcome. The findings for Search replicate Lee and Newell (2011) and provide additional evidence for the flexibility that ΔI and HyGene's specific parameters provide in search order.

Chapter 5: Modeling human inference: A novel behavioral experiment

This chapter focuses on the application of Search, ΔI, and HyGene to data from participants in a behavioral task. After unique training periods on a novel task environment, participants are allowed information on a single cue and asked to choose between a pair of stimuli. Data from this experiment allow for comparison of the three models in an environment that potentially requires the full flexibility of HyGene. Given the inconsistency in search orders observed in previous studies, I expect the inconsistent individual search orders allowed under HyGene give this more complicated model an advantage relative to Search and ΔI. I also expect ΔI to guess slightly more often than Search, assuming some probability of Δ_1 exceeding the difference in cue values for the first cue.

The current data also give a second, convergent method for validating the models. While participant judgments and cue values will be used to fit each model, participants also generate information on which cue is searched first for each trial. Existence of cue choices allow for comparison observed cue search and predicted cue search from each model. Potential inconsistency in the first cue searched for each participant which favors the assumptions built into HyGene, which allows for

varying search order based on the probe vector. Search and ΔI should both produce more consistent predictions about which cue is searched first relative to the more flexible HyGene model.

5.1 Methods

Participants

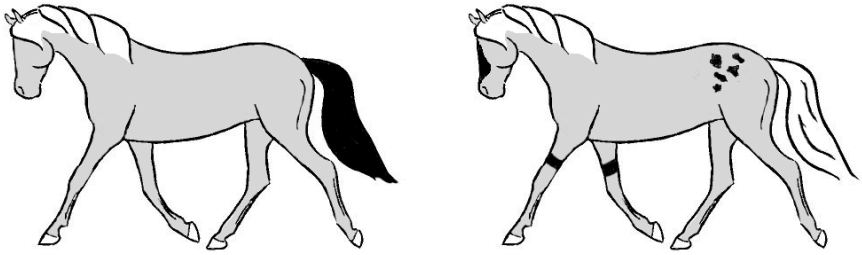
Thirty-eight participants (60% female) from the Psychology department subject pool at the University of Maryland, College Park took part in this experiment. Participants received partial course credit for their participation.

Stimuli

Stimuli for this experiment consisted of line drawings of ponies that were identical except for four dichotomous cues: nose color, leg stripes, hind spots, and tail color. All combinations of four dichotomous cues yields a total of 16 unique figures, see Figure 5.1 for maximally different examples.

The ecology in this experiment was designed so that CV, DR, and τ/success would each identify a separate, preferred cue[1]. The ecology consists of one set of cue weights and some probability for each unique pony. Stimulus pairs for both learning and test phases were created by uniformly sampling from figures according to the frequencies listed in Table 5.1. This sampling process produces an ecology

[1]Due to a coding error, a second ecology did not accurately allow discrimination between success and τ and is not reported.



(a) All cues absent. (b) All cues present.

Figure 5.1: Comparison of pony drawings use in learning phase.

with summary statistics found in Table 5.2.

	Stimulus	Frequency
1	0000	50
2	0001	5
3	0010	10
4	0011	1
5	0100	24
6	0101	3
7	0110	7
8	0111	1
9	1000	46
10	1001	5
11	1010	9
12	1011	5
13	1100	24
14	1101	6
15	1110	3
16	1111	1

Table 5.1: Frequencies of each stimulus for the test ecology.

Procedure

Participants gave consent and heard the following description:

	Cue 1	Cue 2	Cue 3	Cue 4
CV	0.704	0.871	0.867	0.997
DR	0.502	0.427	0.276	0.232
τ_a	0.205	0.317	0.202	0.231
τ_b	0.319	0.518	0.405	0.525
Success	0.603	0.659	0.601	0.615
p(Present)	0.495	0.345	0.185	0.135
Weight	0.100	0.200	0.200	0.400

Table 5.2: Summary statistics for pony cue ecology.

Welcome to the world of pony consulting. This is a cut-throat industry in which desirable ponies are in high demand, pony buyers are extremely wealthy, and pony sellers are highly protective of their goods.

You are training to become a pony consultant. As a pony consultant, your task is to pick ponies that your clients will like. As with any other competitive consulting industry, you will get [rewarded/penalized] for [satisfied/dissatisfied] clients for whom you select the [right/wrong] pony. Your payment today will be based on your overall performance - your goal is to [earn the most positive/receive the fewest negative] reviews.

To prepare for pony consultant work, you will first practice selecting ponies. You will see pictures on the screen of two different ponies and you must choose the more desirable pony. Ponies vary on four traits: face color, leg stripes, spots, and tail color. Pay attention because once you start working you will need to know which pony traits are considered most desirable so you pick the right ponies for your clients.

After completing training, work in the real industry begins at the pony

auction house. As before, you must select the more desirable pony. However, in the real world where the pony sellers are highly protective of the ponies, you must pay to reveal traits. As a junior consultant, you have only enough budget to reveal one trait for each pair of ponies. After that trait is revealed, you must make your choice.

Now it's time to practice selecting ponies. On the screen you will see pairs of ponies. Use the mouse to indicate which pony is more desirable. After making your choice you will receive feedback - green means you made the correct choice; red means you made the incorrect choice; and yellow means that the ponies were equally desirable.

Participants then completed 40 learning trials. On each learning trial, participants saw two ponies side-by-side and clicked a button under the picture they judged to have higher value based on the set of cues. Following each choice, the buttons disappeared from the screen and a colored border appeared around the selected picture: green for correct, yellow for tied, and red for incorrect. For correct and tied choices, the border remained for 500 ms, while the incorrect border remained for 2000 ms.

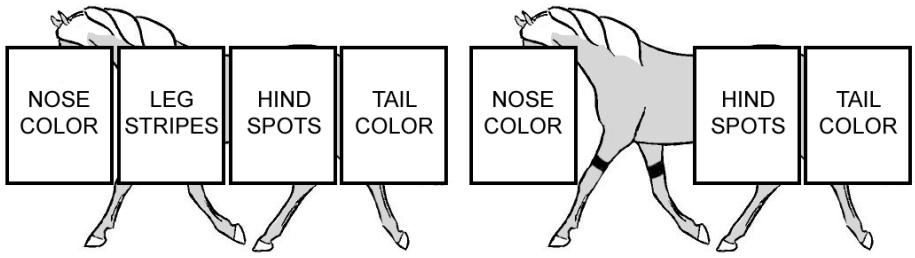
After 40 trials, a research assistant read the following instructions to participants:

Congratulations! You're now a junior consultant. As before, you will see pairs of ponies but this time, their traits are covered. You have enough budget to reveal one trait for each pair of ponies; it is not necessary

to pick the same trait every time. Once the trait is revealed, you must select the more desirable pony.

