universität
wien

MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

„When the Medium Is the Message: How Presentation Formats Influence Information Search"

verfasst von / submitted by
Charley Mingshuo Wu

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of
Master of Science (MSc)

Wien, 2015 / Vienna 2015

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	A 066 013
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Masterstudium Joint Degree Programme MEi :CogSci Cognitive Science
Betreut von / Supervisor:	Prof. Dr. Gerd Gigerenzer
Mitbetreut von / Co-Supervisor:	Dr. Björn Meder Dr. Jonathan Nelson Dr. Flavia Filimon

Preface

This thesis is the sum of more than 18 months of effort and represents my first fledgling attempt at understanding the complex mysteries of cognition.

This work has developed under the wing of Prof. Dr. Gerd Gigerenzer and The Center for Adaptive Behavior and Cognition (ABC) at the Max Planck Institute of Human Development in Berlin. My time as a guest of the ABC group has introduced me to the unique challenges of studying human behavior in an uncertain and changing world. But most of all, it has shown me what it means to be a good scientist.

This thesis would not all be possible without the constant support of my advisers Dr. Björn Meder, Dr. Jonathan D. Nelson, and Dr. Flavia Filimon, to whom I owe an immeasurable amount of gratitude. I give my thanks to Björn for always having the time to answer questions and for showing me the importance of treading carefully in dealing with complex data and a complex world. I owe to Jonathan my thanks in developing the maturity of my scientific thinking, as well as for being a selfless and dedicated mentor. To Flavia, I express my gratitude for asking the difficult questions and for leaving no stone unturned. It was only through standing on their shoulders that any of this is possible.

Before beginning my Masters program, I was merely a theoretical philosopher who knew how to write a few lines of code. In two short years as a Masters student in the MEi:CogSci program in Vienna, I have developed into a cognitive scientist with a fearsome arsenal of technical skills and a sharpened critical mind.

I owe a great deal to Mag. Elisabeth Zimmermann and Prof. Dr. Markus Peschl for designing such a competitive and thoughtful course of study, and for teaching me to be aware of the traps of conventional that hold back the study of cognition. As an international student more than 8000 km from home, Lisl and the Cognitive Science community made Vienna feel like a second home. To all my colleagues, thank you for our fruitful discussions on the nature of mind.

There are numerous people, without whom this work would not be possible. Thanks to John Wong for setting up the servers that hosted the experiments. Thanks to Dr. Amit Kothiyal for spending an afternoon with me working out the complex mathematics behind simple heuristics. Thanks to Prof. Dr. Laura Martignon for being an incredibly patient and thoughtful listener as I presented my initial results. Thanks to the entire ABC research group for their feedback and to Dr. Michelle McDowell for aiding my understanding of the past twenty years of Bayesian reasoning literature. Thanks to the other students who shared the HiWi room with me and kept me motivated, with special thanks to Mara for help with formatting my thesis, Clara for telling me about her hidden stash of snacks during late working evenings, Matthias Gloel for creating the original turtle stimuli, and to Matthias Hofer for feedback on an early draft.

Countless thanks to Agnieszka for teaching me the meaning of *Verwunderung* and reminding me about what is truly important in life.

Lastly, thanks to my parents for nurturing my curiosity about the world, yet teaching me the importance of discipline and hard work in my explorations.

Charley Mingshuo Wu
Berlin
September 2015

This research was supported by Grants ME 3717/2-2 to BM and NE 1713/1-2 to JDN from the Deutsche Forschungsgemeinschaft (DFG) as part of the priority program "New Frameworks of Rationality" (SPP 1516). CW was funded by ERASMUS+ while in Berlin.

Abstract

Even before it is possible to give the right answer, it is first necessary to find the right question. The ability to ask good questions is especially important in situations where resources are limited and search decisions must be chosen carefully. While previous work has shown that the method of presenting information makes a great deal of difference in Bayesian reasoning tasks, the main goal of this thesis is to push these findings further and determine if presentation formats can also influence how humans make information search decisions. A secondary goals is also addressed, concerning the relationship of search behavior to probability judgments and numeracy skill.

Four experiments were conducted, each using the same procedure, but operating in different probabilistic environments. In each experiment, subjects were asked to make an information search query with the goal of maximizing classification accuracy. Subjects were randomly assigned to one of 14 presentation formats, which fall under three main categories: the *probability formats* (numbers expressed as percentages), the *natural frequency formats* (numbers expressed through natural sampling), and *visual formats* (icon arrays, bar graphs, and dot diagrams). For each type of category, we constructed two variant conditions based on the type of information presented, either as a likelihood (probability of the outcome conditioned on the hypothesis) or in terms of the posterior probability (probability of the hypothesis conditioned on the outcome). For the probability and natural frequency formats, we also had an additional variant based on the presence or absence of the complement information (e.g. the probability of the alternative outcome), although this distinction was not applicable to the visual formats.

The results are twofold, with some environments producing a strong effect of presentation format on search decisions (accuracy-maximizing search decisions ranging from 27% to 86% of subjects in Experiment 2), as well as finding no evidence that this effect is mediated either by numeracy or the ability to judge the environmental probabilities. A new *frequency difference heuristic* has also been discovered, which exactly implements the strategy of maximizing classification accuracy, but with much less computational complexity than normative computational models. This heuristic allows us to explain differences between formats, such as why posterior icon arrays and bar graphs resulted in the highest proportion of accuracy-maximizing search decisions.

Zusammenfassung

Noch bevor es möglich ist, die richtige Antwort zu geben, ist es notwendig, die richtige Frage zu stellen. Insbesondere in Situationen, in denen Ressourcen begrenzt sind, ist die aktive Suche nach relevanten Informationen wichtig. Frühere Arbeiten haben gezeigt, dass die Darstellung von Informationen mittels verschiedener Präsentationsformate einen großen Einfluss auf Bayesianisches Schließen hat. Das Hauptziel dieser Arbeit ist herauszufinden, ob Präsentationsformate auch beeinflussen, wie Menschen aktiv nach Informationen suchen. Zwei weitere Ziele sind die Untersuchung des Zusammenhangs von Suchverhalten und Wahrscheinlichkeitsurteilen und der Einfluss interindividueller Differenzen im Verständnis von numerischen Informationen („numeracy“).

Vier Experimente wurden durchgeführt, jeweils mit der gleichen Vorgehensweisen und identischen Materialien, aber in verschiedenen probabilistischen Umwelten. In jedem Experiment wurden die Probanden gebeten, eine Suchentscheidung zu treffen (Auswahl von einem von zwei möglichen genetischen Tests, deren Ergebnis potentiell informativ sein könnte), um mittels der erhaltenen Informationen ein Objekt einer von zwei möglichen Kategorien zuzuordnen. Ziel war die Klassifikationsgenauigkeit zu maximieren. Die statistische Struktur der Umwelt wurde mittels einem von 14 Präsentationsformaten dargestellt. Diese Formate lassen sich drei Hauptkategorien zuordnen: „Wahrscheinlichkeitsformate“ (Zahlen in Prozentwerten ausgedrückt), „natürliche Häufigkeitsformate“ (Zahlen in Form von natürlichen Häufigkeiten), und „visuelle Formate“ (»Icon Array«, »Bar Graph«, und »Dot Diagram«). Für jedes Format wurden zwei Varianten untersucht: Entweder als „likelihood“ (Wahrscheinlichkeit des Ergebnisses bedingt auf die Hypothese) oder als »Posterior Probability« (Wahrscheinlichkeit der Hypothese gegeben das Ergebnis). Für die verschiedenen Wahrscheinlichkeitsformate und die natürlichen Häufigkeitsformate wurde jeweils eine weitere Variante untersucht, in der die Komplementärwahrscheinlichkeiten (z.B. die Wahrscheinlichkeit des alternativen Ausgangs eines Tests) explizit gegeben wurde.

Die Ergebnisse zeigen, dass in bestimmten Umwelten das Präsentationsformat einen starken Einfluss auf das Suchverhalten hat. Zum Beispiel wählten in Experiment 2 in Abhängigkeit vom Präsentationsformat zwischen 27% bis 86% der Probanden den informativeren Test, d.h., denjenigen, der die Klassifikationsgenauigkeit erhöht. Die Studien zeigen auch, dass dieser Effekt weder durch die akkurate Beurteilung der statistischen Struktur der Umwelt, noch durch individuelle Differenzen im Umgang mit numerischen Information („numeracy“) vermittelt ist. Eine neue »Frequency Difference Heuristic« wird vorgestellt, die es ermöglicht, den Test mit der höheren erwarteten Klassifikationsgenauigkeit zu identifizieren, aber komputational einfacher ist als normative Modelle. Diese Heuristik ist eine mögliche Erklärung für den Einfluss der unterschiedlichen Formate auf das Suchverhalten, zum Beispiel warum »posterior icon arrays« und »posterior bar graphs« zu dem höchsten Anteil rationaler Suchentscheidungen führten.

Contents

List of Figures

List of Tables

List of Acronyms

OED Optimal Experimental Design 7

KL-Divergence Kullback-Liebler Divergence 11

AMT Amazon's Mechanical Turk platform 28

HITs Human Interaction Tasks 28

BNT Berlin Numeracy Test 32

MAE Mean Absolute Error 29

HIV Human Immunodeficiency Virus 4

Introduction 1

"Judge a man by his questions rather than by his answers."

— Voltaire

Even before it is possible to give the right answer, it is first necessary to find the right question. How we choose our questions determines the kind of information that is acquired, thus, knowing how to pose good questions is essential for making informed decisions. The concept of information search applies to any decision making task where the goal is to acquire information about the world, such as a wild animal foraging for food, or a human's eye movements trying to pick out a familiar face in the crowd.

The ability to ask good questions is especially important in situations where resources are limited and search queries must be chosen carefully. Consider for example, a doctor choosing a test for diagnosing a patient. Not all information search queries are equally useful, and a good doctor must carefully consider the individual true and false positive rates of each test along with the base rate of the disease in order to decide how best to diagnose a patient. Yet there is a wealth of medical decision making literature showing that doctors are notoriously bad at reasoning about probabilistic information, specifically in integrating base rates and new data (e.g. from a test result) to form an updated belief about a hypothesis (Anderson & Gigerenzer, 2012; Covey, 2007; Forrow et al., 1992; Gigerenzer, 1998; Gigerenzer et al., 2007; Hoffrage et al., 2000). This type of probabilistic reasoning task is referred to as Bayesian reasoning or Bayesian inference (see Meder & Gigerenzer, 2014 for a review).

If doctors have such difficulties reasoning about basic probabilities, how likely are they to succeed at an information search task, which represents a much more complex problem? More broadly, how do humans in general make search decisions and how can we guide them to ask better questions?

1.1 The Medium Is The Message

1.1.1 Aiding Bayesian Reasoning Through Presentation Formats

Gigerenzer and Hoffrage (1995) have shown that the poor performances in Bayesian reasoning tasks were largely due to an inadequacy of the *presentation format*, rather than a *cognitive bias* (Tversky & Kahneman, 1974). By changing how we express statistical information, from the standard probability format (expressed in terms of a single-event probability) to using natural frequencies (expressed as outcomes of natural sampling), it is possible to make Bayesian reasoning problems simpler and drastically improve the proportion of accurate probability judgments (Hoffrage et al., 2002). Indeed, even 6th grade children given natural frequencies are able to solve Bayesian inference problems with the same level of accuracy as adults who were given the standard probability format (Zhu & Gigerenzer, 2006).

An important difference is that natural frequencies express information in terms of *natural sampling*, and in doing so, represents the Bayesian reasoning problem in terms that are more similar to how an individual experiences events in daily life (Meder & Gigerenzer, 2014), as well as simplifying the necessary calculations (Gigerenzer & Hoffrage, 1995). The use of natural sampling evokes the ecological rationality of how humans have traditionally acquired information about the world prior to the recent advent of probability theory (Gigerenzer, 2000). Information acquired through natural sampling is conveyed through the absolute frequencies of events, which inherently preserves base rate information. This is in contrast to how the standard probability format expresses information, by systematically re-normalizing probabilistic information to an artificially fixed base rate (Kleiter, 1994).

1.1.2 Bayesian Reasoning Terminology

Bayesian Reasoning refers to a task where a prior belief is integrated with new data (e.g., a test outcome) in order to form an updated belief. Often this is tested experimentally by asking a subject to infer a single-point posterior probability estimate from information about the prior probabilities and likelihoods (also known as *Bayesian inference*).

Base Rate or *prior probability* is the prior belief about a hypothesis, unconditioned on any evidence, e.g., the prior probability that a patient has a disease, before any tests are performed. Written as $P(c_i)$ where c_i is one of the n possible classes in C that an unknown item could belong to.

Likelihoods refer to the probabilities of a feature or query outcome, conditioned on a specific hypothesis, e.g., the likelihood that a patient will receive positive a test result q_j given that they actually have the disease c_i. Written as $P(q_j|c_i)$ where q_j is one of the m possible outcomes of a query Q.

Posteriors refer to the probability of a hypothesis conditioned on the data, e.g., the probability that a patient has the disease c_i, given the test outcome q_j. Written as $P(c_i|q_j)$.

Natural Frequencies are a method of presenting probabilistic information that uses the *joint frequencies* of two events to describe an event, e.g., the number of patients who have a positive test result as well as actually having the disease. Written as, $N(c_i \wedge q_j)$. This format preserves base rate information and is an approximation of *natural sampling*.

1.1.3 Natural Sampling Inspired Formats

The set of presentation formats that improve Bayesian inference also includes formats that use visualizations based on natural frequencies, which have been shown to solicit superior estimation accuracy when compared to merely numerical representations using only words and numbers (Brase, 2009; Galesic et al., 2009; Micallef et al., 2012). Additionally, since the natural frequency format expresses information as an approximation of natural sampling (Hoffrage et al., 2002), there have also been studies that examined the effect of experience-based learning, where subjects directly experienced single events that were naturally sampled from the probabilistic environment (Lindeman et al., 1988; Medin & Edelson, 1988). These studies showed that yet again, natural sampling inspired formats prove to be more effective at soliciting accurate judgments about Bayesian statistics compared to the standard probability format.

1.2 From Bayesian Reasoning to Information Search

Due to the wealth of literature showing that natural sampling and natural frequency formats systematically improve Bayesian inference skills, the main focus of this investigation is to examine if the effect of presentation formats also transfers to information search tasks, which requires the understanding of a more complex probabilistic environment. Whereas all of the previous research on the effects of presentation formats on Bayesian inference is based solely on soliciting a single probability estimate, information search represents a decision making task where one is required to consider many aspects of a probabilistic environment (see Fig 1.1). For a normative model to determine which query is most useful, it must integrate the expectation of all possible query outcomes across multiple queries, and ultimately judge which query has the highest expectation of useful information. This present study seeks to push the science further in regards to the effects of presentation formats on human reasoning and finally address the open question posed by Hoffrage et al. (2002) about how far do the

effects of the natural frequency format extend beyond Bayesian reasoning tasks (i.e. merely estimating the posterior probability of a binary variable).

1.2.1 An Example of an Information Search Task

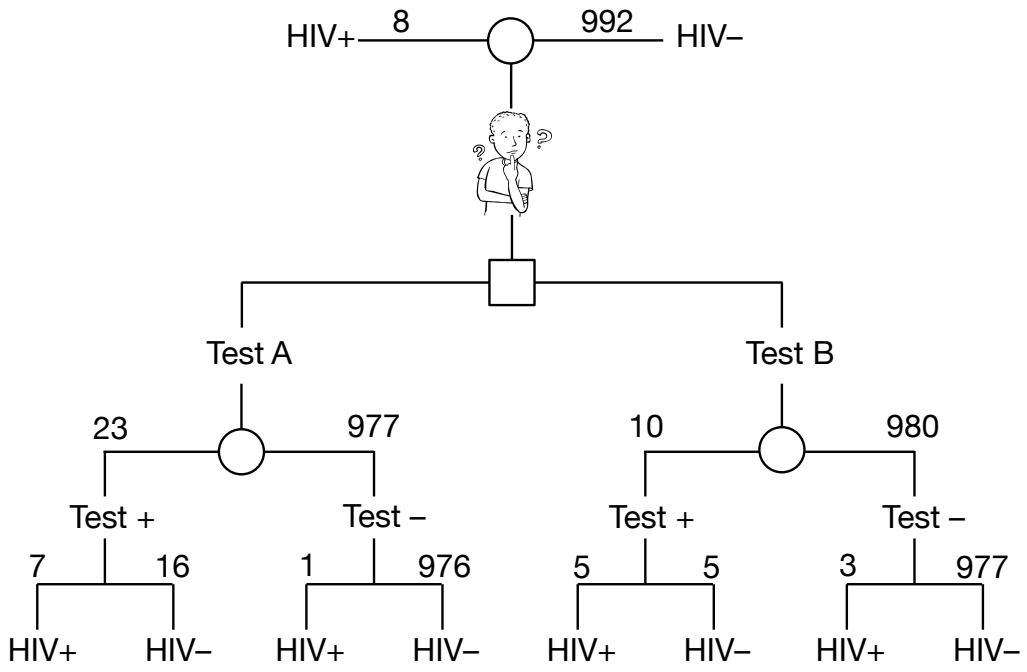


Figure 1.1: An example of an information search task shown as a natural frequency tree. Suppose a doctor receives a new patient who wants to be tested for HIV. Given the existing data above, which test is better for diagnosing the patient?

Suppose a doctor in a far away and fictional country receives a patient who is uncertain if he/she has contracted *Human Immunodeficiency Virus* (HIV) (Fig. 1.1). Out of a random sample of 1,000 individuals from the country's population, 8 are HIV+ while 992 are HIV-. The doctor must decide which test is better for diagnosing the patient. *Test A* has a high false positive rate, with 16 out of 23 positive results for patients who are HIV-. However, Test A has a very low false negative rate, with only 1 out of 977 negative results for patients who are HIV+. On the other hand, Test B has a lower false positive rate (only 5 out of 10), but also has more false negatives (3 out of 980). Which test is more useful? *This is not a trivial task* even when presented in terms of natural frequencies. Compared to an analogous Bayesian reasoning task, which would involve computing the posterior probability of a single test outcome, an information search task requires an individual to reason about a larger set of statistical information.

1.2.2 Previous Studies

Currently, the only research on the effects of presentation format on information search comes from Nelson et al. (2010), which showed that a session of 1-2 hours of experience-based learning yielded vast improvements to the proportion of search queries maximizing classification accuracy, when compared to subjects who were only given the standard probability format (proportion of accuracy-maximizing search decisions ranging from 28% to 82%). Thus, a practical basis for this paper comes from the desire to find descriptive presentation formats that are as effective as experience-based learning in soliciting useful information search queries, but without requiring extensive training times (see Table 1.1).

Table 1.1: Existing research on presentation formats.

Format	Bayesian Reasoning	Information Search
Standard Probability	✗ (Gigerenzer & Hoffrage, 1995)	✗ (Nelson et al., 2010; Skov & Sherman, 1986; Baron et al., 1988)
Natural Frequencies, Icon Array, Bar Graph, etc...	✓ (Cosmides & Tooby, 1996; Galesic et al., 2009; Brase, 2009; Micallef et al., 2012)	?
Experience-Based Learning	✓ (Medin & Edelson, 1988; Lindeman et al., 1988)	✓ (Nelson et al., 2010; Meder & Nelson, 2012)

Note. Previous studies on the effects of presentation formats on Bayesian reasoning and information search. ✓ denotes that the format has been found to be helpful, whereas ✗ denotes that it has not. **?** indicates that no previous studies have investigated the effects of descriptive formats (such as natural frequencies, icon arrays, or bar graphs) on information search.

1.3 Research Questions

The work presented in this thesis examined two main questions. First, to what extent search decisions are influenced by presentation format, and secondly, if this effect is mediated by the ability to provide accurate judgments about the underlying environmental probabilities or by an individuals ability to reason about numbers in general (numeracy skill). The first question serves the practical purpose of allowing us to make claims about how statistical

information *should* be presented in order to improve the rationality of search decisions, while the second question addresses the theoretical problem of developing an understanding of how information search decisions are related to Bayesian reasoning.

One approach to modeling human search behavior is to first describe the usefulness of each question and then to select the most useful question. This intuition is captured by the framework of Bayesian *Optimal Experimental Design* (OED), which have been widely used to construct normative models of information acquisition in cognitive psychology (Savage, 1954; Baron, 1985; Skov & Sherman, 1986; Klayman & Ha, 1987; Slowiaczek et al., 1992; Oaksford & Chater, 1994; Nelson, 2008). The OED framework provides a probabilistic approach for assessing the usefulness of a question, based on existing knowledge about the environment along with expectations about the possible outcomes (Nelson, 2008). OED models provide a computational-level description of an information search task (Marr, 1982), allowing us to quantify the usefulness of a question and make predictions about human behavior. Currently, OED models provide some of the best available accounts of human search behavior in the probabilistic classification literature (Nelson et al., 2010; Markant & Gureckis, 2012). It has also been shown in some cases that simple heuristic strategies can approximate or even exactly implement OED models, thereby providing a link to more psychological plausible mechanisms (Navarro & Perfors, 2011; Nelson, 2005). We use OED models as tools for analyzing and understanding the inductive problem of how to value a question, while at the same time looking towards heuristics as potential explanations about the underlying cognitive processes.