Every time you select a pony that your client really [likes/dislikes], you will receive a [positive/negative] performance review. At the end of the pony auction, the number of [positive/negative] reviews will determine your pay - the [more positive/fewer negative] reviews you earned, the more sweets you get.

Participants then completed 160 test trials. Each test trial began with two masked stimuli (Figure 5.2). Participants were allowed to select a single cue to uncover by clicking on the corresponding named button, which removed the cover from only that cue. They then made their choice between the stimuli based on that single cue. After each decision, a tally of the earned points at the bottom of the screen updated.



(a) All cues masked. (b) One cue masked.

Figure 5.2: Stimulus states for test phase.

Modeling

Models are fit only to the 38 participants in the first ecology. HyGene, Search, and ΔI required slight modification to accommodate heterogeneous training samples. An implicit assumption with previous datasets is that participants have experience with the ecologies of interest. This assumption is instantiated by fitting the models using CV and DR calculated on the entire ecology (Search and ΔI) and including an episodic memory weights for all pairwise comparisons between objects in the ecology (HyGene). For this set of simulations, model CV and DR are calculated separately for each participant using only the 40 pairs of stimuli seen during training. The episodic memory for HyGene is also limited to this set of training stimuli. Participants saw limited feedback during test and cannot be expected to have previous experience with the artificial test environment created for this study. Failure to account for the different experiences participants had with the environment potentially biases the subject-varying parameters fit to test responses.

The current experiment allowed participants to search only a single cue before making a decision. The models are also modified to make a choice after inspecting only a single cue despite determining a search order for the entire cue population.

5.2 Results

Participants' accuracy improved over the course of training in spite of the short duration. Table 5.3 and Figure 5.3 show that on average, accuracy increases from 74.5% to 87.8% over the course of training. The high accuracy demonstrates the

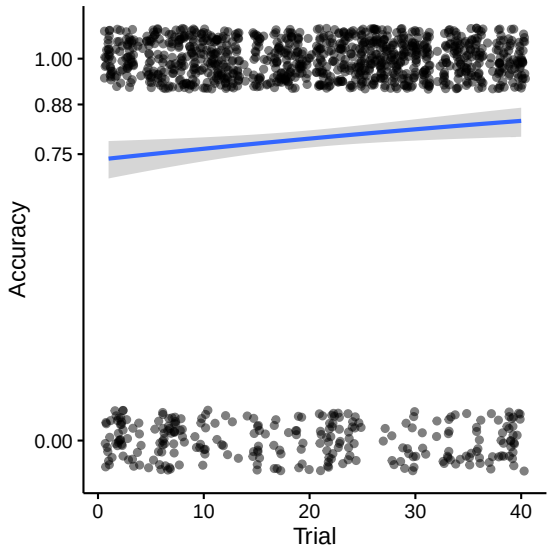


Figure 5.3: Jittered scatterplot and logistic regression prediction for accuracy by trial during training. The intermediate tick marks on the y-axis show the average predicted accuracy for the first and last trials based on the multilevel logistic regression model in Table 5.3.

overall ease of the task; the outcome is perfectly predicted from the cues assuming full knowledge of the environment. All of the cues are also positively related to value and participants have access to all cues during training. Given a dearth of plausible alternatives, these results suggest that participants are increasing accuracy on average by learning about the cue ecology.

		Fixed Effects		
	Coefficient	Std. Error	z	$Pr(>\|z\|)$
Intercept	1.07	0.185	5.793	6.93×10^{-9}
Trial	0.023	0.009	2.451	0.014
		Varying Effects		
Group	Coefficient	Variance	Std. Dev.	
Subject	Intercept	0.554	0.744	
	Trial	0.001	0.038	

Table 5.3: Summary of multilevel logistic regression predicting accuracy using trial and varying both intercept and the effect of trial by participant.

One important feature of these data is the variation in search orders within

participants (Figure 5.4). Ideally, a model of decision making would both predict participant choices and mimic the decision process used by participants. Search and ΔI are specified with the assumption that participants use a consistent search order for each participant, a feature that is contradicted by the search patterns in these data. If HyGene is emulating the process of memory search participants are using to make decisions, then the model should both fit the choices participants made and show similar distributions of cue choices.

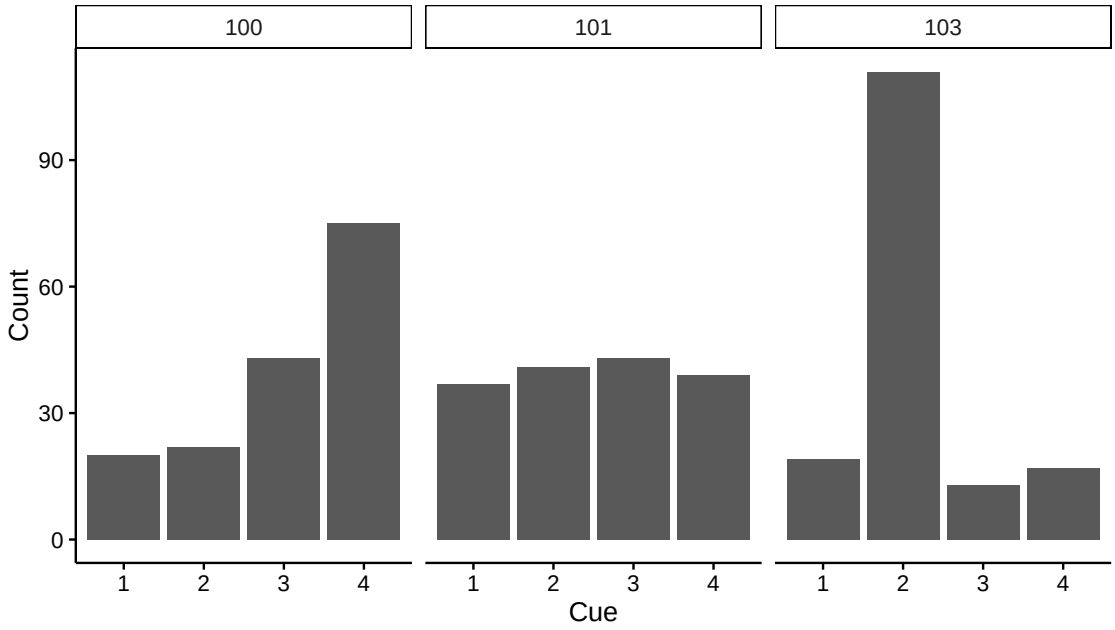


Figure 5.4: Distributions of first cue searched during the test phase for three example participants.