2.1 Modeling Usefulness

The usefulness of a question can be expressed in terms of the expectation of the usefulness of each of the possible outcomes (Savage, 1954). We describe several different OED models,

which each use different *informational utility functions* (also known as sampling norms; Nelson, 2005) to model the usefulness of a query outcome. We use Q to denote a query, where $q_j, ..., q_m$ are the m possible outcomes of the query. Equation 2.1 shows the generic form used by OED models to calculate the expected usefulness of a query, $eu(Q)$.

$$eu(Q) = \sum_{j=1}^{m} P(q_j)u(q_j) \tag{2.1}$$

OED models have been able to explain human behavior in a wide range of information acquisition tasks, such as Wason's selection task (Ginzburg & Sejnowski, 1996; McKenzie, 2004; Nelson & Movellan, 2001; Oaksford & Chater, 1994), categorization tasks (Nelson et al., 2010; Markant & Gureckis, 2012), as well as explaining human eye movement in visual search tasks (Butko & Movellan, 2008; Nelson & Cottrell, 2007; Rehder & Hoffman, 2005).

2.1.1 Different Types of Utility

There are many possible goals in an information search task, just as there are many ways to define utility. Because it is outside the scope of this study to claim which model is the normatively better strategy—specific to a certain search environment—or which model is the most accurate description of human cognition, we present an information search task where the goal has been explicitly defined as *the maximization of classification accuracy*. This goal exactly corresponds to the probability gain model (or any equivalent strategy), which Nelson et al. (2010) suggests is the primary basis for how humans subjectively value information in categorization tasks.

It should also be noted that we are only considering disinterested utility functions, which do not factor in asymmetrical differences in payoffs or search costs (Meder & Nelson, 2012). In medical decision making for example, a correct diagnosis may not balance out a false negative, and different tests may cost different amounts of time or money. However, incorporating payoffs and search costs is outside the scope of this study, because it is a separate question to model how humans perceive and react to these differences. Instead, we use a cover story (introduced in Chapter 3) where payoffs are neutral and there are no differences in search cost.

2.2 Informational Utility Functions

2.2.1 Probability Gain

The probability gain model values a search query in terms of its expected probability of making an accurate classification, seeking to maximize classification accuracy. This has been described by Martignon & Krauss (2003) as maximizing the average validity, because probability gain will average the likelihood of making an accurate judgment across all the possible outcomes of a query. In each of the probabilistic environments outlined in this thesis, the probability gain strategy most highly values the query for which all possible outcomes are an improvement over the initial best guess (i.e., relying solely on base rate information prior to performing a search query) in terms of facilitating classification accuracy of an unknown item. Probability gain has been used as a normative benchmark for a variety of information acquisition tasks, such as in experience-based categorization tasks (Nelson et al., 2010), explaining eye movements where the goal is to find a target in a cluttered environment (Najemnik & Geisler, 2008), and in medical test selection task (Baron et al., 1988).

The utility of a query outcome computed by probability gain, $u_{PG}(q_j)$, is the increase in the probability of making a correct classification of an unknown item $C = \{c_i, ..., c_n\}$, where c_i is one of n possible classes.

$$u_{PG}(q_j) = \max_{i=1}^{n} P(c_i|q_j) - \max_{i=1}^{n} P(c_i) \tag{2.2}$$

The first term in equation 2.2 is the probability of making an accurate classification given the query outcome. The second term is the probability of making a correct classification before any query is made, based solely on the prior class probabilities (i.e., base rate). The max function used in both terms, merely chooses the probability corresponding to the maximally likely class, since in a one-shot classification decision, choosing the most likely class is a reasonable strategy.

2.2.2 Information Gain

Information gain quantifies how much a query outcome reduces the uncertainty about the class membership of an unknown item, where uncertainty[1] is defined using Shannon entropy

[1]Uncertainty is an umbrella term with various meanings. In the sense modeled by Shannon entropy, uncertainty is taken as the uncertain outcome of a risky event, where all outcomes and the probabilities of their occurrence are known. This is quite different from the Knightian sense of uncertainty, where potential outcomes are known but their probabilities are not (Knight, 1923). See Meder et al. (2013) for a discussion on different types of uncertainty.

(Lindley, 1956). The information gain of a query outcome, $u_{IG}(q_j)$, is modeled as the amount it reduces Shannon entropy about the true class of the unknown item. For a binary variable, Shannon entropy is highest when the probability distribution is equiprobable, and approaches zero as one of the probabilities reaches certainty. This drop-off of Shannon Entropy changes logarithmically with the probability distribution, and because of this, information gain displays a very strong preference for queries where one of the possible results returns a high-level of certainty about the true class of the unknown item. While the reduction of Shannon entropy can also improve classification accuracy, there are situations where information gain and probability gain disagree about which query is more useful[2] (Nelson, 2005; Nelson et al., 2010). In Experiments 2 and 3, we specifically study search behavior in environments where the query with the highest expected reduction of Shannon entropy does not maximally increase the expected classification accuracy.

Formally, for a query outcome, q_j, the utility given by information gain is calculated as the difference between the initial entropy, $H(C)$, and the posterior entropy given the results of the query outcome, $H(C|q_j)$.

$$u_{IG}(q_j) = H(C) - H(C|q_j) \tag{2.3}$$

The initial entropy is the initial uncertainty about which class an unknown item belongs to, prior to any questions being asked. The choice of $\log_2$ is arbitrary, but allows us to express information gain in terms of bits.

$$H(C) = \sum_{i=1}^{n} P(c_i) \log_2 \frac{1}{P(c_i)} \tag{2.4}$$

The posterior entropy given the results of a query (equation 2.5) can be described as the Shannon Entropy of the resulting posterior probability distribution.

$$H(C|q_j) = \sum_{i=1}^{n} P(c_i|q_j) \log_2 \frac{1}{P(c_i|q_j)} \tag{2.5}$$

2.2.3 Impact

Impact provides a method of calculating the usefulness of a question based on its ability to change one's beliefs about the environment (Wells & Lindsay, 1980; Nelson, 2005). In

[2]Using simulations we compared the divergence of optimality between information gain and probability gain across different environments. For all possible binary search tasks (i.e., binary class with two binary features), there is agreement in 51% of environments, disagreement in 4%, and no preference for one or both utility functions in 45% of environments.

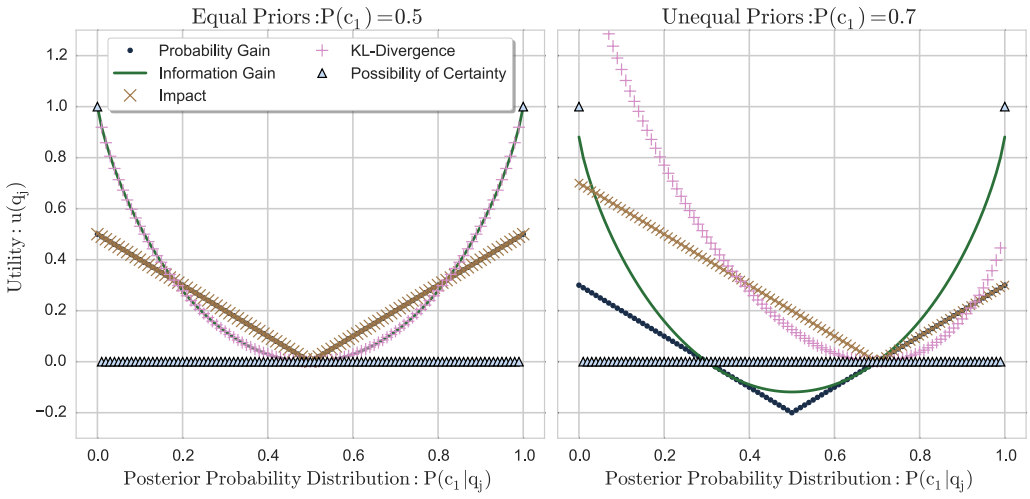


FIGURE 2.1: The utility of a query outcome q_j across the binary probability distribution of a single posterior probability $P(c_i|q_j)$. On the left, the priors are set to $P(c_i) = 0.5$ for two possible classes of C, and on the right, the priors are unequal, with one of the classes, c_1 having a prior probability of $P(c_1) = 0.7$. *Note.* Probability gain and Impact exactly overlap, as well as information gain and KL-Divergence in the figure on the left.

the case of a classification task, this would be the change from the prior to the posterior probability of a hypothesis, after a query has been made.

Formally, this is measured as the absolute difference between the prior probability, $P(c_i)$, and the posterior probability for a specific outcome of a query, $P(c_i|q_j)$. This absolute difference is than divided by n, which is the total number of possible classes in C. In the unique case of a binary category classification task where the base rates are equiprobable (used in Experiment 1), impact has been shown to be exactly analogous to probability gain (Nelson, 2005; see Fig 2.1 left). However, in Experiment 2 and 3 in this thesis, we introduce search environments where probability gain and impact disagree about which query is more useful.

$$u_{Impact}(q_j) = \frac{1}{n}\sum_{i=1}^{n} abs\,|P(c_i|q_j) - P(c_i)| \tag{2.6}$$

2.2.4 Kullback-Leibler Divergence

Kullback-Liebler Divergence (KL-Divergence) also quantifies the usefulness of a question as a function of its ability to change one's beliefs about the environment (Kullback & Leibler, 1951). However, rather than using the absolute difference between the prior and posterior beliefs, KL-Divergence takes the log of the ratio (equation 2.7). This log is then weighted by the posterior probability $P(c_i|q_j)$. KL-Divergence is exactly equivalent to information gain,

because they yield the same expected utility values for every possible question (Oaksford & Chater, 1994); however, they calculate different utilities for each query outcome, $u(q_j)$ (see Fig 2.1 right). In the context of search behavior in the probabilistic environments described in this paper, when we describe a query as being optimal in terms of information gain, it should be understood that KL-Divergence also considers the same query to be optimal.

$$u_{KL}(q_j) = \sum_{i=1}^{n} P(c_i|q_j) \log_2 \frac{P(c_i|q_j)}{P(c_i)} \tag{2.7}$$

2.2.5 Probability of Certainty.

The probability of certainty model was first described in Nelson et al. (2010) as a type of heuristic strategy for valuing the usefulness of a question, where a query is useful insofar as it carries the probability of a certain result (equation 2.8). A query outcome has a utility of 1 if there is certainty about the hypothesis, e.g., $P(c_i|q_j) = 1$, otherwise it is 0. Crupi et al. (2015) showed that a utility function numerically very similar to probability of certainty can be derived by substituting a high-order Tsallis (1988) entropy in place of Shannon entropy, within the information gain model. Thus, probability of certainty can also be viewed as a possible OED approach for evaluating the usefulness of questions. The probability of certainty model is of interest to Experiments 3 and 4, where we introduce certain query outcomes.

$$u_{PofC}(q_j) = \begin{cases} 1, & \text{if } \max\limits_{i=1}^{n} P(c_i|q_j) = 1 \\ 0, & \text{otherwise} \end{cases} \tag{2.8}$$

2.3 Heuristic Explanations

While the OED models are used to provide normative guidance about information search tasks, we also consider the possibility that human search behavior is best modeled using simple heuristic strategies. These strategies are cognitive shortcuts that do not make use of all the available information, but instead, use features of the presentation formats to arrive at a search decisions. Under certain conditions, heuristic strategies can exactly implement or approximate the same behavior as the more complex OED models (Meder & Nelson, 2015; Klayman & Ha, 1987; Navarro & Perfors, 2011). The term heuristics can be applied to any strategy for information search that directly identifies which query to select, without calculating intermediate quantities such as uncertainty (i.e., Shannon entropy), expected classification accuracy, or change in hypothesis strength. At the end of this thesis (Chapter 5),

we introduce a newly discovered heuristic in the general discussion, which exactly implements the probability gain model.

2.3.1 Likelihood Difference Heuristic

The likelihood difference heuristic (Nelson, 2005; Slowiaczek et al., 1992, p. 402) is one such heuristic strategy, which merely chooses the query with the maximal difference in feature likelihoods. Although it seems simple, the likelihood difference heuristic has been shown to exactly implement impact (for a proof see Nelson, 2005). The expected utility of a query can be modeled in the same terms as the OED models, but does need to consider base rates, posteriors, or the marginal likelihood of each query outcome (Equation 2.9).

$$eu_{likelihoodDifference}(Q) = \max_{j=1}^{m} |P(q_j|c_1) - P(q_j|c_2)| \tag{2.9}$$

In a medial diagnosis scenario, one could execute the likelihood difference heuristic by taking the true positive rate of a test and subtracting the false positive rate; this remainder would be the expected utility of the test. This process can be repeated for all tests, and the query with the largest difference between true and false positive rates would be the most useful test, according to both the likelihood difference heuristic and impact[3].

2.4 Initial Assumptions and Hypotheses

Alongside the goal of finding presentation formats that can influence search behavior, we also wanted to examine *how* a potential influence would be mediated. Specifically, if the influence of format would be mediated by a participant's understanding of the probabilistic environment, or by their ability to reason about numbers in general (numeracy skill). Because of the existing literature showing that presentation formats can improve Bayesian reasoning, we hypothesized that formats would also influence the ability to make explicit probability judgments about the task environment. Moreover, we hypothesized that probability judgment accuracy would be a mediator variable for search behavior. Additionally, research on numeracy by Cokely et al. (2012) has found that numeracy skill is a strong predictor for a variety of different cognitive tasks, such as risk perception in numerical and visual presentation formats (Bodemer et al., 2014; Brown et al., 2011). Thus, we expected to find a link between numeracy and the accuracy of probability judgments. The potential relationships shown in Fig. 2.2 represent our initial hypotheses.

[3]One caveat is that the likelihood difference heuristic only applies for binary classification tasks with binary features, whereas impact can apply in situations where multi-valued features and any number of categories (Nelson, 2005).

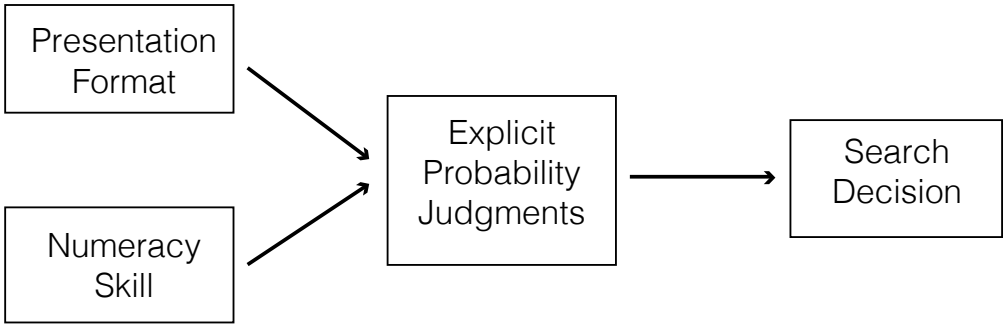


Figure 2.2: The hypothesized relationships between format, numeracy skill, probability judgments, and information search.

Presentation Formats 3

In order to examine the effect of presentation format on search behavior, we devised an information search task involving the classification of artificial turtle stimuli. This allowed us to compare different presentation formats in their ability to solicit search decisions that corresponded to the goal of maximizing classification accuracy. The Turtle Island story represents a two category, binary feature search task, where there are two possible queries, each of which has a binary outcome. This has the same structure as the HIV example, given in Chapter 1.

3.1 Turtle Island Cover Story

For each experiment described in this thesis, participants were told that 100 turtles live on a remote island, with each animal being either a "Freqosian" or "Bayosian" turtle. The two species of turtles look identical, but differ in their genetic makeup, specifically in the distributions of the DE and the LM genes. Therefore, the identification of a turtle can be aided with the help one of two DNA tests, each of which has a binary result. The DE test yields truthful information on whether the D form or the E form of the DE gene is present in the turtle, while the LM test yields truthful information whether the L form or the M form of the LM gene is present. While both species each possess one form of the DE and one form of the LM gene, the extent that these gene forms are present vary between the two species of turtles. Because the difference across species is the *distribution* of the DE and LM genes, knowing that a turtle has a specific form of a gene provides *probabilistic* information about the true species of the turtle. This is a binary information search task, where there are two possible classes, $C = \{Bayosian, Freqosian\}$, and two possible search

queries, $Q = \{DE, LM\}$, with each query having two possible outcomes, $Q_{DE} = \{D, E\}$ and $Q_{LM} = \{L, M\}$. The goal is to choose the test that maximizes the probability of correctly classifying a randomly encountered turtle based on the usefulness of each test. Appendix C provides screenshots from the experiment, with this same information as it was presented to participants.

3.2 Overview of Formats

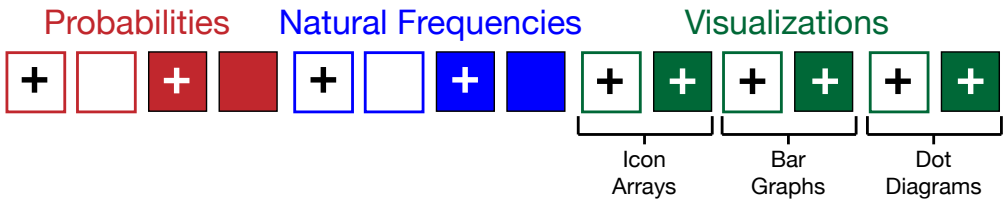


Figure 3.1: The design of the 14 presentation formats we examined, which included four probabilities, four natural frequencies, and six visual formats. Likelihood variations are shown as empty boxes, whereas posterior variations are shown as filled in boxes. "+" denotes that a format includes the complement information.

We studied 14 different presentation formats (see Fig 3.1), which fall under 2 main mediums: *numerical formats* (probabilities or natural frequencies expressed with words and numbers) and *visual formats* (visualizations of natural frequencies). Within these categories, we applied several variations using a between-subject factorial design (described below). Each different presentation format preserves informational equivalence across the various conditions, but not computational equivalence (Simon, 1978). All of the necessary information for computing the expected utilities for each possible OED model or heuristic strategy is included within each format. However, there are different advantages and disadvantages relative to what information needs to be extracted or computed from the format. For example, each of the informational utility functions that we have presented requires the posterior probabilities, i.e., the probability of a turtle being Bayosian or Freqosian given the outcome of a particular DNA test, $P(class|outcome)$, in the computation of the utility of a query outcome. Since the computational complexities of calculating the posterior probabilities are not equivalent across all the presentation formats, this may be a factor influencing search behavior, if indeed knowing the posterior probabilities is an important aspect in how humans make search decisions (i.e., if there is an overlap between OED models and human psychological processes). One of the main hypotheses presented by Hoffrage and colleagues (2002) to explain the advantage of natural frequencies over the standard probability format is that it reduces the computational complexity of computing the posterior

probabilities. Thus, it may also be the case that individuals presented with formats that facilitate the calculation of posterior probabilities are able to make better information search decisions, independent from other factors.

3.2.1 Numerical Formats

The probability and natural frequency formats (red and blue boxes in Fig 3.1, respectively) collectively comprise the numerical formats, which use words and numbers to express probabilistic information. The eight numerical formats were constructed through a 2 (format type: probabilities or natural frequencies) by 2 (information type: likelihoods or posteriors) by 2 (complement: absent or present) factorial design. An example of each type of format is shown in Table 3.1 below, while the full description is available in Appendix B. Format type as a factor tests the difference between using probabilities (numbers expressed as percentages) and natural frequencies (numbers expressed as outcomes of natural sampling), which have been shown to yield substantial differences in Bayesian reasoning tasks (Gigerenzer & Hoffrage, 1995). We also varied the type of information presented, with a likelihood and a posterior variation. The likelihood formats provide the base rate, i.e., $P(class)$, along with the likelihoods of each test outcome conditioned on the class, ie., $P(outcome|class)$. The posterior formats provide the marginal probabilities, i.e., $P(outcome)$ independent of class, as well as the posterior probability of an item belonging to a class conditioned on the test outcome, i.e., $P(class|outcome)$. It should be noted that the likelihood and posterior variants of the natural frequency formats have the same set of numbers (since they are outcomes of natural sampling and not re-normalized into absolute probabilities), but are grouped by class in the likelihood variant, and grouped by test outcome in the posterior variant (see Table 3.1). We also had an additional variant based on the presence or absence of the complement information (i.e., the complementary information for each of the binary variables: if $P(Bayosian) = 0.7$ then the complement provides the information $P(Freqosian) = 0.3$). In previous studies, it was found that people were more sensitive to the informativeness of an answer when provided with the complement (Rusconi & McKenzie, 2013). This factor was not applied to the visual formats, because it did not seem appropriate to exclude the complement from the chosen visualizations.

Table 3.1: An example of the numerical formats, expressed in terms of the Experiment 1 task environment

Format	Quantity	Example
 Standard Probability	$P(class)$	Consider a turtle picked at random from the **100** turtles on the island: The probability that it is a Bayosian turtle is **50%**.
	$P(outcome\|class)$	If a turtle is a Bayosian, then the probability that it has the D form of the DE gene is **10%**.
 Standard Probability with Complement	$P(class)$	Consider a turtle picked at random from the **100** turtles on the island: The probability that it is a Bayosian turtle is **50%**, and the probability that it is a Freqosian turtle is **50%**.
	$P(outcome\|class)$	If a turtle is a Bayosian, then the probability that it has the D form of the DE gene is **10%**, and the probability that it has the E form is **90%**.
 Posterior Probability	$P(outcome)$	Consider a turtle picked at random from the **100** turtles on the island: The probability that it has the D form of the DE gene is **45%**.
	$P(class\|outcome)$	If a turtle has the D form of the DE gene, then the probability that it is a Bayosian turtle is **11%**.
 Posterior Probability with Complement	$P(outcome)$	Consider a turtle picked at random from the **100** turtles on the island: The probability that it has the D form of the DE gene is **45%**, and the probability that it has the E form is **55%**.
	$P(class\|outcome)$	If a turtle has the D form of the DE gene, then the probability that it is a Bayosian turtle is **11%**, and the probability that it is a Freqosian turtle is **89%**.
 Natural Frequency	$N(class)$	Out of the **100** turtles on the island, **50** are Bayosian turtles.
	$N(outcome \wedge class)$	Out of the **50** Bayosian turtles, **5** turtles have the D form of the DE gene.
 Natural Frequency with Complement	$N(class)$	Out of the **100** turtles on the island, **50** are Bayosian turtles and **50** are Freqosian turtles.
	$N(outcome \wedge class)$	Out of the **50** Bayosian turtles, **5** turtles have the D form of the DE gene, and **45** turtles have the E form of the gene.
 Posterior Frequency	$N(outcome)$	Out of the **100** turtles on the island, **45** have the D form of the DE gene.
	$N(class \wedge outcome)$	Out of the **45** turtles with the D form of the DE gene, **5** are Bayosian turtles.

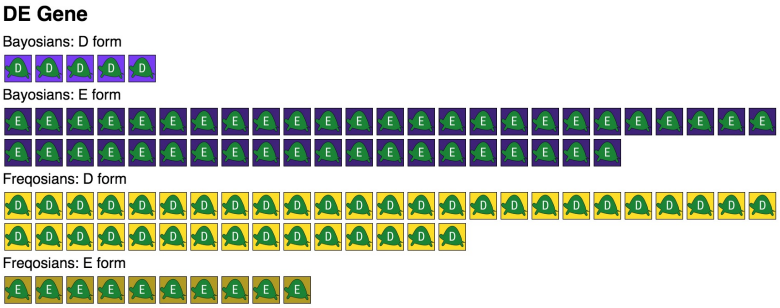
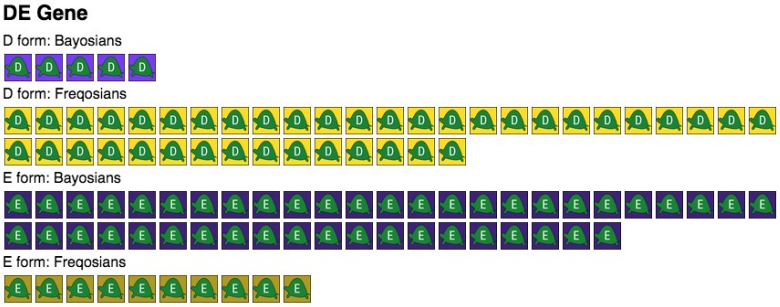
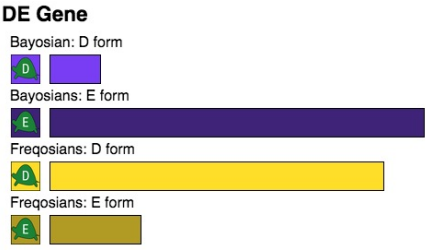
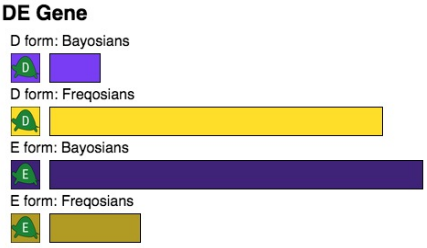
+ Posterior Frequency with Complement	$N(outcome)$	Out of the **100** turtles on the island, **45** have the D form of the DE gene, and **55** turtles have the E form of the gene.
	$N(class \wedge outcome)$	Out of the **45** turtles with the D form of the DE gene,**5** are Bayosian turtles and **40** are Freqosian turtles.

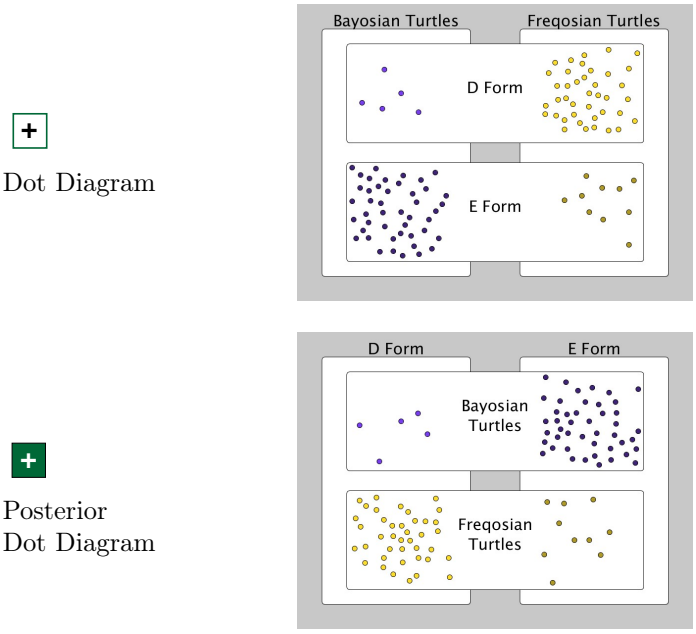
Note. $N(\cdot)$ refers to the natural frequency of an item. Likelihood formats are shown as empty boxes, whereas posterior formats are shown as filled in boxes. "+" denotes that a format includes the complement information.

3.2.2 Visual Formats

Additionally, six visual formats were created through a 3 (format type: icon array, bar graph, or dot diagram) by 2 (information type: likelihoods or posteriors) factorial design. An example of each format is provided in Table 3.2 below. Each type of format visualizes the corresponding natural frequency format (either the likelihood with complement or the posterior with complement variations). Icon arrays and bar graphs are two of the most represented visualizations explored in the Bayesian reasoning literature (Galesic et al., 2009; Gaissmaier et al., 2012; Brase, 2009; Ancker et al., 2006), using the number of icons or the length of a bar (respectively) to represent the number of data points with each set of features (i.e., each relationship between test outcome and class). We also introduce the dot diagram as a novel format inspired by Euler Diagrams (alternatively called "Venn diagrams"), which have been found to be successful at improving Bayesian reasoning (Sloman et al., 2003; Brase, 2009; Micallef et al., 2012). However, one main difference is that Euler diagrams present only the positive result of a binary query, whereas our dot diagram is designed to show both possible test outcomes equivalently in order to fairly present an information search task as opposed to a Bayesian inference task. The dot diagram uses Uniform Poisson Disk Sampling (Lagae & Dutré, 2008) to place the individual dots in an approximation of a uniform random distribution. Dots (each representing a single item) are distributed within containers that highlight the subset relationships between class and test outcomes. For the likelihood and posterior variants, the visual formats were grouped by class (likelihoods) or by test outcome (posterior) in the same way as the natural frequency formats. The colors used in all our visual formats were defined using a tool called Paletton (http://paletton.com), which allowed us to construct a symmetrically divergent and colorblind friendly palette.

Table 3.2: An example of the visual formats, expressed in terms of the Experiment 1 task environment.

Format	Example
+ Icon Array	
+ Posterior Icon Array	
+ Bar Graph	
+ Posterior Bar Graph	



Note. Likelihood formats are shown as empty boxes, whereas posterior formats are shown as filled in boxes. "+" denotes that a format includes the complement information.

3.3 Differences Between Presentation Formats

3.3.1 Design Features of Numerical Formats

One main difference that we expected to play a role in influencing search decisions was between those presenting natural frequencies (including visual formats) and those presenting absolute probabilities. This expectation is based on the ample literature finding an effect of natural frequencies on Bayesian inference (for a review, see Meder & Gigerenzer, 2014). Because natural frequencies are designed to approximate and summarize natural sampling, we hypothesized that certain natural frequency formats would also lead to similar performance as the experience-based learning results from Nelson and colleagues (2005), which were also based on natural sampling.

The differences between providing likelihood information or providing posterior information has no true analogue in the Bayesian reasoning literature, because the posterior information would already contain the answer to the task. However, we formed the weak hypothesis that the posterior variations might be better suited for an information search task, since all of the relevant OED models make use of the posterior probabilities in the calculation of usefulness. The likelihood difference heuristic is an exception however, and it could also be the case that heuristic strategies would not require the extraction of posterior information in order to arrive at the same search decisions as the OED models.

Lastly, Rusconi and McKenzie (2013) have shown that the inclusion of the complement information improves sensitivity to the informativeness of an answer, which we hypothesized would also be a potential mediator for improving search behavior.

3.3.2 Design Features of Visual Formats

Aside from the differences between the numerical formats, we also hypothesized that there would be individual differences between visual formats. Thus our visual presentation formats were specifically chosen to cover differences between three main design features[1]: *spatial extent*, *countability*, and *part-to-whole information*. Table 3.3 provides a comparison of each visual format type used in this study in terms of design features.

Spatial Extent. Empirical work by Cleveland and McGill (1985) has shown that graphical representations exploiting basic perceptual abilities, such as length comparisons, can engage automatic visual perceptual mechanism, thus reducing the required mental computation and leading to higher quantitative reasoning capabilities. This is one hypothesis we tested in our study, whereby icon arrays and bar graphs convey quantitative information using the spatial extent of an array of icons or the length of a bar. In contrast, the dots in the dot diagrams are restricted to a container of fixed size, and convey quantitative information without spatial extent. Ancker et al. (2006) found that randomly distributed stimuli similar to the dot diagram are not as effective for the purpose of judging proportions, although they may help convey the concept of chance.

Countability. Icon arrays are dissimilar to bar graphs because the quantitative information contained in an array of icons are countable, whereas the length of a bar is not. If the countability of visual formats has an effect on Bayesian reasoning or information search, we would expect to see noticeable differences between the icon array and bar graph formats. The dot diagrams are somewhere between icon arrays and bar graphs in terms of countability, because while the dots are countable in principle, the random placement makes it relatively more difficult to arrive at the exact number. Stone et al. (2003) proposed that the inability to derive the exact numerical quantities represented by a visual format accounted for a reduced ability to accurately perceive risk. In contrast, Micallef et al. (2012) found that Euler diagrams with randomly distributed dots (similar to our dot diagrams) solicited the

[1]There has also been work by Gaissmaier et al. (2012) that found no effect of iconicity (i.e., realism of representations) on accuracy in Bayesian inference tasks, and thus was not a factor that we considered in the choice of our stimulus.

highest Bayesian reasoning performance, compared to other variations that used ordered icon arrays and facilitated counting.

Part-to-whole Information. We specifically designed the dot diagrams to provide accessible information about the part-to-whole relationships between features of the environment, e.g., that the quantity of Bayosian turtles with the D form of the DE gene represents a portion of the total number of Bayosian turtles, as well as a portion of the total number of turtles with the D form. Part-to-whole information in the icon arrays and bar graphs are not easily accessible, because each combination of gene form and species is expressed as a separate entity. A review by Ancker et al. (2006) on the effectiveness of visual formats in conveying health-related risk proposed that part-to-whole information played a substantial role in improving risk perception.

Table 3.3: Design Features of Visual Formats

Design Feature	Format Type		
	Icon Arrays	Bar Graphs	Dot Diagrams
Spatial Extent	+	+	—
Countability	+	—	+/–
Part-to-whole Information	—	—	+

Note. '+' denotes that the format type embodies the design feature, whereas '—' denotes that it does not. The dot diagrams are described using '+/-' in the countability row, because while they are countable in theory, the random distribution of the dots makes this process more difficult than the ordered icon arrays.

Experiments 4

4.1 Search Environments

We conducted four different experiments following the same procedures, but each using a unique yet structurally analogous probabilistic environment. Different types of environments may lead to different types of behavior, and we avoided restricting our findings to a single environment. The four environments presented in this study represents our attempts at gaining a more generalizable understanding of search behavior (Table 4.1).

Experiment 1 tests a (theoretically) simple information search task, in which probability gain, information gain, impact, and KL-Divergence all agree about which query is most useful. It is thus a kind of 'baseline' experiment. Experiment 2 tests a potentially more difficult information search task, in which the higher probability gain query is suboptimal from the perspective of information gain, impact, and KL-Divergence. In Experiments 1 and 2, no query offers the possibility of a certain result. In Experiments 3 and 4, we also investigate whether the possibility of a certain results influences test-selection behavior. In Experiment 3, the optimal query (from the perspective of probability gain) is suboptimal from the perspective of information gain, impact, KL-Divergence, and the probability of certainty. Thus, Experiment 3 is an attempt to address whether (or the presentation formats for which) people can identify the highest-probability gain query in circumstances where all the other competing models prefer a different query. Experiment 4 addresses whether (or the circumstances under which) the possibility of a certain outcome may be a dominating factor in information search behavior, as opposed to a weaker contributing influence. In Experiments 2 and 3, the higher probability gain query is in opposition to all the other OED

models, in order to ensure that a subject's search preference is definitively aligned with or in opposition to the normative guidance of probability gain (including all possible heuristic explanations that approximate or exactly implement probability gain). In Experiment 4, the possibility of a certain outcome is set in opposition to all the competing models, to also allow for a definitive understanding of the effect that a certain outcome has on search behavior.

Table 4.1: Overview of search environments.

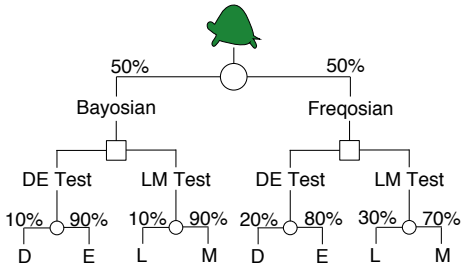
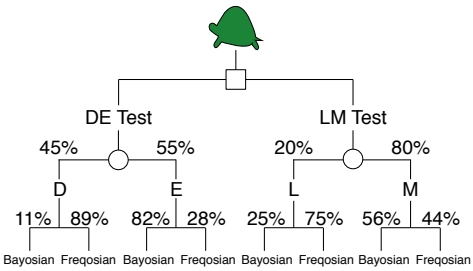
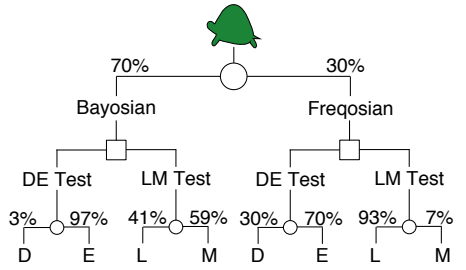
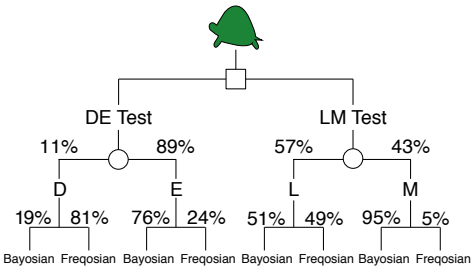
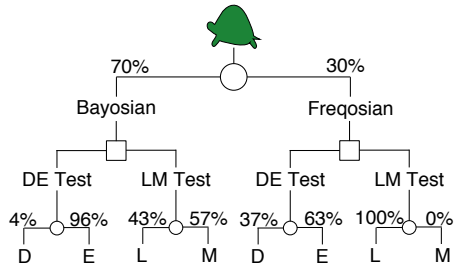
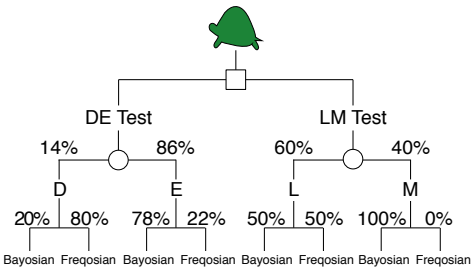
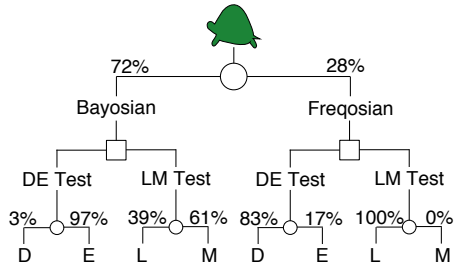
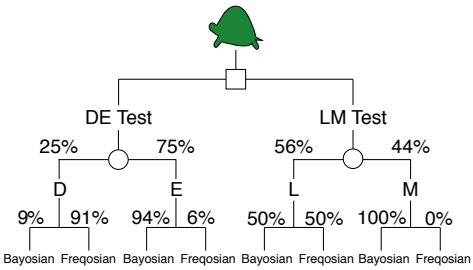
Experiment	Model(s) where the DE test is most useful	Model(s) where the LM test is most useful	Model(s) with no preference
1	Probability gain Information gain Impact KL-Divergence	—	Probability of Certainty
2	Probability gain	Information gain Impact KL-Divergence	Probability of Certainty
3	Probability gain	Information gain Impact KL-Divergence Probability of Certainty	—
4	Probability gain Information gain Impact KL-Divergence	Probability of Certainty	—

Note. "—" denotes that none of the relevant models fit the criterion. A full description of expected utilities of each search query and the utilities of each query outcome is available in Table A1 (Appendix A).

4.2 Environmental Probabilities

Table 4.2 (below) visualizes each environments using probability trees. The likelihood trees shown on the left describe the search environment first in terms of base rate, and then the likelihoods of each test outcome conditioned on the species. The posterior trees on the right describe the environment first in terms of test outcome, and then the posterior probability a turtle belong to the Bayosian or Freqosian species, conditioned on the test outcome.

Table 4.2: Search environments visualized as probability trees.

Experiment	Likelihood Tree	Posterior Tree
1		
2		
3		
4		

Note. The likelihood tree includes base rate information as well as the likelihoods for each species of turtle possessing a specific gene form, $P(q_j|c_i)$. The posterior tree shows the marginal probabilities, i.e., the probability of each test outcome unconditioned on the species, $P(q_j)$, and the posteriors, i.e., the probability that a turtle belongs to a species given a test outcome, $P(c_i|q_j)$. Boxes denote decision nodes and circles denote chance nodes.

4.3 Experiment 1

4.3.1 Environment

Experiment 1 was conducted in order to explore the limits of search behavior. Because all of the models of information search agree on which query is more useful, we expected the results of this experiment to provide information about the upper limits of performance, in terms of the proportion of participants who are able to identify the query maximizing classification accuracy. Additionally, Experiment 1 is used to describe the canonical experimental design replicated for all other experiments. Table 4.2 (top row) visualizes the task environment given to the participants, which was constructed such that all of the relevant OED models agree that the DE test has the higher expected usefulness. Table 4.3 shows the expected utilities for each OED model.

Table 4.3: Expected Utilities of OED Models: Experiment 1

OED Model	$eu(DE)$	$eu(LM)$	Preferred Query
Probability Gain	**0.35**	0.1	DE test
Information Gain, KL-Divergence	**0.40**	0.05	DE test
Impact	**0.35**	0.1	DE test
Probability of Certainty	0	0	—

Note. "—" denotes no preference between the two queries; The higher of the two expected utility values are shown in **bold**. The expected utilities for probability of certainty are zero, because there are no certain outcomes in this environment. The range of possible expected utilities for probability gain and impact are between 0 and 0.5, while the others range between 0 and 1.

4.3.2 Method

Participants and Design. This experiment, along with all the others presented in this thesis, were computer-based experiments carried out using *Amazon's Mechanical Turk platform* (AMT), which is an online marketplace for work that requires human intelligence. *Human Interaction Tasks* (HITs) are self-contained tasks posted on the AMT marketplace, which workers can complete in return for a monetary reward. In the first experiment, 703 HITs were published exclusively to workers who had completed at least 1000 HITs and had at least a 95% acceptance rate on previous tasks. After excluding participants that had failed to complete all of the required tasks, as well as those who self-identified as having only a limited or elementary understanding of English, 677 participants were included in the experiment (see Table 4.4 for demographic information). There was no overlap of participants between

experiments, because once a HIT was completed, the subject's AMT worker ID was added to the exclusion criteria for subsequent studies. To match the majority of previous studies on Bayesian reasoning using a flat fee (McDowell & Jacobs, 2015), each participant who completed the HIT was paid a flat of $1.50 USD. While the reward seems small compared to labratory studies, Paolacci et al. (2010) found that an hourly rate of $1.66 USD to be sufficient motivation for scientific studies run on the AMT platform. The entire experiment consisted of original software, programmed using Javascript, HTML, jQuery, PhP, and AJAX.

Participants were randomly assigned to one of 14 different information presentation formats, with the number of participants per format ranging from 47 to 50. Within each presentation format, participants were randomly assigned to one of 16 different randomizations of the probability values entailing the same informational utilities, in order to avoid confounds with aspects of the stimulus (e.g., the naming of the turtle species, the naming of the genes, and choice of colors in the visual formats). These randomizations are chosen from a class of equivalent environmental conditions, resulting from a shuffling of the prior probabilities for each class and for the conditional probabilities of the form of each gene[1]. For the analyses, the 16 randomizations were recoded to match a single canonical variation, described in Table 4.2 above.