	Model		
Metric	HyGene	Search	ΔI
log $\mathcal{L}$	-4215	-4215	-4215
Penalty	0.004	0.003	0.005
DIC	8431	8431	8431

Table 5.4: Model comparisons for HyGene, Search, and ΔI on empirical data.

Model	Parameter	Value
HyGene	γ	0.5
	μ_β	0.027, 0.06, 0.09, 0.081
	σ_β	0.114, 0.11, 0.105, 0.095
Search	γ	0.5
	μ_w	0.482
	σ_w	0.724
ΔI	γ	0.5
	μ_Δ	0.515, 0.529, 0.502, 0.497
	σ_Δ	0.747, 0.831, 0.864, 0.699
	μ_w	0.516
	σ_w	0.711

Table 5.5: Median fixed effects for all models fit empirical data. γ is the probability of choosing consistent with the terminating model prediction (i.e., not failing to apply the model), μ_w is the average relative weight of CV and DR, and σ_w is the standard deviation of relative weight parameters. μ_Δ and σ_Δ give the mean and standard deviation of the delta parameters for each cue, indicating the distribution of differences in cue values needed to terminate search. μ_β and σ_β give the distributions for the weights in HyGene, indicating average search order.

Table 5.4 gives the log likelihood, penalty, and DIC for each of the three models on these data. These data are equally likely under each of the three models. The additional parameters in ΔI and HyGene have almost no effect in this circumstance, so the penalty used for DIC is nearly equivalent for all three models as well. For all three models, γ is exactly .5 (Table 5.2). While the β parameters for HyGene are defined relative to the training data and are unaffected by this, the fixed effects for Search and ΔI reflected the prior distributions because they have no effect on the likelihood calculation when $\gamma = .5$.

Other evidence suggests that participants are not guessing (Figure A.1). To get an idea of how the models would fit on the assumption that people were not always guessing at the outcome, each model was fit to the data with the γ parameter fixed

Metric	Model HyGene	Search	ΔI
log $\mathcal{L}$	−5406	−5406	−5396
Penalty	0	0	24.46
DIC	1.081×10^4	1.081×10^4	1.082×10^4

Table 5.6: Model comparisons for HyGene, Search, and ΔI on the empirical data with fixed $\gamma = .75$.

Model	Parameter	Value
HyGene	μ_β	0.027, 0.06, 0.09, 0.081
	σ_β	0.114, 0.11, 0.11, 0.095
Search	μ_w	0.515
	σ_w	0.729
ΔI	μ_Δ	0.505, 0.506, 0.505, 0.504
	σ_Δ	0.385, 0.432, 0.374, 0.448
	μ_w	0.454
	σ_w	0.684

Table 5.7: Median fixed effects for all models fit to empirical data with fixed $\gamma = 0.75$. μ_w is the average relative weight of CV and DR, and σ_w is the standard deviation of relative weight parameters. μ_Δ and σ_Δ give the mean and standard deviation of the delta parameters for each cue, indicating the distribution of differences in cue values needed to terminate search. μ_β and σ_β give the distributions for the weights in HyGene, indicating average search order.

at .75. The results are in Tables 5.6 and 5.2. These models all have lower average likelihoods after the forced increase for γ. While ΔI has a slight advantage relative to Search and HyGene in likelihood, it also has a much larger effective number of parameters, giving it larger (and less favorable) DIC.

Figure 5.5 plots the probability of choosing each cue first for each participant for each of the models and the empirical data based on the search orders from the revised models that fix γ at .75. The task only allowed participants to search a single cue, so validation must be limited to predictions regarding the first cue searched. A successful model of cognition should mimic the patterns present in

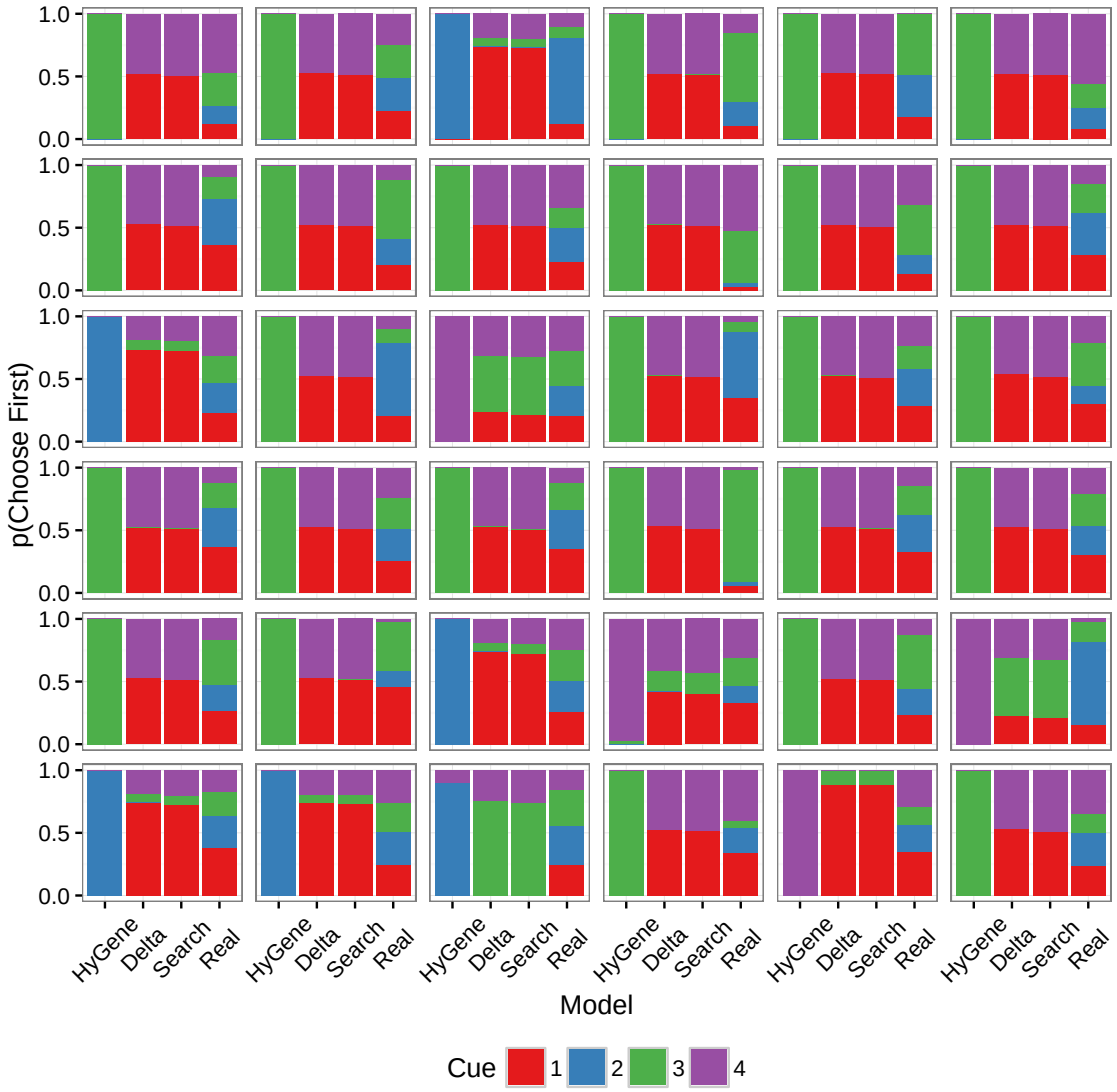


Figure 5.5: Probability of choosing each of the four cues first by source, faceted by participant (with the right bar showing empirical cue choice distributions). Two subjects omitted for space.