We had three dependent variables. The first was the information search decision, where a subject chose either the DE test of LM test, one of which solely aligns with the higher probability gain query[2]. The second variable consisted of the results of the probability estimation task, with which we used *Mean Absolute Error* (MAE) to measure judgment accuracy. The third variable was the result of a numeracy test, measured with scores ranging from 1 - 7. We hypothesized that presentation format and numeracy score would be correlated with probability judgment error, which in turn would be a predictor for the proportion of probability gain queries.

4.3.3 Procedure and Materials

The experiment was designed to take between 15-20 minutes. After participants gave their consent to participate in the study, they read the *Turtle Island* cover story (see Appendix C for screenshots from the experiment). Before moving on, participants were asked a

[1]For example, swapping the query outcomes for the DE and LM test change which query is most useful, respective to each of the OED models. We then recoded the test selection and probability judgments relative to the assigned randomization.

[2]The higher probability gain query is always identified as the DE test in the canonical versions of the search environments described in this thesis; however, participants did not necessarily perceive it as such, due to the use of 16 assigned randomizations of the environmental probabilities.

Table 4.4: Demographics of AMT Participants

Experiment	Gender	Age	Education
1 (n=677)	Female: 50% Male: 50%	Median: 32 Mean: 35 Range: [18-71]	High School: 12% Some University: 39% Bachelor's Degree or Higher: 49% Other: 1%
2 (n=817)	Female: 46% Male: 54%	Median: 32 Mean: 34 Range: [18-76]	High School: 14% Some University: 34% Bachelor's Degree or Higher: 51% Other: 0.5%
3 (n=683)	Female: 45% Male: 55%	Median: 31 Mean: 34 Range: [18-83]	High School: 15% Some University: 33% Bachelor's Degree or Higher: 52% Other: 0.4%
4 (n=681)	Female: 48% Male: 52%	Median: 31 Mean: 34 Range: [18-73]	High School: 14% Some University: 37% Bachelor's Degree or Higher: 48% Other: 1%

Note. Gender, age, and education were self reported.

simple question to ensure that they had understood the nature of the task (following the recommendations of Bramley et al., 2015). Specifically, they were asked one of four randomly selected questions about the binary nature of the DNA test, i.e., "If a turtle does **not** have the D form of the DE gene, then which form of the gene does it have?". Responses were selected from one of the following options: "D form", "E form", "L form", or "M form". participants were required to correctly answer this question before continuing the study. If an incorrect answer was given, participants were asked to re-read the instructions and attempt a new randomly selected question.

Presentation Formats. After being given background on the Turtle Island scenario, participants received information on the statistical structure of the task environment in one of the 14 different presentation formats, which provided information about the entire population of 100 turtles. A full description of each format is provided in Appendix B (see Table B1 for Experiment 1).

Search Task. After being presented with information on the environmental probabilities, participants were asked to choose the DNA test that they believed would yield the highest chance of correctly classifying the species of a turtle picked at random from the island. Specifically, they were asked: "Which test is better for having the highest chance of correctly

classifying the animal as either a Freqosian or a Bayosian turtle?" This question corresponds to an accuracy goal, for which probability gain is the appropriate model of informational utility. In the probabilistic environment described above, the DE test has a higher probability gain than the LM test (Table 4.3). In Experiment 1 the DE test also entails a higher expected utility relative to information gain, impact, and KL-Divergence; however, we test other relationships in subsequent experiments.

Probability Judgments. After the search task, participants were asked to give their estimates for the 11 different probabilities of the search environment (one base rate, two marginals, four likelihoods, and four posteriors). To avoid potential memory confounds, the assigned presentation format describing the search environment was once again provided while answering the probability judgment questions. The questions were arranged into blocks, with three questions referring to the prior probability species, i.e., $P(class)$, and the marginal probability of the test outcomes, ie., $P(outcome)$; four questions concerning the likelihood of a test outcome given the species, i.e., $P(outcome|class)$; and four judgments regarding the posterior probabilities of the species given the test outcome, i.e., $P(class|outcome)$. The order of the three blocks were randomized across participants (see Table 4.5 for example questions). Responses were given using a visual analog scale slider. As the slider was adjusted, text displayed the two complementary values of the slider (e.g., "Bayosian turtles: 70%" and "Freqosian turtles: 30%") corresponding to its position (see Appendix D for the full set of probability judgment questions and screenshots of the input method). Thus, participants could provide answers with the help of implicit representations using the visual analog scale, or explicit numeric representations using the displayed text, or both.

Table 4.5: Probability Judgment Questions. See Appendix C for a full description

Question	Example
Prior $P(class)$	What are the chances that a turtle picked at random from the island is a Bayosian or Freqosian turtle, respectively?
Marginal $P(outcome)$	What are the chances that a turtle picked at random from the island has the D or E form, respectively, of the DE gene?
Likelihood $P(outcome\|class)$	What are the chances that a Bayosian turtle has the D or E form, respectively, of the DE gene?
Posterior $P(class\|outcome)$	What are the chances that a turtle picked at random, which has the D form of the DE gene, is a Bayosian or Freqosian turtle, respectively?

Numeracy Test. Subsequently, participants were asked to complete a numeracy test. The goal of this test was to assess participants' abilities to comprehend and reason with numerical information, which has been shown to correlate positively with accuracy in statistical reasoning tasks (Schwartz et al., 1997; Bodemer et al., 2014; Brown et al., 2011). Our initial hypothesis was that numeracy may mediate an individual's ability to perform Bayesian reasoning, with the consequence of indirectly improving the capacity to evaluate the usefulness of a question. We used a hybrid test consisting of the Schwartz et al. (1997) numeracy test and the adaptive *Berlin Numeracy Test* (BNT) (Cokely et al., 2012). The choice to combine the two numeracy tests follows the recommendation of Cokely et al. (2012), who found that this method resulted in the best discriminability specific to the broad range of education backgrounds within an AMT sample population. The BNT test is well suited for discriminating among highly educated populations, whereas the Schwart et al. test is better for discriminating relative numeracy among relatively less educated individuals. The three Schwartz et al. questions were presented first, in random order, yielding a scoring range between zero and three. The subsequent adaptive BNT is comprised of between two to four questions (depending on the responses), and yields a score between one and four. The scores for the two tests were added together, yielding a combined numeracy score in the range of 1 to 7. See Appendix D: Table E1 for a full description.

Demographic Information. Lastly, participants provided demographic information, which is summarized in Table 4.4.

4.3.4 Results: Experiment 1

Information Search Decisions. The search environment in Experiment 1 was specifically chosen to test the upper-limits of performance on an information search task, relative to the methods and subject population, since all relevant OED models agree on which query is most useful. Thus, it was expected that the majority of participants would pick the more useful query (DE test), regardless of format. Aggregated across all formats, 80% of participants correctly chose the test with the higher probability gain, with a range of 71% - 88% for individual formats. We can view this performance as a reasonable upper-bound to the proportion of probability gain queries we can expect from each of the subsequent experiments presented in this paper. Addressing our primary research question about the effect of format on search decisions, a one-way between participants ANOVA was conducted to compare the effect of presentation formats on search decisions. We found no significant differences between any formats, in terms of the proportion of participants who choose the

more useful search query [$F(13,664)$=.99, p=.46]. Fig 4.1 (A) shows the complete results of the information search task, with the height of the bars representing the proportion of probability gain queries for each format.

Formats and Probability Judgments. To test whether presentation formats influenced Bayesian reasoning, we conducted a one-way between participants ANOVA comparing the effect of presentation format on the *Mean Absolute Error* (MAE) of the probability judgment task[3]. We found that no single format performed significantly better or worse than any other format when aggregated over the 11 questions that made up the probability judgment task [$F(13,664)$=1.18, p=.29]. Fig 4.1 (B) shows the distribution of MAE split by presentation format. We did however find differences between formats on specific subsets of probability questions. Likelihood formats performed significantly better than the posterior formats on the likelihoods questions, $t(675)$=-6.22, p<.001, as well as on the base rate question, $t(675)$=-5.05, p<.001, whereas the posterior formats performed better at the posterior questions $t(675)$=3.70, p<.001. This effect is even more pronounced when looking only at the subset of probability formats, where participants were explicitly given the correct answers to certain questions (e.g., with congruence between the presentation format and the answer format) as opposed to the natural frequency and visual formats, where information is expressed in joint probabilities and must be re-normalized into absolute probabilities. These results confirmed that presentation formats have an effect on specific probability judgments; however, no single format had a distinct advantage when aggregated across all 11 probabilities that make up a two-query search environment. This gives a very different perspective than if we had focused solely on a single posterior probability judgment, which is typical in the Bayesian reasoning literature (for a review see McDowell & Jacobs, 2015).

Probability Judgments and Search Decisions. To test the hypothesis that a better understanding of the environmental probabilities would facilitate the identification of the higher probability gain query, we conducted an analysis on the relationship between these two factors. We found a moderate negative correlation (r_{pb}=-.16) between probability judgment error and search decisions, indicating that lower error was correlated with a higher proportion of probability gain queries (Fig 4.1 C left; participants split into quartiles based on judgment error, where the height of the bars show the proportion of probability gain queries). Additionally, we found that participants who chose the DE test had a lower group mean error on the probability judgment task than those who chose the alternative,

[3]Mean Squared Error yielded equivalent results in all cases, although MAE provides a better representation of the magnitude of error.

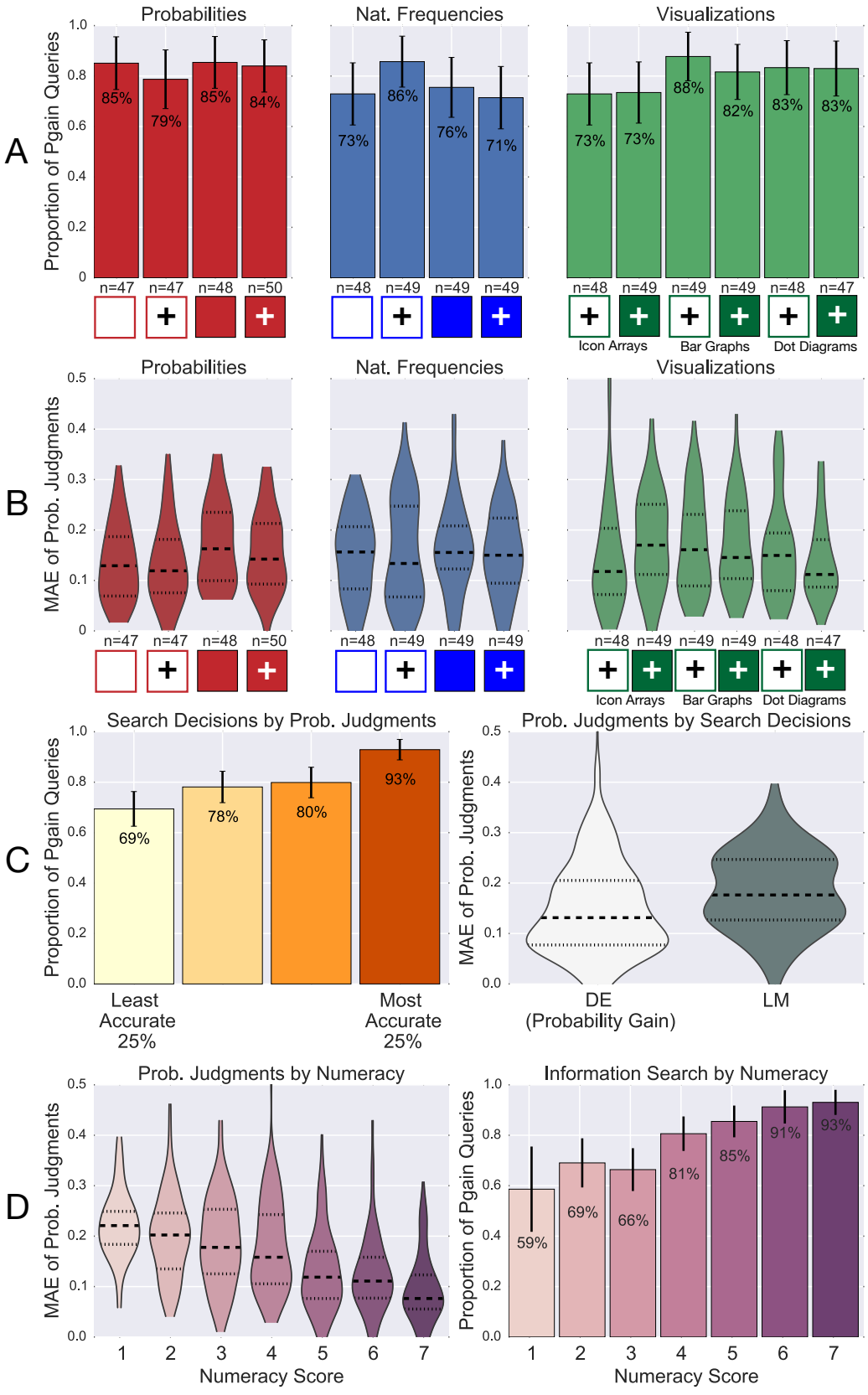


Figure 4.1: Results from Experiment 1: Empty boxes denote likelihood formats, whereas posterior formats are shown as filled in boxes. "+" denotes that a format includes the complement information. (A) the proportion of probability gain queries for each format, with Agresti-Coull 95% confidence intervals; (B) distribution of MAE on the probability judgment task for each format; (C left) search decisions with participants split into quartiles based on probability judgment performance, in increasing order of performance; (C right) violin plots of MAE with participants split by search decision; (D left); violin plot of MAE split by numeracy score; (D right) search decisions split by numeracy score.

$t(680)$=4.3, p<.001 (Fig 4.1 C right; distribution of error, split by search decision). Thus, for Experiment 1 we found evidence suggesting that lower probability judgment error was correlated with a higher proportion of queries maximizing classification accuracy.

Numeracy. Lastly, we investigated the role of numeracy as it relates to probability judgments and information search. The results of the numeracy task confirmed our initial hypothesis that numeracy facilitates accurate probability judgments, showing that higher numeracy skill is correlated with lower error on the probability judgment task (r=-.42; Fig 4.1 D left). We also found that higher numeracy scores were a predictor for a larger proportion of probability gain search decisions (r_{pb}=.25; Fig 4.1 D right). Thus, Experiment 1 shows that higher numeracy skill may facilitate both probabilistic reasoning as well as information search decisions.

Summary. Experiment 1 allowed us to identify the reasonable upper-limit of performance on an information search task (overall 80% accuracy-maximizing queries; range of 71% to 88% across formats), specific to the experimental methods and the AMT subject pool. In this search environment—where all of the relevant OED models agree on which query is more useful—we found no effect of presentation format on search decisions, whereas performance on the probability judgment task and numeracy skill were good predictors for the proportion of correctly identified probability gain queries. Having established a performance baseline on a search environment with unanimous agreement amongst OED models, let us turn to the subsequent experiments where we introduce disagreement.

4.4 Experiment 2

4.4.1 Environment

While Experiment 1 shows that the majority of participants were able to identify the more informative query in an environment where there was no conflict between different OED models, Experiment 2 was conducted using a divergent search environment (see Table 4.2 for the probability trees). Specifically, we used computer simulations to search the parameter space over to find an environment where probability gain maximally disagreed with information gain, with the constraint that no queries result in a certain outcome[4].

[4] A two feature, binary class search environment can be fully described using a single prior and four likelihood probabilities. We randomly generated 1 million random variations of these five probabilities and calculated the expected utilities for each query relative to probability gain and information gain. The environment used in Experiment 2 had the largest pair-wise disagreement strength between probability gain and information gain, under the constraint that all posterior probabilities belonged in the range of 0.05 and 0.95. See Nelson et al. (2010) for a full description of this process.

The DE test has higher probability gain, while the LM test has higher information gain, KL-Divergence, and impact (Table 4.6). As was the case with all experiments, participants were explicitly asked to choose the query that would result in the highest classification accuracy, corresponding to only the higher probability gain strategy.

Table 4.6: Expected Utilities of OED Models: Experiment 2

OED Model	$eu(DE)$	$eu(LM)$	Preferred Query
Probability Gain	**0.07**	0	DE test
Information Gain, KL-Divergence	0.10	**0.19**	LM test
Impact	0.11	**0.22**	LM test
Probability of Certainty	0	0	—

Note. "—" denotes no preference between the two queries; The higher of the two expected utility values is shown in **bold**. The expected utilities for probability of certainty are zero, because there are no certain outcomes in this environment. The range of possible expected utilities for probability gain and impact are between 0 and 0.5, while the others range between 0 and 1.

4.4.2 Method

Participants and Design. The procedure was identical to the previous experiment. 942 HITs were published and submitted, and after excluding those who had failed to complete all the required tasks as well as those who had limited or elementary understanding of English, 817 participants were included in the experiment. There was no overlap of participants between experiments. Table 4.4 shows the demographics of the subject population. The number of participants randomly assigned to each format ranged from 42 to 65.

4.4.3 Results: Experiment 2

Information Search Decisions. In contrast to Experiment 1 in which all OED models agreed on the more useful query, in Experiment 2 we found that search decisions were strongly influenced by presentation format. Selection of the higher probability gain query varied between 27% and 86% across the different formats (Fig 4.2 A). The standard probability format (likelihood variation without complement) resulted in the lowest ability to identify the probability gain test, with only 27% of participants making this choice. This proportion was virtually identical to the standard probability format results from Nelson et al. (2010, Experiment 1, Conditions 1 and 2), which used the same presentation format and found that 27% and 30% of participants, respectively, could identify the query with higher probability gain. On the other hand, the posterior icon array (82% probability gain) and posterior bar

graph (86% probability gain) formats met and even exceeded the levels of experience-based learning from Nelson et al. (2010, Experiment 1: 80% in Condition 1 and 82% in Condition 2).

Figure 4.2 A shows that there was both high variability between subsets of formats as well as within subsets (e.g., probability formats, natural frequency formats, and visualizations). For instance, within the set of numeric probability formats, the proportion of choices for the probability gain test ranged from 27% to 77%. In two of the four probability formats, the information search choices differed from chance: participants in the standard probability condition reliably failed to identify the higher probability gain test ($p<.001$; binomial test[5]), whereas participants in the posterior with complement condition reliably succeeded in identifying the higher probability gain test ($p<.001$; binomial test). Search decisions in the other two conditions did not differ from chance.

The natural frequency formats did not vary as much as the probability and visual formats, but still ranged from 51% to 67%. Only in the posterior with complement condition did search decisions differ from chance ($p=.01$; binomial test). This indicates that although the natural frequency formats tend to be better than the standard probability format, they are not particularly helpful for evaluating the usefulness of tests in certain search environments.

Within the visual formats, the proportion of probability gain decisions ranged from 43% to 86%. In four out of the six conditions, participants were able to reliably identify the accuracy-maximizing query at greater than chance levels. The posterior icon array was most effective (86%, $p<.001$; binomial test), followed by the posterior bar graph (82%, $p<.001$), the likelihood bar graph (72%, $p<.001$), and the likelihood icon array (70%, $p=.007$). The two types of dot diagrams did not seem to be particularly helpful for identifying the more useful test, with search decisions at chance level for both conditions.

Summary of Presentation Formats and Search Decisions. Information formats seem to matter a great deal for soliciting useful information search decisions when there is disagreement between OED models. We found strong differences across as well as within different types of probability, natural frequency, and visual formats. Our findings suggest that visual formats such as icon arrays and bar graphs provide effective means to help people identify the higher probability gain test, particularly in the posterior variation (grouped by outcome rather than by class). When posterior icon arrays and posterior bar graphs were used, the proportion of participants correctly identifying the higher probability gain query was similar to the experienced-based results from (Nelson et al., 2010). The key advantage

[5]A binomial test is used to test for significant differences from chance for a binary variable, such as a search decision between two possible queries.

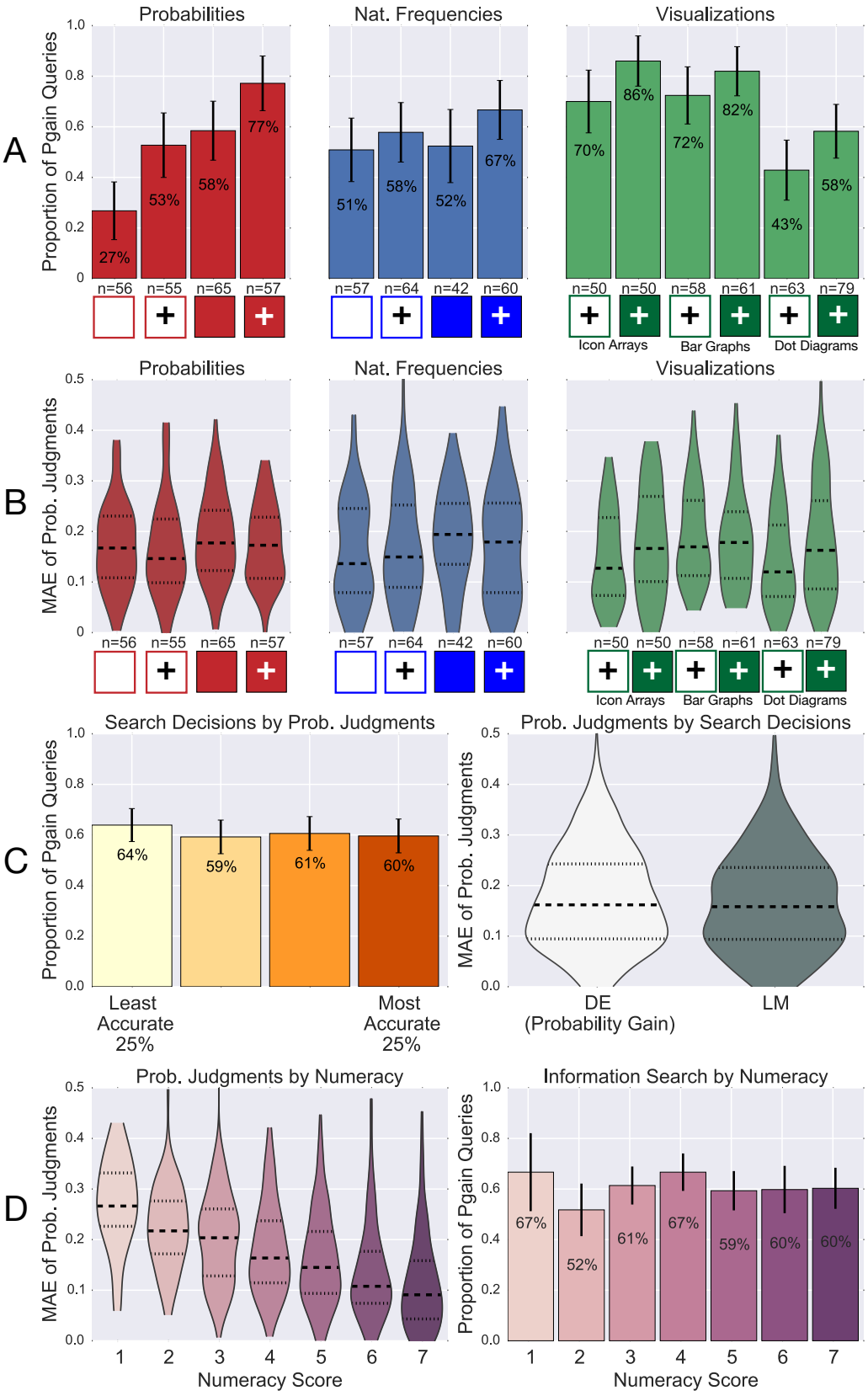


Figure 4.2: Results from Experiment 2: (A) proportion of probability gain queries (B) distribution of MAE; (C left) search decisions with participants split into quartiles based on probability judgment performance, in increasing order of performance; (C right) MAE split by search decision; (D left); MAE split by numeracy score; (D right) search decisions split by numeracy score.

of these graphical formats is that they do not require several hours and hundreds of trials of learning, as opposed to experience-based learning, while at the same time eliciting the same (correct) intuitions about the relative usefulness of a query.

Several of the numerical formats were also found to be relatively effective. The results suggest that people can utilize numeric information conveyed in the form of probabilities if the complement is added and if posterior probabilities are given, rather than in the form of priors and likelihoods associated with the standard probability format. Consistent with previous research on Bayesian reasoning (Gigerenzer & Hoffrage, 1995) and information search (Nelson et al., 2010), the standard probability format led to the lowest proportion of correctly identified search decisions, with a performance that was significantly worse than chance.

Formats and Probability Judgments. Is there a similar relationship between information format and accuracy on the probability judgment task? Results for Experiment 2 replicated the finding from the Experiment 1 in that no single format performed systematically better or worse on the probability judgment task, when aggregated across all 11 probability judgment questions (Fig 4.2 B) [$F(13,804)$ =1.15, p=.31]. Experiment 2 results replicated the sanity check results from Experiment 1, finding that likelihood formats performed better than posterior formats on the likelihood questions, t(815)=-9.33, p<.001, and base rate questions, $t(815)$=-5.73, p<.001, whereas the posterior formats performed better on the posterior questions, $t(815)$=7.66, p<.001.

Probability Judgments and Search Decisions. In contrast to Experiment 1, there was virtually no correlation (r_{pb}=.04) between probability judgment error and the proportion of probability gain decisions (Fig 4.2 C left). We also found a similar lack of correlation between judgment error and search decisions when broken down by question type (i.e, base rate, marginals, likelihoods, and posteriors). This suggests that the ability to make accurate probability judgments is not a sufficient criterion for improving search decisions. In Experiment 2, there were also no significant differences in the group means of probability judgment error between participants who chose the higher probability gain query compared with those who did not, $t(815)$=-1.2, p=.24 (Fig 4.2 C right).

Numeracy. In Experiment 2, numeracy was again a consistent predictor for MAE on the probability judgment task (r=-.4; Fig 4.2 D left), replicating the results from Experiment 1, showing participants with higher numeracy skill consistently had lower error in judging the environmental probabilities. However, in Experiment 2 there was no correlation between

numeracy and the proportion of probability gain queries (r_{pb}=.003; Fig 4.2 D right), suggesting that numeracy facilitates probability judgment accuracy, but not necessarily information search.

Summary. Experiment 2 used a search environment where the optimal query relative to probability gain was set in opposition to the optimal query for information gain, impact, and KL-Divergence. In both experiments we found that probability judgment error aggregated over the entire set of 11 questions was not influenced by format, but solely by numeracy skill. On the other hand, in contrast to the results of Experiment 1, presentation format had a strong influence on search behavior in Experiment 2. Also in contrast to Experiment 1, we found no indication that either numeracy or probability judgment ability were helpful for identifying the higher probability gain query in Experiment 2.

What do the divergent results from Experiment 1 and Experiment 2 imply? One possibility is that when faced with a more ambiguous search environment——one where there is disagreement among the OED models as to which search query is more useful——the effect of numeracy and the ability to accurately estimate the underlying probabilities fails to be a predictor for search decision. Instead, only format has an effect on search decisions. Moreover and in contrast to our initial hypothesis, it was not necessary for participants be able to accurately provide *explicit* probability estimates in order to make good search decisions. Participants who were presented with the posterior icon array or the posterior bar graph were able to reliably identify the more useful query, even though they were not any better at the probability estimation task than those assigned to other presentation formats. This result raises the concern that existing literature on Bayesian reasoning is too focused on the ability to make explicit calculations, whereas decision making tasks such as information search may be independent of this ability.

At this point, we have found that presentation formats make a great deal of difference in search decisions, while at the same time finding that the ability to make accurate probability judgments is not necessarily correlated with search behavior. To further explore the relationship between presentation format and search behavior as a function of the probabilistic environment, we conducted two additional experiments using new environments. In Experiment 3, we "stacked the deck" against probability gain by including the possibility of a certain test result to the search query in opposition to the higher probability gain query. In Experiment 4, we investigated how much the probability of certainty, by itself, is capable of influencing search behavior.

4.5 Experiment 3

4.5.1 Environment

In this experiment we introduced the possibility of a certain outcome, where out of the two possible search queries, there exists a query outcome that provides certainty about the species of a turtle. Experiment 3 used a search environment that was almost identical to Experiment 2, but adjusted slightly so that one of the outcomes of the LM test would result in certainty. This set the higher probability gain test in opposition to the query with the higher usefulness for the probability of certainty model, along with all other OED models (Table 4.7). Table 4.2 shows this search environment visualized as probability trees, where the M result of the LM test gives 100% certainty that the turtle belongs to the Bayosian species. However, the M test result is only present in 40% of the turtles, whereas the alternative outcome (L result) gives maximal uncertainty about the species (50-50 outcome) with a 60% likelihood. Thus, the LM test is a risky choice and results in a lower expected classification accuracy than the DE test (70% and 78% respectively, corresponding to a probability gain of 0.00 and 0.08). The DE test remains the more useful query according to the probability gain model, while the LM test has the higher information gain, impact, KL-Divergence, and probability of certainty (Table 4.7).

Table 4.7: Expected Utilities of OED Models: Experiment 3

OED Model	$eu(DE)$	$eu(LM)$	Preferred Query
Probability Gain	**0.08**	0	DE test
Information Gain, KL-Divergence	0.13	**0.28**	LM test
Impact	0.14	**0.24**	LM test
Probability of Certainty	0	**0.4**	LM test

Note. The higher of the two expected utility values are shown in **bold**. The range of possible expected utilities for probability gain and impact are between 0 and 0.5, while the others range between 0 and 1.

4.5.2 Method

Participants and Design. The procedure was identical to the previous experiments. 706 HITs were published and submitted, of which 683 were included in the experiment after excluding those who had failed to complete all of the required tasks as well as those identifying as having limited or elementary English skills. Table 4.4 shows the demographics of the subject population. The number of participants randomly assigned to each format ranged from 48 to 52.

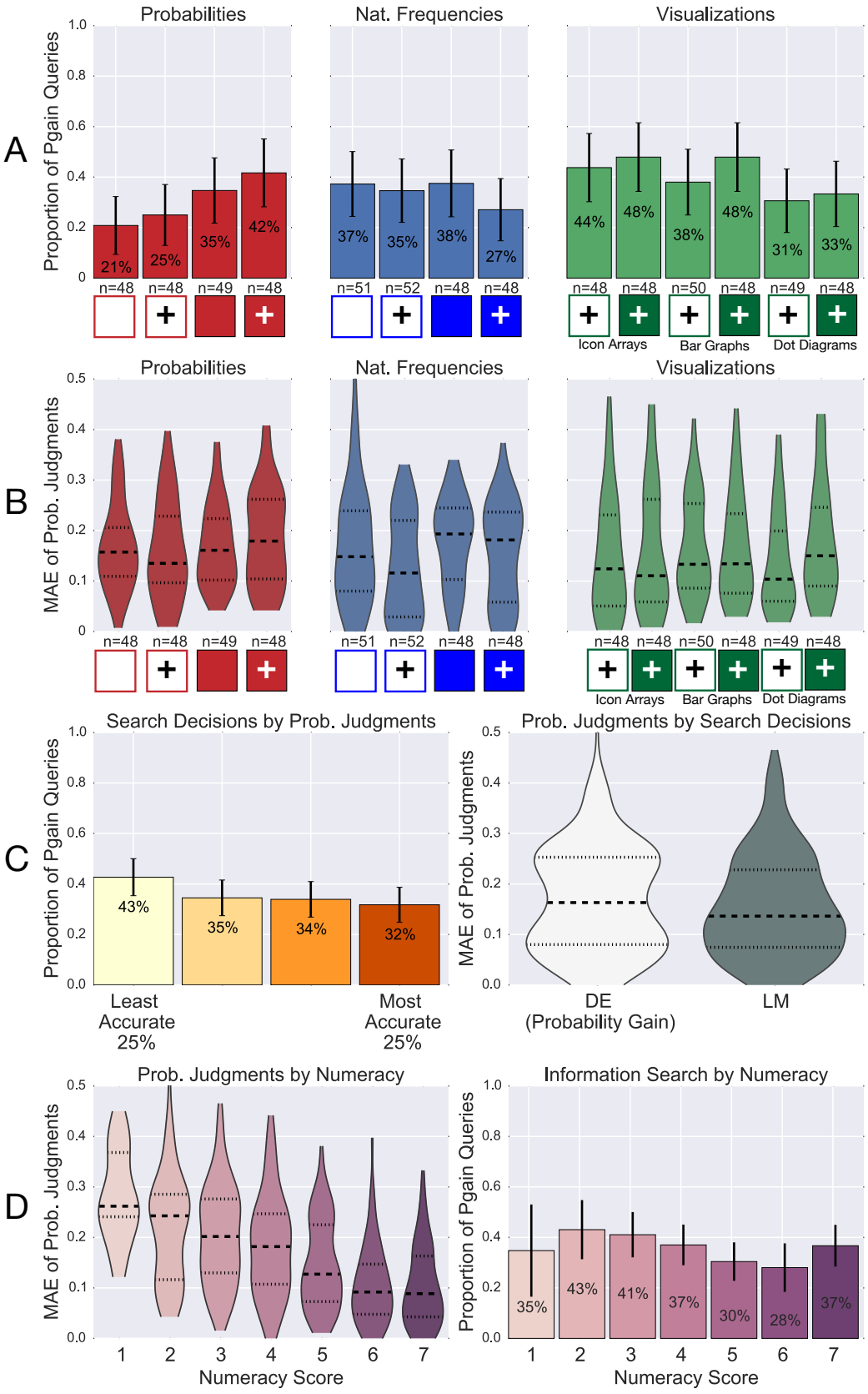


Figure 4.3: Results from Experiment 3: (A) proportion of probability gain queries (B) distribution of MAE; (C left) search decisions with participants split into quartiles based on probability judgment performance, in increasing order of performance; (C right) MAE split by search decision; (D left); MAE split by numeracy score; (D right) search decisions split by numeracy score.

4.5.3 Results: Experiment 3

Information Search Decisions Fig 4.3 (A) shows the effect of format on the proportion of search decisions that corresponded to the higher probability gain query. The relative rankings of the formats are highly correlated with Experiment 2 (r_s=.76), but all formats in Experiment 3 had a much lower proportion of probability gain decisions compared to the previous two experiments. Overall, only 36% participants chose the higher probability gain query (compared to 80% in Experiment 1 and 61% in Experiment 2). As was the case in Experiment 2, the posterior icon array and posterior bar graph solicited the highest proportion of probability gain queries (48% for both); however this lower proportion was not statistically different from chance. In fact, none of the formats resulted in participants showing a reliable preference for the probability gain query that was above chance. The only groups of participants who made search decisions that were statistically different from chance were those who had a strong preference for the lower probability gain test.

Formats and Probability Judgments. Consistent with all previous experiments, we found that format did not have a strong effect on the probability estimate task when aggregated across all 11 questions (Fig 4.3 B) [$F(13,670)$=1.26, p=.24]. Moreover, we also replicated the sanity check results from the previous experiments, showing that the likelihood formats performed better than posterior formats on the likelihood questions, $t(681)$=-7.73, p<.001, and base rate questions, $t(681)$=-6.02, p<.001, whereas the posterior formats performed better on the posterior questions, $t(681)$=4.46, p<.001.

Probability Estimates and Search Decisions. Similar to Experiment 2, but in contradiction to Experiment 1, we found no relationship (r_{pb}=.06) between judgment error and search decisions (Fig 4.3 C left). We also found a similar lack of correlation between judgment error and search decisions when breaking up the probability judgment task into different subsets of questions (i.e., base rate, marginals, likelihoods, and posteriors). There was also no significant difference in the distribution of error between participants who choose the higher probability gain query compared with those who did not, $t(681)$=-1.6, p=0.1 (Fig 4.3 C right). These results resemble the findings from Experiment 2, which also used a divergent environment, but are different from the results in Experiment 1, where all models agreed on which query was more useful.

Numeracy. Numeracy was correlated with performance on the probability estimation task (Fig 4.3 D left, r=-.45), replicating all previous experiments. However, as was the

case in Experiment 2, we found no relationship between numeracy and the proportion of probability gain search decisions (r_{pb}=-.06; Fig 4.3 D right).

Summary. When the possibility of certainty is introduced in opposition to the higher probability gain query, we see a relationship between format and search decisions that looks on the surface quite similar to Experiment 2, but with the proportion of probability gain choices reduced substantially across the board. None of the formats led to greater-than-chance rates of selecting the higher probability gain query. This suggests that the possibility of a certain outcome makes a substantial difference in search behavior, across all types of formats. In order to examine the effect of certainty by itself, Experiment 4 examines search behavior in an environment with the probability of certainty set in opposition to all the other OED models.

4.6 Experiment 4

4.6.1 Environment

We conducted one final study to isolate the effect of the probability of certainty on search behavior. Table 4.2 the task environment as probability trees, where probability gain, information gain, impact, and KL-Divergence consider the DE test to be the more useful query, but the LM test has a 44% chance of yielding an M result, which carries with it the implication that the turtle is certainly Bayosian. As was the case in Experiment 3, the LM test is a risky choice, because the L outcome results in maximal uncertainty (50-50) about the species of the turtle, and occurs with a 56% probability. Thus, the LM test has a probability gain of 0, and is also seen as less useful by the other OED models, except probability of certainty (Table 4.8).

Table 4.8: Expected Utilities of OED Models: Experiment 4

OED Model	$eu(DE)$	$eu(LM)$	Preferred Query
Probability Gain	**0.21**	0	DE test
Information Gain, KL-Divergence	**0.49**	0.29	DE test
Impact	**0.32**	0.25	DE test
Probability of Certainty	0	**.44**	LM test

Note. The higher of the two expected utility values are shown in **bold**. The range of possible expected utilities for probability gain and impact are between 0 and 0.5, while the others range between 0 and 1.

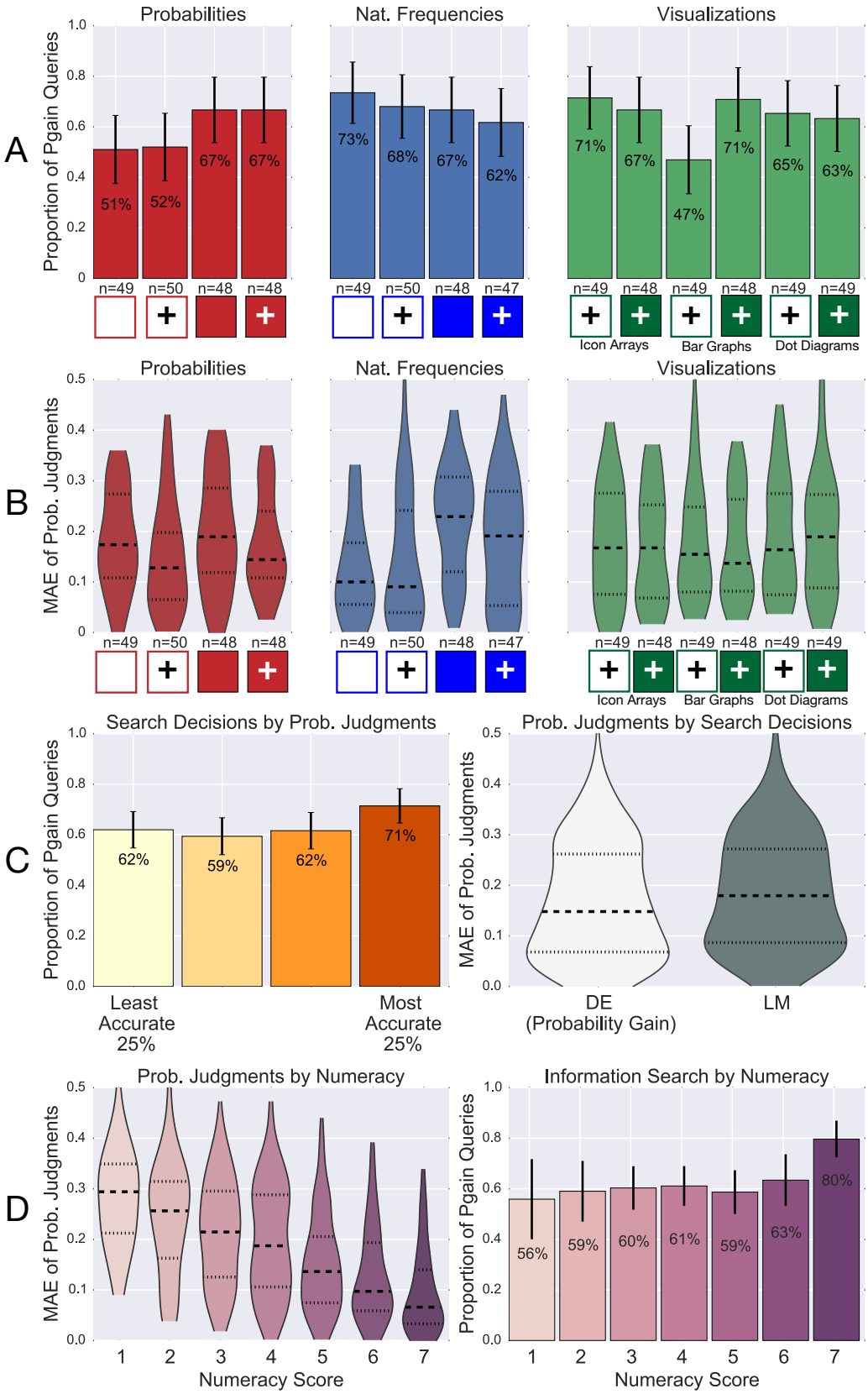


FIGURE 4.4: Results from Experiment 4: (A) proportion of probability gain queries (B) distribution of MAE; (C left) search decisions with participants split into quartiles based on probability judgment performance, in increasing order of performance; (C right) MAE split by search decision; (D left); MAE split by numeracy score; (D right) search decisions split by numeracy score.

4.6.2 Method

Participants and Design. The procedure was identical to the previous experiments. 708 HITs were published and submitted, of which 681 were included, after excluding those who failed to complete all required tasks or self-identified as having poor English skills (elementary or limited ability). Table 4.4 shows the demographics of the subject population. The number of participants randomly assigned to each format ranged from 47 to 50.

4.6.3 Results: Experiment 4

Information Search Decisions. The results of the information search task most closely resembled Experiment 1, which used a similar environment in which probability gain, information gain, impact, and KL-Divergence agreed on the more useful query. Like in Experiment 1, none of the formats significantly differed from each other (Fig 4.4 A) [$F(13,668)=1.386$, $p=.171$]. The formats that strongly differed from chance with a preference for the probability gain query were the natural frequency (likelihood variation), likelihood icon array, and posterior bar graph formats ($p<0.01$, Binomial Test). Overall, 64% of participants chose the higher probability gain query, with every single individual format having a lower proportion of probability gain queries than in Experiment 1.