the data. Almost all participants searched all of the cues first at least once. Some participants had near-uniform rates of first cue search while others were much more likely to choose one or two cues more often than the others. The preferred cue differed by participant; some of the non-uniform choosers preferred cue four (highest CV) while others preferred cue one (highest DR). The three models are also fit to

the same training information as the participants, so a really successful model would exhibit all of these patterns and also accurately predict the frequencies with which each participant chose to search cues first. Prediction of cue search behavior is a particularly high bar, given that the model likelihood does not depend on these search orders; first cue search is an emergent property for all three models.

Search and ΔI both make very similar predictions for first cue searched. Both of these models predict that cues one and four will be searched with some probability for most participants, with occasional looks at cue three. Both models also show some variability is uniformity of cue choice: some participants show a strong preference for a single cue while others choose between three of the available cues more evenly. The variance in uniformity does not follow observed patterns in participant cue choice, however; Search and ΔI do not follow a given participants probabilities of choosing individual cues very closely, if at all. Neither of these models predicts even a single look at the second cue, a notable departure from the data.

HyGene makes very different predictions that are inaccurate in different ways. This model usually only identifies a single cue that a participant will search, though it occasionally shows a second cue with a non-zero probability of being searched first. HyGene does sometimes predict the second cue as being the preferred cue, though, which is consistent with the observed choices and never predicted by Search or ΔI. The modal pick for HyGene is also reasonably consistent with the modal choice for each participant.

5.3 Discussion

The current chapter focuses on fitting Search, ΔI, and HyGene to human responses rather than generated data. While simulated data was useful in understanding the mechanics for each model, the goal is to understand something about human behavior. I focus on each model in turn before attempting to reconcile the inferences from the differing explanations.

For these experimental data, HyGene predicts very consistent search orders within participant. HyGene also predicts that no one will search the DR cue. Despite obvious deviations from the observed data, the single cue searched by HyGene often agrees with the first- or second-most searched cue in reality. Though HyGene's search orders are too consistent, the pattern of cue choices suggests that the ordering mechanism is picking up on the same environmental structure as the participants. HyGene's median values of μ_β are quite small compared with σ_β. Despite this, HyGene searches cues in a very consistent order. This suggests unmodeled but positive covariances between β parameters, which encode the overlap between the cues in predicting magnitude of interest. Especially on the small training samples in this experiment, collinearity between βs may be reflected in cue search.

Search heavily favors initial search of the first (DR) and fourth (CV) cues. Search also occasionally searches the third cue, which is preferred by no suggested cue ordering metric, but avoids the τ/success cue. This behavior is the result of a large uncertainty on the relative weighting of CV and DR described by σ_w. Search was formulated as a model about the individual differences in cue search based on

fixed, point values of w for each participant. Despite this, the variability w within participant is what allows variation in cue search order and increases similarity between modeled and actual cue search orders. Search is interesting because of the extremism it shows. The model probabilistically searches the CV cue, which has very poor discrimination, or the DR cue, which has the poorest CV, despite the existence of cue two, which goes entirely unsearched and is a compromise between these features.

Delta Inference makes predictions that are largely consistent with the Search model. The Δ parameter distributions once again have density above one, so some of the time a differentiating cue is being ignored. Large values are Δ are infrequent, however, and do very little to change search order. The relative weighting parameters are very similar to those in the Search model and yield very similar first-cue choices.

Across all three models, participants are guessing quite often. The high rate of guessing is likely a function of the ecology, which contains a large number of tied values and no cues of exactly zero weight. Assuming the correct valence is learned for each cue, searching any cue will yield above-chance accuracy. Participants may have picked up on this strategy and been insufficiently motivated to maximize accuracy on the task by using a more difficult strategy. This would be consistent with findings from Newell et al. (2009), which also used an environment for which different strategies would yield only small differences in accuracy.

The current experimental setup may not be an ideal test for these models. All four cues have positive predictive value according to the CV, DR, and success. If the

cues are too similar, participants may be unwilling or unable to distinguish between them. The difference in accuracy between conditions provides some protection from this criticism. The only difference in information for a given trial, and the only way accuracy could differ, is through cue choices. Despite the variation in first cues chosen, the difference in accuracy between conditions provides indirect evidence for some consistent pattern in cue use. Another potentially limiting factor is the single cue that participants are allowed during the test phase. Forcing a single cue choice could change cue ordering behavior. While there is no way of knowing whether limiting available information alters cue utilization, this is a potentially interesting question for future research. If limiting information changes decision making, then it may be an important feature to include in future decision models. While no a priori reason exists for limiting information to alter cue use, it could explain the failure of all three models to capture cue choices in this study.

5.3.1 Summary

Beyond a high proportion of guessing, these three models do not agree on much regarding participant behavior. No single model is a particularly good predictor of actual participant search orders despite having nearly equivalent likelihoods for these data. This is partially by design, each of these models exists to explain decision making at a different level. In this case, however, it produces models that are mutually incompatible while capturing qualitatively different features in cue search.

Chapter 6: General Discussion

> The road to wisdom? — Well, it's plain
> and simple to express:
> Err
> and err
> and err again
> but less
> and less
> and less.
>
> *Piet Hein*

Search, Δ Inference, and HyGene have disparate theoretical motivation but predict human behavior with similar success. Being simplifications, it is neither surprising nor discouraging that they also fail to account for potentially important response patterns (Box and Draper, 1987). These limitations do not preclude the culling of useful information from computational models, however; the ways in which these models fail tells us something about what aspects of decision making could be explained by omitted components.

Chapter two summarizes the performance of Search, ΔI, and HyGene when fit to data generated from each of the three models using two well-defined, underlying ecologies. ΔI consistently outperforms Search and HyGene both in penalized and unpenalized likelihood despite a large effective number of parameters. Searching past a continuous cue that barely discriminates between alternatives is potentially

more important for model performance than flexible order of search. Both Search and ΔI had fixed search orders for these data. Given the focus on psychology rather than normative model performance, these results serve merely as a baseline for understanding how different aspects of these models are related.

Each model is next fit to simulated participants with varying search order governed by a weighted combination of CV and DR. The GCT is a commonly-used dataset for decision research with unknown but naturalistic ecological structure and strong relationships between the cues and criterion values. For these data, Search fit better than ΔI and HyGene, though all three models identify very similar distributions of search orders, albeit with varying precision and accuracy. While this still tells us nothing about the psychology of decision making as such, it suggests limitations with using ΔI on dichotomous cues and shows that, even with mis-specification, HyGene's ordering mechanism generates reasonable posterior distributions of cue search behavior.

The penultimate chapter compares Search, ΔI, and HyGene when fit to behavioral data using an ecology designed to assess cue preference. The models in this chapter suggest, above all else, that participants in this study were guessing a large percentage of the time despite statistical evidence of greater-than-chance accuracy. The three models produce highly similar average fit statistics, suggesting comparable success in explaining the data. HyGene also produces very different search behavior than Search and ΔI when focusing only on the first cue searched. Search and ΔI settle on values of w that vary the first cue searched between the first and fourth cue (those highest on DR and CV, respectively) with occasional looks

at the third cue (highest on none of the included metrics) depending on the participant. HyGene produces search behavior that is very consistent within participant, but would have participants search the second, third, or fourth cue depending on their training set. These different behaviors are uniquely inconsistent with observed search behavior; both should select among all four cues, though they omit different cues, and both should show more variability among cues within participant, though HyGene shows less variability than Search and ΔI.

Search, ΔI, and HyGene exhibit some consistency in the posterior estimates of their parameters. The guessing parameter, γ, is estimated quite consistently across all models when fit to the same data. Consistency in γ suggests that the models agree on the error rate with which participants apply TTB to the modeled search orders. Though ΔI tends to converge on a slightly higher mean value, this is caused by an increased number of guesses. In terms of the model, when $TTB(a_j, b_j) \notin (a, b)$, the model chooses either outcome with a 50% chance. ΔI has a higher proportional of true guesses because of the possibility of $\Delta > 1$; true guesses do not affect the posterior of γ. Error of application and guessing are only equivalent when $\gamma = .5$, otherwise the model is more likely to produce a correctly-applied response.

In the GCT with participants simulated from Lee and Newell (2011), all three models identify very similar search orders. Despite the true search order being dictated by a relative weighting of CV and DR, the HyGene cue ordering mechanic approximates the search orders quite well. HyGene has more diffuse distributions of τ order due to misspecification, but still manages to find some search orders that are closer to the true order compared with Search and ΔI due to the varying search

order within participant. The observed consistency in search order suggests that whatever variance in cue ordering is due to CV and DR can be at least partially recovered using the current version of HyGene's search of episodic memory.

Taken together, this agreement causes problems for the interpretation of the models in chapter four. For these data, HyGene's predicted search orders are quite different from Search and ΔI when fixing γ above the guessing threshold. The initial, low value for γ, however, suggests that regardless of the searched cue, the models predict that participants are effectively guessing. This consistent inference across the models, despite performance well above chance for nearly all participants, indicates that some important aspect of decision making behavior is entirely ignored by these models.

6.1 Psychological Plausibility

Like all research, these studies have limitations. The three tested models represent a very small subset of the possible models that encode cue ordering in a two-alternative, forced-choice context. Other models like SSL (Rieskamp and Otto, 2006), mixtures of models (Scheibehenne et al., 2013), or cognitive neuroscience-inspired models (Donoso et al., 2014), could more closely resemble the decision process that people use. Search, ΔI, and HyGene are interesting particularly because of their similarity. While these studies are unlikely to uncover the true generating model for participants' responses, features of the process that are consistent with available explanations are more likely to be true of human decision making.

Further work in decision making must focus on psychological plausibility.

Though hierarchical Bayesian modeling of individual differences is a step beyond deterministic models, aspects of Search, ΔI, and HyGene are still potentially optimistic about the limits of human cognition. Search and ΔI, for instance, make use of CV and DR calculations, which require that people either store or calculate those for relevant cues when making a decision. Calculation of CV and DR is less psychologically plausible than something like HyGene, which is based directly on memory search and therefore has convergent evidence for the cue ordering mechanism. One could go farther by including limitations on working memory, including temporal dynamics, or modeling the learning process within these models.

These models could be further constrained. For example, psychological constructs like working memory could be assessed and included as data in these models (Lee, 2010), rather than assumed and fit as free parameters. Adding features such as working memory to all of these models potentially increases the psychological and biological plausibility of these models.

6.2 Modeling Search Order

The CV and DR weighting mechanism in the Search model is restrictive. The Search method of weighting only allows search orders that are some combination of CV and DR. The data from Chapter 4 serve as an existence proof that, at least some of the time, participants will select cues that are not ever predicted by this metric. While there are mathematical and historical reasons for focusing on CV and DR, a model allowing all possible search orders would sacrifice some ease of interpretation for a more accurate estimate of variance in search order attributable

to individual differences. The w parameter is convenient, it is interpreted as a participants' relative preference for two cue metrics. A Dirichlet distribution or Gaussian process model for search order would allow for all possible search orders but would be nearly impossible to summarize with a single value. If the goal for the Search model is to give plausible estimates of sources of uncertainty in decision processes, then abandoning w for a more flexible mechanism makes sense.

Another problem with the use of w to establish cue order is that it maps non-linearly onto search order. Depending on the distribution of CV and DR in a given ecology, changes in w cause completely unpredictable changes in search order. The non-linear mapping of w onto cue order gives little reason to believe that the posterior distribution on w will be continuous and unimodal, either. Despite the apparently simple interpretation and obvious relationship with the success metric, relative weighting of CV and DR is a troublesome method for understanding search order.

6.3 Contamination

Lee (2010) also suggests a process for detecting contamination. Though the first two chapters focused on simulated data, chapter four's empirical data almost certainly includes features that are not the direct and sole result of a person making his or her best judgments about the task. The models in this dissertation include a misapplication parameter, γ, which fills this role in a limited way. Lee and Newell (2011) interpret this term as an error of application, with γ being the probability of choosing explicitly counter to the TTB prediction (and separate from guessing).