Formats and Probability Judgments. For the first time, we found that format had an effect on the probability estimate task when aggregated across all 11 questions (Fig 4.4 B) [$F(13,668)=2.32$, $p=.005$]. A post hoc Tukey test showed that the likelihood variation of the natural frequency formats (both with and without complement) had significantly lower judgment error than the posterior frequency format (without complement) at $p<.05$; none of the other formats significantly differed from each other. We replicated the sanity check results from the previous experiments, showing that the likelihood formats performed better than posterior formats on the likelihood questions, $t(679)=-7.18$, $p<.001$, and base rate questions, $t(679)=-8.0$, $p<.001$, whereas the posterior formats performed better on the posterior questions, $t(679)=3.1$, $p=.002$. These differences between subsets of questions have proven to be consistent across all four experiments.

Probability Judgments and Search Decision. No significant correlation was found between judgment error and search decisions (Fig 4.4 C left, $r_{pb}=-.07$), with a similar lack of correlation for the different subsets of questions. There was also no significant difference in the distribution of error between participants who choose the higher probability gain query compared with those who did not, $t(679)=1.8$ $p=.07$ (Fig 4.4 C right). This

reinforces the notion that our hypothesized relationship between probability judgment error and information search is false, or at least is not consistently found across possible search environments.

Numeracy. As was the case with all previous experiments, numeracy was correlated with performance on the probability estimation task (r=-.47; Fig 4.4 D left). Thus, we can conclude that the correlation between numeracy and accuracy in making explicit probability judgments to be a robust finding, which is independent of the task environment. On the other hand, the relationship between numeracy and search decisions has proven to be heavily dependent on the task environment. Experiment 1 found a high correlation, while Experiments 2 and 3 found no correlation. For Experiment 4, the results of are somewhere in the middle, with a small correlation between numeracy and search decisions (Fig 4.4 D right, r_{pb}=.12). This finding suggests that numeracy can be a good predictor for search decisions when the more useful query is relatively obvious, and the various OED models are in agreement. However, this correlation disappears in more ambiguous environments, where there is a disagreement between models regarding which query is most useful.

Summary. There was no effect of format on search decisions in Experiment 4, with the results most closely resembling Experiment 1. Because the overall proportion of probability gain queries was lower (64% in Experiment 4 compared to 80% in Experiment 1), it appears that the possibility of a certain result had an effect on search preference across all formats. We also found an effect of format on probability judgments for the first time, where the natural frequency format, both with and without complement, performed significantly better than the posterior frequency format. Yet being more accurate at estimating probabilities had no bearing on the ability to identify the higher probability gain query: in this experiment we found no correlation between error on the probability judgment task and search decisions. Thus, while the natural frequency format has been shown to be better suited to deriving posterior probabilities than the standard probability format (Gigerenzer & Hoffrage, 1995; Cosmides & Tooby, 1996; Meder & Gigerenzer, 2014), and is also shown in this experiment to be more accurate at the entire aggregate of 11 probabilities than its posterior counterpart—none of these factors matter when it comes to selecting the more useful query. What did matter however, was that the possibility of a certain outcome eliminated the role that probability judgments play in explaining search behavior, when compared to Experiment 1, which used a similar environment with the exception of a certain outcome. We also found a smaller correlation between numeracy and search decisions compared to Experiment 1, suggesting that the relationship between these two variables may

be dependent on the level to which different search strategies disagree. Higher numeracy is correlated with a higher proportion of probability gain queries when there is no disagreement between competing models of informational utility (e.g., Experiment 1, r_{pb}=.25), but this effect diminishes as we introduce a larger number of disagreement between various models (see Table 4.9 for a summary of results across all experiments). Lastly, we replicated for the fourth time a correlation between numeracy and probability judgments, affirming the hypothesis that this is a stable relationship across multiple types of search environments.

4.7 Summary of All Experiments

Having examined the effects of presentation formats on information search and Bayesian reasoning across several different task environments, let us briefly summarize the results. Fig 4.5 shows the results aggregated over all experiments. For an overview of the statistical tests performed, see Table 4.9.

Effect of Formats. We found a strong effect of presentation format on search behavior in Experiment 2, but no significant differences in the other experiments, although the relative rankings of formats in Experiment 3 were highly similar (r_s=.76). When aggregated over all experiments, a one-way between participants ANOVA found significant differences between formats on the information search task [$F(13,2845)$=4.11, p<.001], with a Tukey posthoc test showing that the standard probability format was significantly worse at identifying the higher probability gain query than all other formats, with the exception of the likelihood probability with complement format. Over all experiments, two visual formats, the posterior bar graph and the posterior icon array yielded the highest proportion of correctly identified probability gain queries (71% and 69% respectively), while the posterior probability with complement format performed best out of the numerical formats (68%). Even though these best-performing formats did not always achieve significant levels of reliability in identifying the higher probability gain query in each individual experiment, they represent the best options from a general perspective.

We also found significant differences between formats on the probability judgment task [$F(13,2845)$=2.87, p<.001]. The results of a Tukey posthoc test found that the likelihood variations of the natural frequency formats (with and without complement) had the lowest overall judgment error, which was significantly lower than six other formats: posterior probability, posterior probability with complement, posterior frequency, posterior icon array, likelihood bar graph, and posterior bar graph. This seems to confirm the results of the Bayesian reasoning literature, that the natural frequency formats are very helpful for

deriving single point estimates of probabilities; however, they are not particularly useful in information search tasks compared to their visual counterparts. This strongly points to important differences between numerical and visual representations of natural frequencies that influence information search behavior.

Probability Judgments and Information Search. In contradiction to our original hypothesis, we found no correlation between probability judgment error and search decisions (r_{pb}=-.02). Comparing the relationship between search decisions and judgment error on subsets of the probability judgment task (i.e., base rate, marginals, likelihoods, and posteriors) produced similarly low correlations. We did find a significant relationship in Experiment 1 between search decision and judgment error, but the relatively low correlation coefficient (r_{pb}=.16) suggests that probability judgment accuracy is unlikely to be a strong factor in search performance. Rather, the aggregated data suggests that the ability to provide accurate probability estimates was not a sufficient criterion for identifying the higher probability gain query. Instead, presentation formats played a larger and direct role in guiding search behavior.

Numeracy. The relationship between numeracy and Bayesian reasoning was very robust in all experiments. Not surprisingly, this correlation is also present aggregated across the four experiments (r=.43). However, this relationship does not necessarily extend to information search, where we found a very low correlation (r_{pb}=.05) when aggregated over all experiments, suggesting that the influence numeracy on information search is very limited.

Within individual experiments, we found a significant correlation between numeracy and search behavior in Experiment 1 and 4 (r_{pb}=.25 and r_{pb}=.12, respectively), although this relationship disappears when we introduce disagreements between models (Experiments 2 and 3). When models disagree about which query is more useful, it seems that numeracy skill is not very helpful in differentiating between different search behaviors, and ultimately in identifying the query that is most useful, respective to the goal of maximizing classification accuracy.

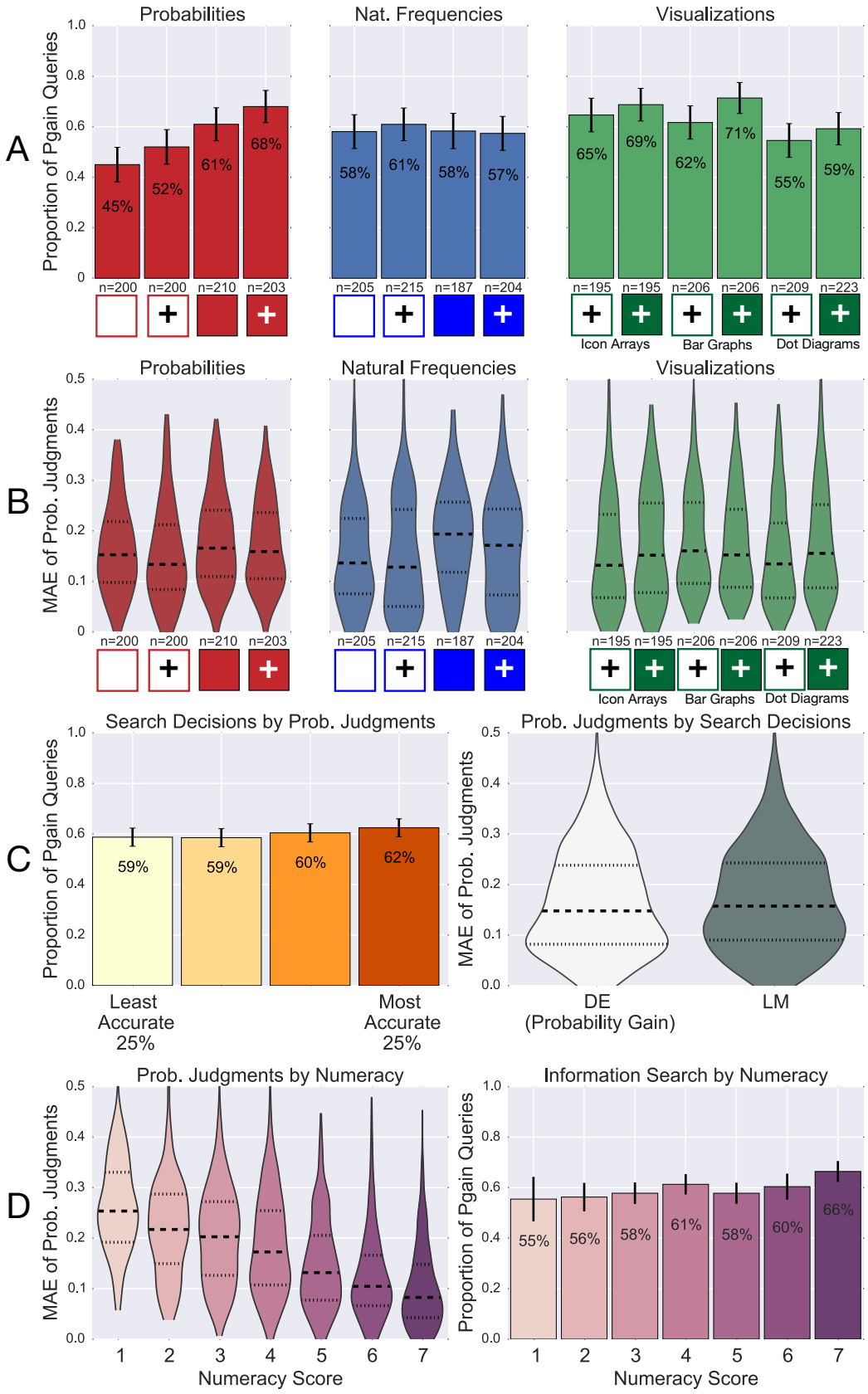


Figure 4.5: Results over all experiments: (A) proportion of probability gain queries (B) distribution of MAE; (C left) search decisions with participants split into quartiles based on probability judgment performance, in increasing order of performance; (C right) MAE split by search decision; (D left); MAE split by numeracy score; (D right) search decisions split by numeracy score.

Table 4.9: Summary of Results from all Experiments

Experiment	Overall Proportion of Probability Gain Queries	Influence of Formats on:		Probability Judgments and Search Decisions	Relationship between Numeracy and:	
		Search Decision	Probability Judgments		Probability Judgments	Search Decision
1 (n=677)	80%, range: [71%–88%]	F(13,664)=0.99, p=.46	F(13,664)=1.18, p=.29	r_{pb}=-.16	r=-.42	r_{pb}=.25
2 (n=817)	61%, range: [27%–86%]	F(13,804)=6.57, p<.001	F(13,804)=1.15, p=.31	r_{pb}=-.04	r=-.40	r_{pb}=.003
3 (n=683)	36%, range: [21%–48%]	F(13,670)=1.38, p=.16	F(13,670)=1.26, p=.24	r_{pb}=-.06	r=-.45	r_{pb}=-.06
4 (n=681)	64%, range: [47%–73%]	F(13,668)=1.39, p=.17	F(13,668)=2.32, p=.005	r_{pb}=-.07	r=-.47	r_{pb}=.12
Overall (n=2858)	60%, range: [45%-71%]	F(13,2845)=4.106, p<.001	F(13,2845)=2.87, p=.001	r_{pb}=-.02	r=-.43	r_{pb}=.05

*Note.*Mean absolute error aggregated across all 11 questions is used as the measure of accuracy for the probability judgment task, thus numeracy being negatively correlated with probability judgments signifies that higher numeracy skill is a predictor for lower error.

5.1 "When" Formats Influence Search Rather than "If"

The results of our experiments showed that presentation formats can have a strong effect on information search behavior (Experiment 2), with some formats found to be as effective as experience-based learning (posterior icon array and posterior bar graph). However, the influence of format is greatly diminished when (1) the more useful question is relatively obvious due to unanimous agreement amongst OED models, and (2) the possibility of a certain outcome is introduced.

5.1.1 When OED Models Models Agree and When They Disagree

When OED models all agree on the more useful query query (Experiment 1), we found that the proportion of probability gain queries could be largely accounted for by numeracy and probability judgments, whereas presentation format had little effect. In contrast, when we introduced disagreement between OED models in Experiment 2, *no amount of numeracy or ability to explicitly judge the underlying probabilities could be linked to a higher proportion of probability gain queries*. Thus, in environments with disagreement amongst OED models, being able to do the math and understanding the probabilities do not seem to be sufficient to improve search decisions. Instead, only the use of clever presentation formats could help reveal which query was more useful. Visual formats such as icon arrays or bar graphs were found to be particularly effective (especially in the posterior variations), although certain numerical formats also performed quite well when they presented the posterior probabilities and included the complement information.

5.1.2 The Possibility of a Certain Outcome

When the possibility of certainty was introduced into a search environment in opposition to probability gain (Experimetn 3), the proportion of probability gain queries was substantially reduced across all formats (see Table 4.9). This was the case even though all OED models (except probability of certainty) were in agreement about the more useful query in Experiment 4.

As was addressed initially, there are many different ways to define the usefulness of a question, relative to different goals. OED models and heuristic strategies approximate these goals, and even though we explicitly defined the task relative to the goal of maximizing classification accuracy (probability gain goal), people may still have a high level of variance in the strategies used in selecting search queries. The probability of certainty model is one such strategy that has been found to have an influence on search behavior, as demonstrated in Experiment 4.

5.2 Relationship Between Probability Judgments and Information Search

Our second major result was that there does not exist a catch-all combination of format, probability judgment accuracy, and numeracy skill, which could act as a stable predictor for the proportion of probability gain queries across all four environments that we studied. Whereas the correlation between higher numeracy skill and lower error on the probability judgment task remained invariable across all four experiments, all the other relationships seemed to be dependent on the nature of the search environment. The performance on the probability judgment task only lead to a higher proportion of optimal search decisions in Experiment 1, where all relevant OED models agreed on the more useful question. Indeed, Experiment 1 was conceived as a baseline study in order to test the upper limits of human performance, and as such, the higher probability gain query was rather obvious compared to the other experiments. Thus, it is not surprising that understanding the probabilities in the task environment seemed to be a sufficient criterion to arrive at the optimal search query. However, in the other environments (Experiments 2, 3, and 4), identifying the higher probability gain query is still not trivial, even with an understanding of the probabilitic environment, due to disagreements amongst potential strategies.

These results require us to rethink the relationship between probability judgments and information search that we had originally hypothesized. Our initial assumptions were based on using explicit probability judgments as an indicator of how well an individual understood

the probabilistic environment. However, our results suggest that (1) explicit probability judgments are insufficient to evaluate how well individuals can integrate probabilistic information in an information search task, or alternatively, (2) there may be a large amount of variance in the different strategies employed by humans to determine the more useful query, washing away the predictive power of probability judgments.

5.3 The Frequency Difference Heuristic

The use of heuristic strategies could explain how the different presentation formats influenced search decisions directly, without being mediated by probability judgments. Gigerenzer & Hoffrage (1995) argued that natural frequency formats facilitated Bayesian inference tasks because they simplified the computations required to compute a posterior probability. In the same vein, we discovered a simple heuristic strategy that we believe invariably identifies the higher probability gain query for all variations of a two-category, binary feature search task (see Conjecture 5.3.1), but through simpler computations than any of the OED models. The FREQUENCY DIFFERENCE HEURISTIC is a relevant strategy for all natural frequency formats, although it is especially salient for visualizations of natural frequencies. This newly discovered strategy can be executed by merely choosing the query with the largest absolute difference of natural frequencies, when comparing each query outcome. We can model this in terms of other OED models, where the expected utility of the query, Q, is the absolute difference between the joint frequencies $N(c_1 \wedge q_j)$ and $N(c_2 \wedge q_j)$ for the outcome q_j, where the absolute difference is largest (equation 5.1).

$$eu_{freqDiff}(Q) = \max_{j=1}^{m} |N(c_1 \wedge q_j) - N(c_2 \wedge q_j)| \tag{5.1}$$

This heuristic strategy is most salient for the posterior variations of the natural frequency and visual formats, because $N(c_1 \wedge q_j)$ and $N(c_2 \wedge q_j)$ are grouped together and easier to compare, which is a possible explanation for why the posterior variations yielded the highest proportion of probability gain queries in Experiment 2. There is also a possibility that the poor performance of the dot diagrams in the search task could be explained by this heuristic. The presentation of the dots—placed in a random uniform distribution within a container of fixed area—makes the task of comparing absolute differences much less intuitive than comparing relative differences (through a density estimate). The reason that comparing relative differences is undesirable, is because it re-normalizes the natural frequencies as absolute frequencies, and loses the additional information about base rate and marginal probabilities that is contained within the natural frequency representation. Comparing

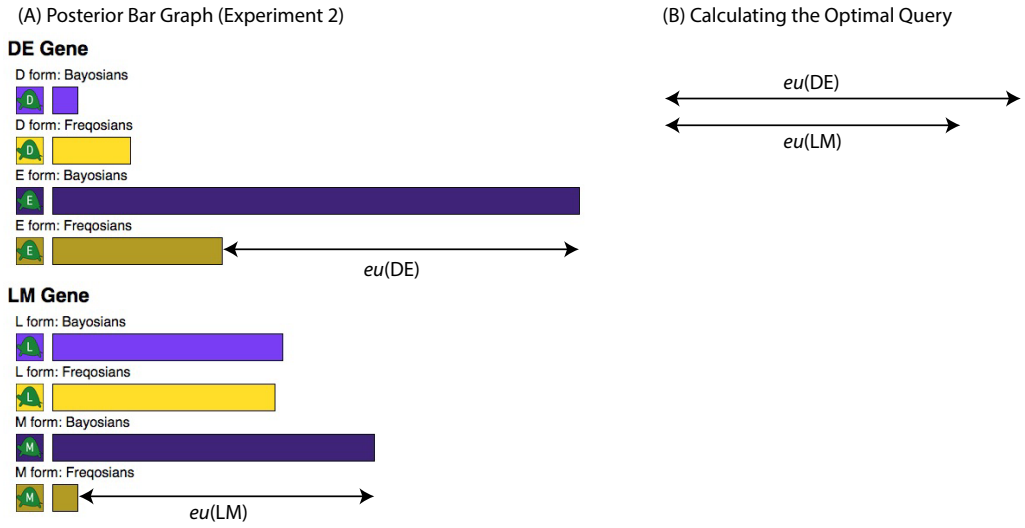


Figure 5.1: An example of the frequency difference heuristic applied to the posterior bar graph from Experiment 2. The higher probability gain query (DE test) is the one with the largest difference between the lengths of the bars.

relative differences does not approximate the probability gain model and is not guaranteed to yield the same search behavior.

While we do not have evidence that this heuristic strategy was employed—or if it was, what proportion of subjects utilized it—the ability to explain differences across formats suggests that it should be considered as a potential strategy used by participants. The fact that visual formats did not have a significant effect on search decisions when the possibility of certainty was introduced (Experiment 3) does not necessarily create any inconsistencies with the heuristic explanation, because the probability of certainty model may itself be a competing heuristic strategy, which was shown to account for a substantial proportion of search behavior (comparing Experiment 1 and 4). Additionally, the likelihood difference heuristic (Nelson, 2005) is also available to the probability formats, which also led to the same search decision as a probability gain query in Experiment 1. Thus, the frequency difference heuristic may be one of several different cognitive strategies employed by participants in the information search task, with the unique quality that it exactly implements probability gain.

5.3.1 Conjecture: the Frequency Difference Heuristic is Exactly Equivalent to Probability Gain Search Decisions

Definition 5.3.1 *Let Q represent one of two possible search queries, $Q = \{D, L\}$, where both queries have binary outcomes, such that $d_j \in \{d_1, d_2\}$ and $l_j \in \{l_1, l_2\}$. The goal of an information search task is to select the query that maximizes the expected probability of*

classifying an unknown item, C, which belongs to one of two possible classes, $c_i \in \{c_1, c_2\}$.