Modeling contamination might make use of additional information, such as reaction time or changes in accuracy over time, to isolate and remove patterns in data that are unrelated to the underlying construct of interest. Removing contaminants prevents the substantive model parameters from attempting to account for variability in the observed data that are actually the result of a mixture process. In addition to guessing, participants may also search cues differently over time (Table A.2), a process that could be motivated by the limited feedback, boredom, or exhaustion. The motivation for removing contaminants is the same as for removing outlying data points. Extreme scores can bias statistical tests. Unfortunately, unprincipled outlier removal can also negatively influence test properties (Antonakis and Dietz, 2011). Removing participants with near-chance accuracy would partially alleviate the problem of modeling the mixture of true decision making behavior and contaminant guessing, but it would also introduce a selection problem into the modeling. Modeling contaminant processes provides a principled method of removing spurious or unrelated patterns in the data. One inference from chapter four is that participants are almost certainly guessing on some trials. Different collection methods and more constrained models would potentially allow us to isolate the guessed trials by participant and focus on trials that are the result of a legitimate decision process.

6.4 Aggregation

Search order is not the only question and might not even be a relevant question. Thought the models under consideration assume sequential cue search, people may combine cues in various, unordered ways to produce decisions. Some modeling has

even been done to capture the trade-off between sequential and simultaneous cue use (Lee and Newell, 2011; Ravenzwaaij et al., 2014). In some cases, both methods produce similar results. Many studies provide evidence that participants search through cues rather than combining them in some way (Newell and Shanks, 2003). The experimental data in chapter four allows only a single cue's information for each decision at test. Even if this scenario is only relevant to a subset of decisions that people make in the wider world, participants in the study had to choose among cues in some way.

A more extensive comparison of the Search, ΔI, and HyGene models would explore aggregation as well. The Stop model, briefly discussed in chapter 1, provides one potential method for understanding the balance between effort and information in decision processes Lee and Newell (2011). The same methods used for cue ordering could be applied as weighting schemes or to alter the stopping rules used for the decision process.

One motivation for exploring non-normative models of decision making is to account for the information-processing constraints that humans impose on the process. Decision processes could be made informationally frugal in a large number of ways. This is a question that ΔI attempts to address. With dichotomous cues, the Δ parameters in our models were just another source of uncertainty in the models, occasionally allowing the search process to continue past a dichotomous cue. With continuously-valued cues, however, Δ provides a mechanism for adaptive equivalence, setting a threshold on what differences are meaningful enough to stop search. Though this level of complication was not justified in the limited environments for

the present dissertation, fitting to multiple environments and examining sources of individual variation or combining the Δ parameterization with other cue ordering methods could be useful.

6.5 Summary

People make seemingly difficult decisions constantly and with relative ease. Experimental work shows that these decisions do not conform to a variety prescriptive or deterministic models, but it is yet unclear how far uncertainty can be reduced in human decision processes. This dissertation shows that three qualitatively different computational models of cue ordering and decision making can fit empirical data with similar success but fail to capture important patterns in search order. Assuming sequential search of cue information, people might differ on relative weighting of cue metrics or generate cue orders based on similarity of a decision to memories of the ecology. Either way, other inputs to the decision process must explain variation in individual cue ordering and people probably aggregate over multiple cues in some instances.

Appendix A: Experimental Differences in Accuracy

	Fixed Effects			
	Coefficient	Std. Error	z	$Pr(>\|z\|)$
Intercept	0.314	0.072	4.334	1.46×10^{-5}
Gain	0.228	0.103	2.223	0.026
	Error Terms			
Group	Coefficient	Std. Dev.		
Subject	Intercept	0.259		
Residual		1.000		

Table A.1: Summary of multilevel logistic regression predicting accuracy using condition and varying intercept by participant. Intercept gives the average accuracy for the loss condition, the difference in accuracy for the gain condition is given by the Gain predictor.

There is some evidence that the gain/loss manipulation alters accuracy. Participants were generally not very accurate on the task, with an overall accuracy of $59.8\% \pm 0.5\%$. A multilevel logistic regression with varying intercepts by participant suggests that accuracy is higher for participants in the gain condition compared to those in the loss condition (Table A.1). Very few participants had average accuracies of less than chance (Figure A.1).

It is possible that, absent sufficient engagement or feedback, participants altered their cue search strategies over time. Table A.2 presents posterior estimates for a multilevel multinomial model predicting participant cue choices in the test phase by trial using a logistic link function with MCMCglmm (Hadfield, 2010). This multinomial model allows for differences in cue choice and the effect of trial

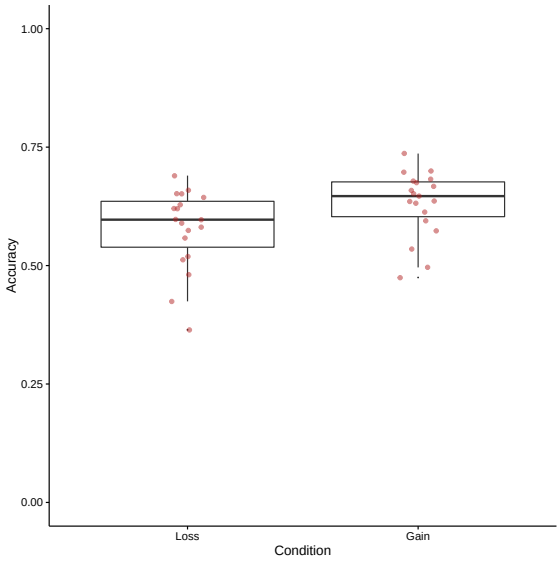


Figure A.1: Boxplots for average participant accuracy by condition.

on cue choice for each participant, which is the equivalent of varying intercepts and slopes in a multilevel regression. The first three parameters estimate the rates of choosing cues two, three, and four relative to cue one on the first trial. Participants appear to choose cues one and three at nearly equal rates for the first trial, while cues two and four are chosen less often than the first cue. The second three parameters estimate the difference in rate of choosing between a given cue and cue one for each additional trial. Relative to cue one, cue two is chosen more often in later trials. Figure A.2 shows the probability of choosing each cue over time, averaged over participants.

Parameter	Mean	95% CI	pMCMC
Cue 2	-0.41	$(-0.7, -0.07)$	0.013
Cue 3	-0.22	$(-0.6, 0.2)$	0.269
Cue 4	-0.48	$(-0.9, -0.07)$	0.025
Cue 2:Trial	0.002	$(-4 \times 10^{-6}, 0.004)$	0.042
Cue 3:Trial	10^{-5}	$(-0.002, 0.002)$	0.967
Cue 4:Trial	6.6×10^{-4}	$(-0.002, 0.003)$	0.697

Table A.2: Fixed effect estimates for a multilevel multinomial model predicting cue choice by time with varying effects by participant. Mean gives the mean of the marginal posterior distribution for each parameter, while the 95% confidence interval gives the 2.5% and 97.5% percentile samples for each marginal posterior distribution. pMCMC is an MCMC approximate of the p-value and gives the probability of observed an estimate of equal or greater magnitude given the estimated standard deviation centered at zero.

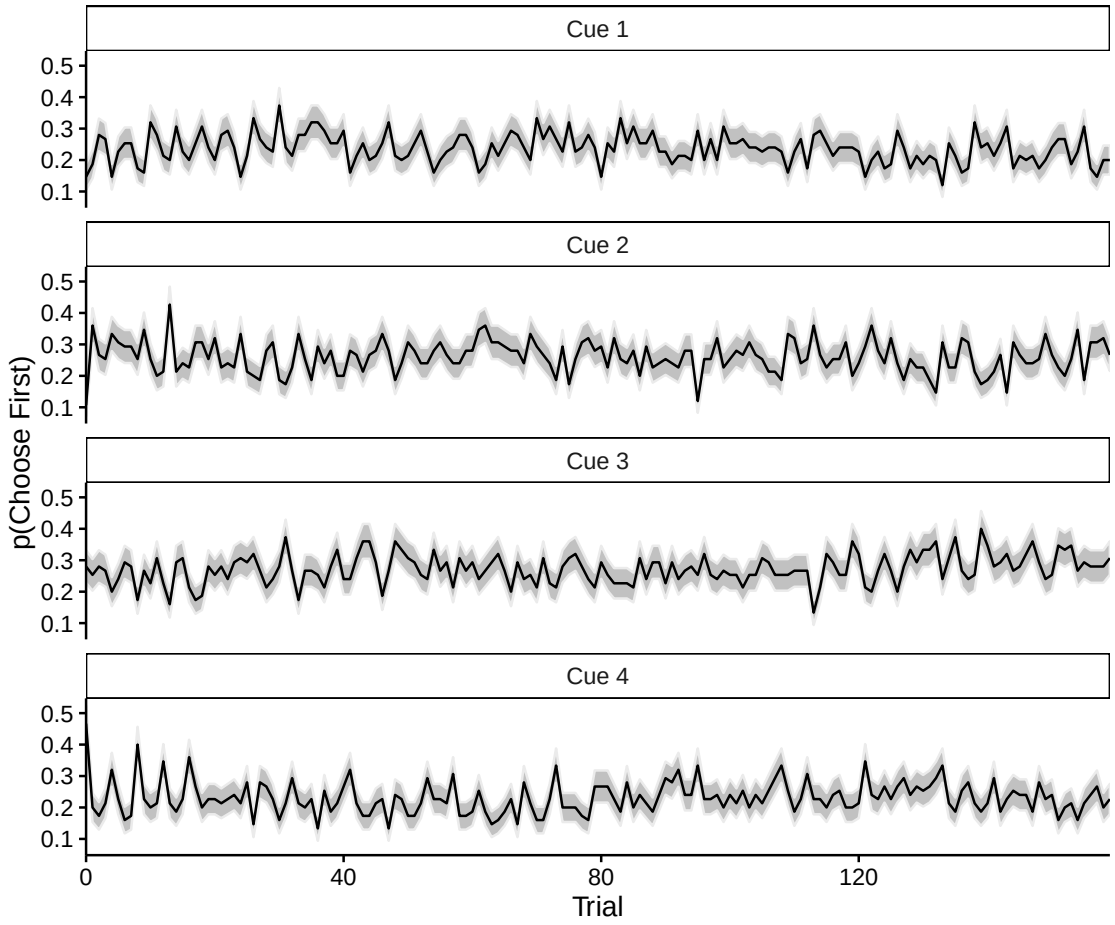


Figure A.2: Probability of choosing each cue, averaged over subjects, over the course of the test trials. Error ribbon represents a single proportion standard error, $\sqrt{p \cdot (1-p)/N}$

Bibliography

Anderson, J. R. (1990). *The Adaptive Character of Thought*. Psychology Press.

Antonakis, J. and Dietz, J. (2011). Looking for validity or testing it? the perils of stepwise regression, extreme-scores analysis, heteroscedasticity, and measurement error. *Personality and Individual Differences*, 50(3):409–415.

Bergert, F. B. and Nosofsky, R. M. (2007). A response-time approach to comparing generalized rational and take-the-best models of decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33(1):107.

Bowen, J. and Qiu, Z.-l. (1992). Satisficing when buying information. *Organizational Behavior and Human Decision Processes*, 51(3):471–481.

Box, G. E. P. and Draper, N. R. (1987). Empirical model-building and response surfaces. In *Probability and Mathematical Statistics: Applied Probability and Statistics*. John Wiley & Sons.

Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science*, 16(3):199–231.

Brunswik, E. (1952). *The conceptual framework of psychology*, volume 1. University of Chicago Press.

Brunswik, E. (1955). Representative design and probabilistic theory in a functional psychology. *Psychological Review*, 62(3):193.

Buttaccio, D. R., Lange, N. D., Thomas, R. P., and 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.

Chater, N., Oaksford, M., Nakisa, R., and Redington, M. (2003). Fast, frugal, and rational: How rational norms explain behavior. *Organizational behavior and human decision processes*, 90(1):63–86.

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist*, 34(7):571.

Dawes, R. M. and Corrigan, B. (1974). Linear models in decision making. *Psychological Bulletin*, 81(2):95.

Donoso, M., Collins, A. G., and Koechlin, E. (2014). Foundations of human reasoning in the prefrontal cortex. *Science*, 344(6191):1481–1486.

Dougherty, M. R., Franco-Watkins, A. M., and 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., Gettys, C. F., and Ogden, E. E. (1999). Minerva-dm: A memory processes model for judgments of likelihood. *Psychological Review*, 106(1):180.

Fellner, G., Güth, W., and Maciejovsky, B. (2009). Satisficing in financial decision makinga theoretical and experimental approach to bounded rationality. *Journal of Mathematical Psychology*, 53(1):26–33.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2014). *Bayesian data analysis*, volume 2. Taylor & Francis.

Gelman, A. and Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Gigerenzer, G. (1993). The bounded rationality of probabilistic mental models. In Manktelow, K. I. and Over, D. E., editors, *Rationality: Psychological and philosophical perspectives*, pages 284–313. Routledge, London.

Gigerenzer, G. (2010). Moral satisficing: Rethinking moral behavior as bounded rationality. *Topics in Cognitive Science*, 2(3):528–554.

Gigerenzer, G. and Brighton, H. (2009). Homo Heuristicus: Why Biased Minds Make Better Inferences. *Topics in Cognitive Science*, 1(1):107–143.

Gigerenzer, G. and Goldstein, D. G. (1996). Reasoning the fast and frugal way: models of bounded rationality. *Psychological Review*, 103(4):650.