Lemma 5.3.1 *Probability gain prefers the query with the highest expected utility, $eu_{PG}(Q) = \sum_{j=1}^{m} P(q_j) \max_{i=1}^{n} P(c_i|q_j) - \max_{i=1}^{n} P(c_i)$. Thus, we can construct a preference function, $\lambda_{PG}(D, L) = eu_{PG}(D) - eu_{PG}(L)$, such that $\lambda_{PG}(D, L)$ is positive when query D is more useful, negative when query L is more useful, and is 0 when there is no preference. Thus a choice function modeling search behavior can be defined as,*

$$Choice_{PG}(D, L) = \begin{cases} D, & \text{if } \lambda_{PG}(D, L) > 0 \\ L, & \text{if } \lambda_{PG}(D, L) < 0 \\ noPref, & \text{otherwise} \end{cases}$$

Lemma 5.3.2 *The Frequency Difference heuristic strategy can describe the expected usefulness a query as $eu_{freqDiff}(Q) = \max_{j=1}^{m} |N(c_1 \wedge q_j) - N(c_2 \wedge q_j)|$. Likewise with the probability gain strategy, we can also construct a preference function, $\lambda_{freqDiff}(D, L) = eu_{freqDiff}(D) - eu_{freqDiff}(L)$, such that $\lambda_{freqDiff}(D, L)$ is positive when query D is more useful, negative when query L is more useful, and is 0 when there is no preference. The equivalent choice function is defined as:*

$$Choice_{freqDiff}(D, L) = \begin{cases} D, & \text{if } \lambda_{freqDiff}(D, L) > 0 \\ L, & \text{if } \lambda_{freqDiff}(D, L) < 0 \\ noPref, & \text{otherwise} \end{cases}$$

Conjecture 5.3.1 *When neither $\lambda_{PG}(D, L)$ nor $\lambda_{freqDiff}(D, L)$ are 0, ie., when both models have a preference for one of the two queries, both models will choose the same query. Thus, $\lambda_{PG}(D, L) * \lambda_{freqDiff}(D, L) > 0$,*

Remark 5.3.1 *Through computer simulations iterating through the entire range of environmental probabilities for a binary search environment with a step size of 0.01, we found that there does not exist any situation where probability gain and the frequency difference heuristic disagree about which query is more useful. We also replicated these results by randomly generating 10 million binary search environments, which produced the same findings.*

5.4 Explaining Differences Between Formats

Aggregated over all experiments, icon arrays and bar graphs solicited a higher proportion of probability gain queries than the dot diagrams, suggesting that spatial extent plays a more important role in search decisions than the part-to-whole information embodied by the dot diagrams. Moreover, we did not find any strong differences between icon arrays

and bar graphs in terms of search behavior, discrediting the hypothesis that the ability to calculate exact numerical quantities from visual representations (i.e., countability) plays any substantial role in perceiving the usefulness of a question. Additionally, icon arrays and bar graphs consistently outperformed the numerical natural frequency formats, even though they represented the same quantities.

One potential hypothesis is that when given a set of numbers, it is not clear whether to compare them in absolute terms (via addition or subtraction) or in relative terms (via division or multiplication). The frequency difference heuristic shows that there exists a computational shortcut for identifying the more useful question, but requires the comparison of absolute differences of natural frequencies, and not relative differences. Thus, it may be the case that icon arrays and bar graphs are able to better facilitate an absolute comparison by using spatial extent to represent a quantity rather than a numerical representation. Additionally, in a review of neuroimaging studies on numerical and visual cognition, Hubbard et al. (2005) make the claim that the problem of computing spatial interactions is formally identical to that of computing addition or subtraction, based on evidence from overlapping activations in the Posterior Parietal Cortex. Thus, visualizations that utilize spatial extent may facilitate perceptual processes that aid in the comparison of absolute differences between natural frequencies.

Part-to-whole information may also be useful in order to correctly compare the relevant information, as evidenced by the difference of performance between likelihood and posterior variations of icon arrays and bar graphs. However we only tested spatial extent and part-to-whole information independently from each other. Future studies may find even higher levels search performance when combining spatial extent and part-to-whole information in a single format.

5.5 Limitations

One important requirement for a subject to choose the higher probability gain query is that they correctly understood the goal of the task. In the design of the experiment, we have attempted to make the information search task as clear as possible, even requiring subjects to confirm they had understood the binary nature of the DNA tests with a simple skill-testing question. However, there may be individual differences in how subjects interpreted the search goal of maximizing expected classification accuracy. Subjects may have had difficulty interpreting the concept of 'expected' classification accuracy resulting from a search query, and merely chosen the query that had the highest 'potential' classification outcome. This corresponds to a max-max strategy, which could explain the overall low proportion

of probability gain queries in environments with the possibility of certainty. Indeed, the frequency difference heuristic is also a form of a max-max strategy, although it maximizes the absolute difference between natural frequencies rather than a single numerical entity.

Additionally, Meder and Nelson (2015) address the difference between short-term and long-term optimality in information search. Whereas the experiments presented in this thesis consist of one-shot information search tasks where subjects are only able to make a single query, real-world search environments often afford the possibility of making multiple sequential search decisions. Through computational modeling, Meder and Nelson (2015) found that step-wise probability gain (i.e., measuring the usefulness of a query based on a single question) is not effective at maximizing expected classification accuracy in a long-term, multi-step search task (see also Meier & Blair, 2013; Bramley et al., 2015). Thus, while our task was a restricted to a single query, it may be the case that individuals are more adapted to an unrestricted and multi-step information search task, which would exclude step-wise probability gain from being a reasonable normative description.

In a review of 20 years of research on natural frequencies, McDowell & Jacobs (2015) found that when there was an incongruence between presentation format and the answer format for a Bayesian inference task (i.e., presenting information in terms of natural frequencies, but asking for a probability judgment as an absolute probability), the effect size of natural frequencies was reduced. This was also the case in our experiments, where subjects were given statistical information in a variety of formats, but always asked to provide probability judgments in terms of absolute probabilities, although the analog scale slider may have had some (limited) congruence with the bar graphs and the icon array. A lack of congruence may have reduced the advantages of natural frequency formats (including visual formats) in accurately estimating probabilities, thereby also diminishing a potential correlation between probability judgments and search decisions.

Lastly, more information was presented to fully describe a binary information search task than in a typical Bayesian reasoning task (e.g., two binary outcomes rather than a single binary outcome). Thus, subjects were required to process more information as well as provide a larger set of probability judgments (11 questions) than the single probability judgment used in previous Bayesian reasoning studies (for a review see Meder & Gigerenzer, 2014). The increased complexity of the probabilistic environment, as well the increased number of probability judgments may have reduced judgment accuracy across the board. Our results may be hitting a floor effect, whereby we are unable to distinguish some of the differences attributable to format. Also, we used a flat fee for this task, which may be sufficient to motivate a Bayesian reasoning task consisting of a single probability estimate,

whereas it may have been insufficient in the case of our set of 11 questions.

5.6 Conclusion

The purpose of this thesis was to examine whether or not presentation formats have an effect on search decisions. Consequently, we found that in some environments (Experiment 2 and to a lesser extent Experiment 3) presentation formats make a great deal of difference in search behavior. Although significant differences between formats were not present in each individual experiment, we found that aggregated over all experiments, the standard probability format was significantly worse at identifying the higher probability gain query than all other formats. Over all, the posterior variations of the icon array and bar graph yielded the highest proportion of queries maximizing classification accuracy (71% and 69%, respectively) while the posterior probability format performed the best out of the numerical formats (68%).

The design of the search environments also allowed us to consider a set of hypotheses based on the unique differences between experiments. We found that when probability gain, information gain, impact, and KL-Divergence all agreed on the more useful query (Experiment 1), numeracy and probability judgment error were able to predict search behavior. However, when we introduced disagreement in Experiment 2, only formats played a role in influencing search decisions. No amount of numeracy skill or probability judgment accuracy was found to helpful in identifying the accuracy-maximizing query.

Additionally, the introduction of a certain outcome (Experiments 3 and 4) reduced the proportion of accuracy-maximizing search decisions across all formats. The results of Experiment 4 showed that the possibility of a certain outcome had a substantial effect on search decisions, even when in opposition to all the other OED models.

The goal of this thesis project was not only to look for a relationship between presentation format and search behavior, but also to explain how these two variables are related. To this end, we designed the experiments to provide an insight into the relationship between format, probability judgments, numeracy, and search behavior. The result of this investigation was that the ability to make explicit probability judgments is not a good indicator of how well individuals can make search decisions about a probabilistic environment. This raises the concern that Bayesian reasoning literature is overly concerned with explicit probability judgments, which may have no relation to how humans make decisions about probabilistic information. In environments where the more useful question is ambiguous and the normative guidance of various OED models disagree about the optimal query, we found that no amount of numeracy skill or ability to make probability judgments was sufficient to make better

search decisions. Instead, only presentation formats played a role in influencing information search.

Lastly, the discovery of the frequency difference heuristic and the explanations that it offers in terms of individual differences between formats suggests that there may be a direct link between presentation format and search behavior, independent of probability judgments and numeracy skill as we originally hypothesized.

References

Ancker, J. S., Senathirajah, Y., Kukafka, R., & Starren, J. B. (2006). Design features of graphs in health risk communication: a systematic review. *Journal of the American Medical Informatics Association*, *13*(6), 608–618.

Anderson, B., & Gigerenzer, G. (2012). Statistical literacy in obstetricians and gynecologists. *Journal for Healthcare Quality*, *00*(1), 1–13. doi: 10.1111/j.1945-1474.2011.00194.x

Baron, J. (1985). *Rationality and intelligence*. Cambridge University Press.

Baron, J., Beattie, J., & Hershey, J. C. (1988). Heuristics and biases in diagnostic reasoning: Ii. congruence, information, and certainty. *Organizational Behavior and Human Decision Processes*, *42*(1), 88–110.

Bodemer, N., Meder, B., & Gigerenzer, G. (2014). Communicating Relative Risk Changes with Baseline Risk: Presentation Format and Numeracy Matter. *Medical decision making : an international journal of the Society for Medical Decision Making*(July), 615–626. doi: 10.1177/0272989X14526305

Bramley, N. R., Lagnado, D. A., & Speekenbrink, M. (2015). Conservative forgetful scholars: How people learn causal structure through sequences of interventions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(3), 708.

Brase, G. L. (2009). Pictorial representations and numerical representations in Bayesian reasoning. *Applied Cognitive Psychology*, *23*, 369–381. doi: 10.1002/acp

Brown, S. M., Culver, J. O., Osann, K. E., MacDonald, D. J., Sand, S., Thornton, A. A., ... others (2011). Health literacy, numeracy, and interpretation of graphical breast cancer risk estimates. *Patient education and counseling*, *83*(1), 92–98.

Butko, N. J., & Movellan, J. R. (2008). I-pomdp: An infomax model of eye movement. In *Development and learning, 2008. icdl 2008. 7th ieee international conference on* (pp. 139–144).

Cleveland, W. S., & McGill, R. (1985). Graphical perception and graphical methods for analyzing scientific data. *Science*, *229*(4716), 828–833.

Cokely, E. T., Galesic, M., Schulz, E., Ghazal, S., & Garcia-Retamero, R. (2012). Measuring risk literacy: The berlin numeracy test. *Judgment and Decision Making*, *7*(1), 25–47.

Cosmides, L., & Tooby, J. (1996). Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. *Cognition*, *58*, 1–73.

Covey, J. (2007). A meta-analysis of the effects of presenting treatment benefits in different formats. *Medical Decision Making*, *27*(5), 638–654.

Crupi, V., Nelson, J., Meder, B., Cevolani, G., & Tentori, K. (2015). Generalized information theory meets human cognition: Introducing a unified framework to model uncertainty and information search.
(Manuscript in preparation)

Forrow, L., Taylor, W. C., & Arnold, R. M. (1992). Absolutely relative: how research results are summarized can affect treatment decisions. *The American journal of medicine*, *92*(2), 121–124.

Gaissmaier, W., Wegwarth, O., Skopec, D., Müller, A.-S., & Broschinski, S. (2012). Numbers can be worth a thousand pictures: individual differences in understanding graphical and numerical representations of health-related information. *Health psychology : official journal of the Division of Health Psychology, American Psychological Association...*, *31*(3), 286–296. doi: 10.1037/a0024850

Galesic, M., Garcia-Retamero, R., & Gigerenzer, G. (2009, March). Using icon arrays to communicate medical risks: overcoming low numeracy. *Health psychology : official journal of the Division of Health Psychology, American Psychological Association*, *28*(2), 210–6. doi: 10.1037/a0014474

Gigerenzer, G. (1998). What are natural frequencies? *Acad Med*, 1–2. doi: 10.1136/bmj.d6386

Gigerenzer, G. (2000). *Adaptive thinking: Rationality in the real world*. Oxford University Press.

Gigerenzer, G., Gaissmaier, W., Kurz-Milcke, E., Schwartz, L. M., & Woloshin, S. (2007). Helping Doctors and Patients Make Sense of Health Statistics. *Psychological science in the public interest*, *8*(2), 53–96.

Gigerenzer, G., & Hoffrage, U. (1995). How to Improve Bayesian Reasoning Without Instruction: Frequency Formats. *Psychological review*, *102*(4), 684–704.

Ginzburg, I., & Sejnowski, T. J. (1996). Dynamics of rule induction by making queries: Transition between strategies. In *Proceedings of the 18th annual conference of the cognitive science society* (pp. 121–125).

Hoffrage, U., Gigerenzer, G., Krauss, S., & Martignon, L. (2002, July). Representation facilitates reasoning: what natural frequencies are and what they are not. *Cognition*, *84*(3), 343–52.

Hoffrage, U., Lindsey, S., Hertwig, R., & Gigerenzer, G. (2000). Statistical Communicating Information. , *290*(December).

Hubbard, E. M., Piazza, M., Pinel, P., & Dehaene, S. (2005). Interactions between number and space in parietal cortex. *Nature Reviews Neuroscience*, *6*(6), 435–448.

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

Kleiter, G. D. (1994). Natural sampling: Rationality without base rates. In *Contributions to mathematical psychology, psychometrics, and methodology* (pp. 375–388). Springer.

Knight, F. H. (1923). *Risk, uncertainty and profit*. Courier Corporation. (Original work published in 1923)

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics*, 79–86.

Lagae, A., & Dutré, P. (2008). A comparison of methods for generating poisson disk distributions. In *Computer graphics forum* (Vol. 27, pp. 114–129).

Lindeman, S. T., van den Brink, W. P., & Hoogstraten, J. (1988). *Effect of Feedback on Base-Rate Utilization* (Vol. 67) (No. 2). doi: 10.2466/pms.1988.67.2.343

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

Markant, D., & Gureckis, T. M. (2012). Does the utility of information influence sampling behavior. In *Proceedings of the 34th annual conference of the cognitive science society* (pp. 719–724).

Marr, D. (1982). *Vision: A computational approach*. Freeman & Co., San Francisco.

Martignon, L., & Krauss, S. (2003). Can l'homme éclairé be fast and frugal? reconciling bayesianism and bounded rationality. *Emerging perspectives on judgment and decision research*, 108–22.

McDowell, M., & Jacobs, P. (2015). *Twenty years of natural frequencies: A systematic review and meta-analysis of the effect of natural frequencies on bayesian reasoning.* Retrieved from `http://www.spudm25.eu/images/docs/Abstract_book_FINAL.pdf` (Subjective Probability, Utility, and Decision Making (SPUDM) Conference)

McKenzie, C. R. (2004). Framing effects in inference tasks—and why they are normatively defensible. *Memory & cognition*, *32*(6), 874–885.

Meder, B., & Gigerenzer, G. (2014). Statistical thinking: no one left behind. In *Probabilistic thinking* (pp. 127–148). Springer.

Meder, B., Le Lec, F., & Osman, M. (2013). Decision making in uncertain times: what can cognitive and decision sciences say about or learn from economic crises? *Trends in cognitive sciences*, *17*(6), 257–260.

Meder, B., & Nelson, J. (2015). *When optimal experimental design models are not optimal: Sequential search, optimality, and heuristics.* (Unpublished manuscript)

Meder, B., & Nelson, J. D. (2012). Information search with situation-specific reward functions. *Judgment and Decision Making*, *7*(2), 119–148.

Medin, D. L., & Edelson, S. M. (1988). Problem structure and the use of base-rate information from experience. *Journal of experimental psychology. General*, *117*(1), 68–85. doi: 10.1037/0096-3445.117.1.68

Meier, K. M., & Blair, M. R. (2013). Waiting and weighting: Information sampling is a balance between efficiency and error-reduction. *Cognition*, *126*(2), 319–325.

Micallef, L., Dragicevic, P., & Fekete, J. D. (2012). Assessing the effect of visualizations on bayesian reasoning through crowdsourcing. *IEEE Transactions on Visualization and Computer Graphics*, *18*(12), 2536–2545. doi: 10.1109/TVCG.2012.199

Najemnik, J., & Geisler, W. S. (2008). Eye movement statistics in humans are consistent with an optimal search strategy. *Journal of Vision*, *8*(3), 4.

Navarro, D. J., & Perfors, A. F. (2011). Hypothesis generation, sparse categories, and the positive test strategy. *Psychological review*, *118*(1), 120.

Nelson, J. (2005). Finding useful questions: on Bayesian diagnosticity, probability, impact, and information gain. *Psychological review*, *112*(4), 979–999. doi: 10.1037/0033-295X.112.4.979

Nelson, J. (2008). Towards a rational theory of human information acquisition. In *The probabilistic mind: prospects for rational models of cognition* (pp. 143–164). doi: 10.1093/acprof:oso/9780199216093.003.0007

Nelson, J., & Cottrell, G. W. (2007). A probabilistic model of eye movements in concept formation. *Neurocomputing*, *70*(13), 2256–2272.

Nelson, J., McKenzie, C., Cottrell, G., & Sejnowski, T. (2010). Experience matters information acquisition optimizes probability gain. *Psychological science*. doi: 10.1177/0956797610372637

Nelson, J., & Movellan, J. R. (2001). Active inference in concept induction. *Advances in neural information processing systems*(13), 45–51.

Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review*, *101*(4), 608.

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on amazon mechanical turk. *Judgment and Decision making*, *5*(5), 411–419.

Rehder, B., & Hoffman, A. B. (2005). Eyetracking and selective attention in category learning. *Cognitive psychology*, *51*(1), 1–41.

Rusconi, P., & McKenzie, C. R. (2013). Insensitivity and oversensitivity to answer diagnosticity in hypothesis testing. *The Quarterly Journal of Experimental Psychology*, *66*(12), 2443–2464.

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

Schwartz, L. M., Woloshin, S., Black, W. C., & Welch, H. G. (1997). The role of numeracy in understanding the benefit of screening mammography. *Annals of internal medicine*, *127*(11), 966–972.

Simon, H. a. (1978). On the forms of mental representation. *Perception and cognition: Issues in the foundations of psychology.*, 3–18.

Skov, R. B., & Sherman, S. J. (1986). Information-gathering processes: Diagnosticity, hypothesis-confirmatory strategies, and perceived hypothesis confirmation. *Journal of Experimental Social Psychology*, *22*(2), 93–121.

Sloman, S. A., Over, D., Slovak, L., & Stibel, J. M. (2003). Frequency illusions and other fallacies. *Organizational Behavior and Human Decision Processes*, *91*(2), 296–309.

Slowiaczek, L. M., Klayman, J., Sherman, S. J., & Skov, R. B. (1992). Information selection and use in hypothesis testing: What is a good question, and what is a good answer? *Memory & Cognition*, *20*(4), 392–405.

Stone, E. R., Sieck, W. R., Bull, B. E., Yates, J. F., Parks, S. C., & Rush, C. J. (2003). Foreground: background salience: Explaining the effects of graphical displays on risk avoidance. *Organizational Behavior and Human Decision Processes*, *90*(1), 19–36.

Tsallis, C. (1988). Possible generalization of boltzmann-gibbs statistics. *Journal of statistical physics*, *52*(1-2), 479–487.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *science*, *185*(4157), 1124–1131.

Wells, G. L., & Lindsay, R. (1980). On estimating the diagnosticity of eyewitness nonidentifications. *Psychological bulletin*, *88*(3), 776.

Zhu, L., & Gigerenzer, G. (2006, January). Children can solve Bayesian problems: The role of representation in mental computation. *Cognition*, *98*(3), 287–308. doi: 10.1016/j.cognition.2004.12.003

Appendix A: Search Environments

Table A1: Expected utilities for each search environment.