Gigerenzer, G., Hoffrage, U., and Kleinbölting, H. (1991). Probabilistic mental models: a brunswikian theory of confidence. *Psychological Review*, 98(4):506.

Gigerenzer, G. and Todd, P. M. (1999). *Simple heuristics that make us smart*. Oxford University Press.

Glöckner, A., Betsch, T., and Schindler, N. (2010). Coherence shifts in probabilistic inference tasks. *Journal of Behavioral Decision Making*, 23:439–462.

Hadfield, J. D. (2010). MCMC methods for multi-response generalized linear mixed models: The MCMCglmm R package. *Journal of Statistical Software*, 33(2):1–22.

Hammond, K. R. (1990). *Functionalism and Illusionism: Can Integration be Usefully Achieved?* University of Chicago Press.

Hilbig, B. E., Erdfelder, E., and Pohl, R. F. (2010). One-reason decision making unveiled: A measurement model of the recognition heuristic. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 36(1):123.

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

Hogarth, R. M. and Karelaia, N. (2007). Heuristic and linear models of judgment: Matching rules and environments. *Psychological Review*, 114(3):733.

Karelaia, N. and Hogarth, R. M. (2008). Determinants of linear judgment: a meta-analysis of lens model studies. *Psychological Bulletin*, 134(3):404.

Lee, M. D. (2008). Three case studies in the Bayesian analysis of cognitive models. *Psychonomic Bulletin & Review*, 15(1):1–15.

Lee, M. D. (2010). How cognitive modeling can benefit from hierarchical Bayesian modeling. *Journal of Mathematical Psychology*.

Lee, M. D. and Newell, B. J. (2011). Using hierarchical Bayesian methods to examine the tools of decision-making. *Judgment & Decision Making*, 6(8).

Lee, M. D. and Zhang, S. (2012). Evaluating the coherence of take-the-best in structured environments. *Judgment and Decision Making*, 7(4):360.

Luan, S., Schooler, L. J., and Gigerenzer, G. (2014). From perception to preference and on to inference: An approach-avoidance analysis of thresholds. *Psychological Review*, 121(3):501–525.

Marewski, J. N. and Schooler, L. J. (2011). Cognitive niches: an ecological model of strategy selection. *Psychological Review*, 118(3):393.

Marr, D. (1982). *Vision: A Computational Investigation*. Freeman New York.

Martignon, L. and Hoffrage, U. (1999). Why does one-reason decision making work. In Gigerenzer, G. and Todd, P. M., editors, *Simple heuristics that make us smart*, pages 119–140. Oxford University Press.

Martignon, L. and Hoffrage, U. (2002). Fast, frugal, and fit: Simple heuristics for paired comparison. *Theory and Decision*, 52(1):29–71.

McCloskey, D. N. (1998). *The Rhetoric of Economics*. Univ of Wisconsin Press.

Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence.* University of Minnesota Press.

Newell, B. R. (2005). Re-visions of rationality? *Trends in cognitive sciences*, 9(1):11–15.

Newell, B. R. and Lee, M. D. (2011). The right tool for the job? comparing an evidence accumulation and a naive strategy selection model of decision making. *Journal of Behavioral Decision Making*, 24(5):456–481.

Newell, B. R., Rakow, T., Weston, N. J., and Shanks, D. R. (2004). Search strategies in decision making: The success of success. *Journal of Behavioral Decision Making*, 17(2):117–137.

Newell, B. R. and Shanks, D. R. (2003). Take the best or look at the rest? Factors influencing "one-reason" decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 29(1):53.

Newell, B. R., Weston, N. J., Tunney, R. J., and Shanks, D. R. (2009). The effectiveness of feedback in multiple-cue probability learning. *The Quarterly Journal of Experimental Psychology*, 62(5):890–908.

Parker, A. M., Bruine de Bruin, W., and Fischhoff, B. (2007). Maximizers versus satisficers: Decision-making styles, competence, and outcomes. *Judgment and Decision Making*, 2(6):342–350.

Payne, J. W., Bettman, J. R., and Johnson, E. J. (1988). Adaptive strategy selection in decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 14(3):534.

Payne, J. W., Bettman, J. R., and Johnson, E. J. (1992). Behavioral decision research: A constructive processing perspective. *Annual Review of Psychology*, 43(1):87–131.

Platzer, C., Bröder, A., and Heck, D. W. (2014). Deciding with the eye: How the visually manipulated accessibility of information in memory influences decision behavior. *Memory & Cognition*, 42(4):595–608.

Plummer, M. (2015). *rjags: Bayesian Graphical Models using MCMC*. R package version 3-15.

Plummer, M. et al. (2003). Jags: A program for analysis of bayesian graphical models using gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing*, volume 124, page 125. Technische Universit at Wien.

Rand, D. G., Peysakhovich, A., Kraft-Todd, G. T., Newman, G. E., Wurzbacher, O., Nowak, M. A., and Greene, J. D. (2014). Social heuristics shape intuitive cooperation. *Nature Communications*, 5.

Ravenzwaaij, D., Moore, C. P., Lee, M. D., and Newell, B. R. (2014). A hierarchical bayesian modeling approach to searching and stopping in multi-attribute judgment. *Cognitive science*, 38(7):1384–1405.

Rieskamp, J. and Otto, P. E. (2006). SSL: a theory of how people learn to select strategies. *Journal of Experimental Psychology: General*, 135(2):207.

Scheibehenne, B., Rieskamp, J., and Wagenmakers, E.-J. (2013). Testing adaptive toolbox models: A bayesian hierarchical approach. *Psychological Review*, 120(1):39.

Schwartz, B., Ward, A., Monterosso, J., Lyubomirsky, S., White, K., and Lehman, D. R. (2002). Maximizing versus satisficing: happiness is a matter of choice. *Journal of Personality and Social Psychology*, 83(5):1178.

Shalizi, C. (2015). Advanced data analysis from an elementary point of view. Unpublished.

Simon, H. A. (1955). A behavioral model of rational choice. *The Quarterly Journal of Economics*, 69(1):99–118.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639.

Stirling, W. C. and Goodrich, M. A. (1999). Satisficing games. *Information Sciences*, 114(1):255–280.

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

Todd, P. M. and Dieckmann, A. (2004). Heuristics for ordering cue search in decision making. In Saul, L. K. and Bottou, L., editors, *Advances in Neural Information Processing Systems*, pages 1393–1400. MIT Press.

Weng, W. and Gelman, A. (2014). Difficulty of selecting among multilevel models using predictive accuracy. *Statistics and Its Interface*, 7(1).

Yoon, K. P. and Hwang, C.-L. (1995). *Multiple Attribute Decision Making: An Introduction*, volume 104. Sage Publications.