Experiment	OED Model	DE Test					LM Test					Preferred Query
		$P(D)$	$u(D)$	$P(E)$	$u(E)$	**eu(DE)**	$P(L)$	$u(L)$	$P(M)$	$u(M)$	**eu(LM)**	
1	Probability Gain	45%	0.39	55%	0.32	**0.35**	20%	0.25	80%	0.06	**0.1**	DE Test
	Information Gain		0.5		0.32	**0.4**		0.19		0.01	**0.05**	DE Test
	KL-Divergence		0.5		0.32	**0.4**		0.19		0.01	**0.05**	DE Test
	Impact		0.39		0.32	**0.35**		0.25		0.06	**0.1**	DE Test
	Probability of Certainty		0		0	**0**		0		0	**0**	—
2	Probability Gain	11%	0.11	89%	0.06	**0.07**	57%	-0.19	43%	0.25	**0**	DE Test
	Information Gain		0.18		0.09	**0.1**		-0.12		0.6	**0.19**	LM Test
	KL-Divergence		0.81		0.01	**0.1**		0.12		0.29	**0.19**	LM Test
	Impact		0.51		0.06	**0.11**		0.19		0.25	**0.22**	LM Test
	Probability of Certainty		0		0	**0**		0		0	**0**	—
3	Probability Gain	14%	0.1	86%	0.08	**0.08**	60%	-0.2	40%	0.2	**0**	DE Test
	Information Gain		0.16		0.12	**0.13**		-0.12		0.88	**0.28**	LM Test
	KL-Divergence		0.77		0.02	**0.13**		0.12		0.51	**0.28**	LM Test
	Impact		0.5		0.08	**0.14**		0.2		0.3	**0.24**	LM Test
	Probability of Certainty		0		0	**0**		0		1	**0.4**	LM Test
4	Probability Gain	25%	0.19	75%	0.22	**0.21**	56%	-0.22	44%	0.28	**0**	DE Test
	Information Gain		0.44		0.51	**0.49**		-0.14		0.86	**0.29**	DE Test
	KL-Divergence		1.3		0.22	**0.49**		0.15		0.47	**0.29**	DE Test
	Impact		0.63		0.22	**0.32**		0.22		0.28	**0.25**	DE Test
	Probability of Certainty		0		0	**0**		0		9	**0.44**	LM Test

Note. For each test, $P(\bullet)$ denotes the probability of the outcome, $u(\bullet)$ denotes the utility of the outcome, and $eu(\bullet)$ denotes the expected utility of the test. All of the OED models compute $eu(\bullet)$ through a normalized mean, that weights each individual outcome utility, $u(\bullet)$, by the probability of the outcome, $P(\bullet)$. '—' denotes that neither of the queries has a higher expected utility, and that there is no preference.

Appendix B: Presentation Formats

Experiment 1

Table B1: Each Type of Format used in Experiment 1

Format	Example
Standard Probability	Consider a turtle picked at random from the **100** turtles on the island: The probability that it is a Bayosian turtle is **50%**. If a turtle is a Bayosian, then the probability that it has the D form of the DE gene is **10%**. If a turtle is a Freqosian, then the probability that it has the D form of the DE gene is **80%**. If a turtle is a Bayosian, then the probability that it has the L form of the LM gene is **10%**. If a turtle is a Freqosian, then the probability that it has the L form of the LM gene is **30%**.
Probability with Complement	Consider a turtle picked at random from the **100** turtles on the island: The probability that it is a Bayosian turtle is **50%**, and the probability that it is a Freqosian turtle is **50%**. If a turtle is a Bayosian, then the probability that it has the D form of the DE gene is **10%**, and the probability that it has the E form is **90%**. If a turtle is a Freqosian, then the probability that it has the D form of the DE gene is **80%**, and the probability that it has the E form is **20%**. If a turtle is a Bayosian, then the probability that it has the L form of the LM gene is **10%**, and the probability that it has the M form is **90%**. If a turtle is a Freqosian, then the probability that it has the L form of the LM gene is **30%**, and the probability that it has the M form is **70%**.

Posterior Probability

Consider a turtle picked at random from the **100** turtles on the island:
The probability that it has the D form of the DE gene is **45%**.
The probability that it has the L form of the LM gene is **20%**.

If a turtle has the D form of the DE gene, then the probability that it is a Bayosian turtle is **11%**.
If a turtle has the E form of the DE gene, then the probability that it is a Bayosian turtle is **82%**.

If a turtle has the L form of the LM gene, then the probability that it is a Bayosian turtle is **25%**.
If a turtle has the M form of the LM gene, then the probability that it is a Bayosian turtle is **56%**.

Posterior Probability with Complement

Consider a turtle picked at random from the **100** turtles on the island:
The probability that it has the D form of the DE gene is **45%**, and the probability that it has the E form is **55%**.
The probability that it has the L form of the LM gene is **20%**, and the probability that it has the M form is **80%**.

If a turtle has the D form of the DE gene, then the probability that it is a Bayosian turtle is **11%**, and the probability that it is a Freqosian turtle is **89%**.
If a turtle has the E form of the DE gene, then the probability that it is a Bayosian turtle is **82%**, and the probability that it is a Freqosian turtle is **18%**.

If a turtle has the L form of the LM gene, then the probability that it is a Bayosian turtle is **25%**, and the probability that it is a Freqosian turtle is **75%**.
If a turtle has the M form of the LM gene, then the probability that it is a Bayosian turtle is **56%**, and the probability that it is a Freqosian turtle is **44%**.

Natural Frequency

Out of the **100** turtles on the island, **50** are Bayosian turtles.

Out of the **50** Bayosian turtles, **5** turtles have the D form of the DE gene.
Out of the **50** Freqosian turtles, **40** turtles have the D form of the DE gene.

Out of the **50** Bayosian turtles, **5** turtles have the L form of the LM gene.
Out of the **50** Freqosian turtles, **15** turtles have the L form of the LM gene.

Natural Frequency with Complement

Out of the **100** turtles on the island, **50** are Bayosian turtles and **50** are Freqosian turtles.

Out of the **50** Bayosian turtles, **5** turtles have the D form of the DE gene, and **45** turtles have the E form of the gene.
Out of the **50** Freqosian turtles, **40** turtles have the D form of the DE gene, and **10** turtles have the E form of the gene.

Out of the **50** Bayosian turtles, **5** turtles have the L form of the LM gene, and **45** turtles have the M form of the gene.
Out of the **50** Freqosian turtles, **15** turtles have the L form of the LM gene, and **35** turtles have the M form of the gene.

Posterior Frequency

Out of the **100** turtles on the island, **45** have the D form of the DE gene.
Out of the **100** turtles on the island, **20** have the L form of the LM gene.

Out of the **45** turtles with the D form of the DE gene, **5** are Bayosian turtles.
Out of the **55** turtles with the E form of the DE gene, **45** are Bayosian turtles.

Out of the **20** turtles with the L form of the LM gene, **5** are Bayosian turtles.
Out of the **80** turtles with the M form of the LM gene, **45** are Bayosian turtles.

Posterior Frequency with Complement

Out of the **100** turtles on the island, **45** have the D form of the DE gene, and **55** turtles have the E form of the gene.
Out of the **100** turtles on the island, **20** have the L form of the LM gene, and **80** turtles have the M form of the gene.

Out of the **45** turtles with the D form of the DE gene, **5** are Bayosian turtles and **40** are Freqosian turtles.
Out of the **55** turtles with the E form of the DE gene, **45** are Bayosian turtles and **10** are Freqosian turtles.

Out of the **20** turtles with the L form of the LM gene, **5** are Bayosian turtles and **15** are Freqosian turtles.
Out of the **80** turtles with the M form of the LM gene, **45** are Bayosian turtles and **35** are Freqosian turtles.

Icon Array

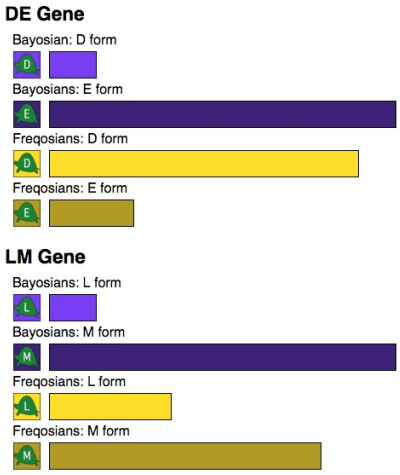
Below you see a graphical representation of the distribution of the DE gene and the LM gene in Bayosian and Freqosian turtles. Each icon represents **1** out of the total **100** turtles on the island.

Posterior Icon Array

Below you see a graphical representation of the distribution of Bayosian and Freqosian turtles for the DE gene and the LM gene. Each icon represents **1** out of the total **100** turtles on the island.

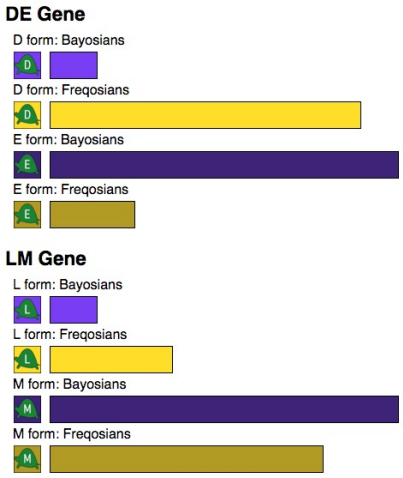
Bar Graph

Below you see a graphical representation of the distribution of the DE gene and the LM gene in Bayosian and Freqosian turtles. The length of each bar (relative to the span of the white space on the page) indicates the proportion of the total **100** turtles on the island.



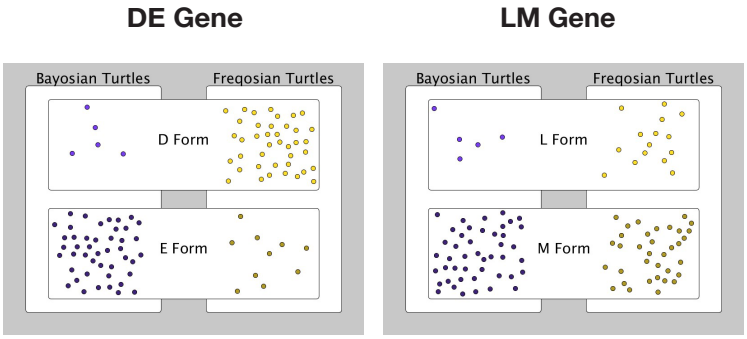
Posterior Bar Graph

Below you see a graphical representation of the distribution of Bayosian and Freqosian turtles for the DE gene and the LM gene. The length of each bar (relative to the span of the white space on the page) indicates the proportion of the total **100** turtles on the island.



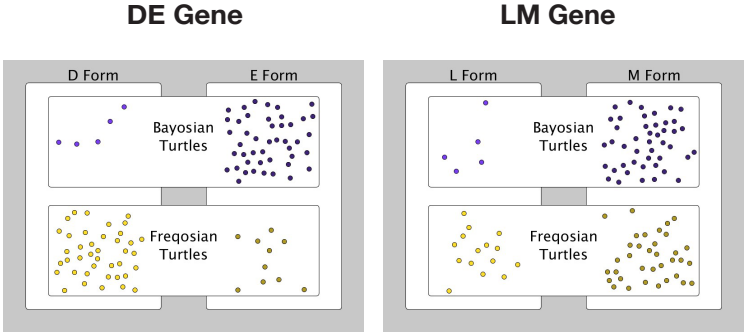
Dot Diagram

Below you see a graphical representation of the distribution of the DE gene and the LM gene in Bayosian and Freqosian turtles. Each dot represents **1** out of the total **100** turtles on the island.



Posterior Dot Diagram

Below you see a graphical representation of the distribution of Bayosian and Freqosian turtles for the DE gene and the LM gene. Each dot represents **1** out of the total **100** turtles on the island.



Note. Likelihood formats are shown as empty boxes, whereas posterior formats are shown as filled in boxes. "+" denotes that a format includes the complement information.

Experiment 2

Table B2: Each Type of Format used in Experiment 2

Format	Example
 Standard Probability	Consider a turtle picked at random from the **100** turtles on the island: The probability that it is a Bayosian turtle is **70%**. If a turtle is a Bayosian, then the probability that it has the D form of the DE gene is **3%**. If a turtle is a Freqosian, then the probability that it has the D form of the DE gene is **30%**. If a turtle is a Bayosian, then the probability that it has the L form of the LM gene is **41%**. If a turtle is a Freqosian, then the probability that it has the L form of the LM gene is **93%**.
 Probability with Complement	Consider a turtle picked at random from the **100** turtles on the island: The probability that it is a Bayosian turtle is **70%**, and the probability that it is a Freqosian turtle is **30%**. If a turtle is a Bayosian, then the probability that it has the D form of the DE gene is **3%**, and the probability that it has the E form is **97%**. If a turtle is a Freqosian, then the probability that it has the D form of the DE gene is **30%**, and the probability that it has the E form is **70%**. If a turtle is a Bayosian, then the probability that it has the L form of the LM gene is **41%**, and the probability that it has the M form is **59%**. If a turtle is a Freqosian, then the probability that it has the L form of the LM gene is **93%**, and the probability that it has the M form is **7%**.
 Posterior Probability	Consider a turtle picked at random from the **100** turtles on the island: The probability that it has the D form of the DE gene is **11%**. The probability that it has the L form of the LM gene is **57%**. If a turtle has the D form of the DE gene, then the probability that it is a Bayosian turtle is **19%**. If a turtle has the E form of the DE gene, then the probability that it is a Bayosian turtle is **76%**. If a turtle has the L form of the LM gene, then the probability that it is a Bayosian turtle is **51%**. If a turtle has the M form of the LM gene, then the probability that it is a Bayosian turtle is **95%**.

Posterior Probability with Complement

Consider a turtle picked at random from the **100** turtles on the island:
The probability that it has the D form of the DE gene is **11%**, and the probability that it has the E form is **89%**.
The probability that it has the L form of the LM gene is **57%**, and the probability that it has the M form is **43%**.

If a turtle has the D form of the DE gene, then the probability that it is a Bayosian turtle is **19%**, and the probability that it is a Freqosian turtle is **81%**.
If a turtle has the E form of the DE gene, then the probability that it is a Bayosian turtle is **76%**, and the probability that it is a Freqosian turtle is **24%**.

If a turtle has the L form of the LM gene, then the probability that it is a Bayosian turtle is **51%**, and the probability that it is a Freqosian turtle is **49%**.
If a turtle has the M form of the LM gene, then the probability that it is a Bayosian turtle is **95%**, and the probability that it is a Freqosian turtle is **5%**.

Natural Frequency

Out of the **100** turtles on the island, **70** are Bayosian turtles.

Out of the **70** Bayosian turtles, **2** turtles have the D form of the DE gene.
Out of the **30** Freqosian turtles, **9** turtles have the D form of the DE gene.

Out of the **70** Bayosian turtles, **29** turtles have the L form of the LM gene.
Out of the **30** Freqosian turtles, **28** turtles have the L form of the LM gene.

Natural Frequency with Complement

Out of the **100** turtles on the island, **70** are Bayosian turtles and **30** are Freqosian turtles.

Out of the **70** Bayosian turtles, **2** turtles have the D form of the DE gene, and **68** turtles have the E form of the gene.
Out of the **30** Freqosian turtles, **9** turtles have the D form of the DE gene, and **21** turtles have the E form of the gene.

Out of the **70** Bayosian turtles, **29** turtles have the L form of the LM gene, and **41** turtles have the M form of the gene.
Out of the **30** Freqosian turtles, **28** turtles have the L form of the LM gene, and **2** turtles have the M form of the gene.

Posterior Frequency

Out of the **100** turtles on the island, **11** have the D form of the DE gene.
Out of the **100** turtles on the island, **57** have the L form of the LM gene.

Out of the **11** turtles with the D form of the DE gene, **2** are Bayosian turtles.
Out of the **89** turtles with the E form of the DE gene, **68** are Bayosian turtles.

Out of the **57** turtles with the L form of the LM gene, **29** are Bayosian turtles.
Out of the **43** turtles with the M form of the LM gene, **41** are Bayosian turtles.

Posterior Frequency with Complement

Out of the **100** turtles on the island, **11** have the D form of the DE gene, and **89** turtles have the E form of the gene.
Out of the **100** turtles on the island, **57** have the L form of the LM gene, and **43** turtles have the M form of the gene.

Out of the **11** turtles with the D form of the DE gene, **2** are Bayosian turtles and **9** are Freqosian turtles.
Out of the **89** turtles with the E form of the DE gene, **68** are Bayosian turtles and **21** are Freqosian turtles.

Out of the **57** turtles with the L form of the LM gene, **29** are Bayosian turtles and **28** are Freqosian turtles.
Out of the **43** turtles with the M form of the LM gene, **41** are Bayosian turtles and **2** are Freqosian turtles.

Icon Array

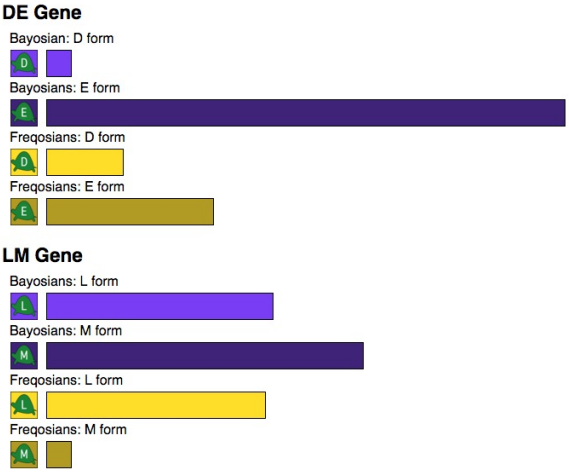
Below you see a graphical representation of the distribution of the DE gene and the LM gene in Bayosian and Freqosian turtles. Each icon represents **1** out of the total **100** turtles on the island.

Posterior Icon Array

Below you see a graphical representation of the distribution of Bayosian and Freqosian turtles for the DE gene and the LM gene. Each icon represents **1** out of the total **100** turtles on the island.

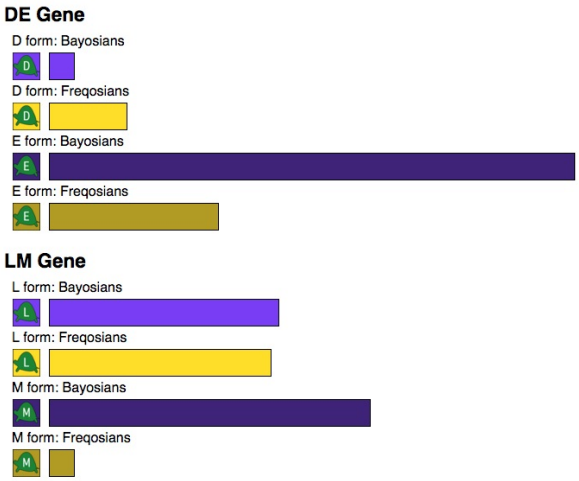
Bar Graph

Below you see a graphical representation of the distribution of the DE gene and the LM gene in Bayosian and Freqosian turtles. The length of each bar (relative to the span of the white space on the page) indicates the proportion of the total **100** turtles on the island.



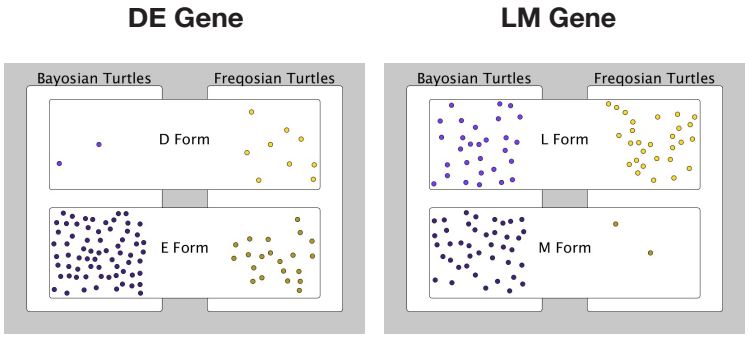
Posterior Bar Graph

Below you see a graphical representation of the distribution of Bayosian and Freqosian turtles for the DE gene and the LM gene. The length of each bar (relative to the span of the white space on the page) indicates the proportion of the total **100** turtles on the island.



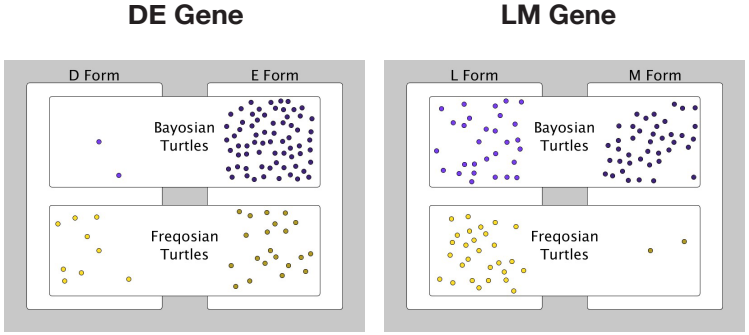
Dot Diagram

Below you see a graphical representation of the distribution of the DE gene and the LM gene in Bayosian and Freqosian turtles. Each dot represents **1** out of the total **100** turtles on the island.



Posterior Dot Diagram

Below you see a graphical representation of the distribution of Bayosian and Freqosian turtles for the DE gene and the LM gene. Each dot represents **1** out of the total **100** turtles on the island.



Note. Likelihood formats are shown as empty boxes, whereas posterior formats are shown as filled in boxes. "+" denotes that a format includes the complement information.

Experiment 3

Table B3: Each Type of Format used in Experiment 3

Format	Example
 Standard Probability	Consider a turtle picked at random from the **100** turtles on the island: The probability that it is a Bayosian turtle is **70%**. If a turtle is a Bayosian, then the probability that it has the D form of the DE gene is **4%**. If a turtle is a Freqosian, then the probability that it has the D form of the DE gene is **37%**. If a turtle is a Bayosian, then the probability that it has the L form of the LM gene is **43%**. If a turtle is a Freqosian, then the probability that it has the L form of the LM gene is **100%**.
 Probability with Complement	Consider a turtle picked at random from the **100** turtles on the island: The probability that it is a Bayosian turtle is **70%**, and the probability that it is a Freqosian turtle is **30%**. If a turtle is a Bayosian, then the probability that it has the D form of the DE gene is **4%**, and the probability that it has the E form is **96%**. If a turtle is a Freqosian, then the probability that it has the D form of the DE gene is **37%**, and the probability that it has the E form is **63%**. If a turtle is a Bayosian, then the probability that it has the L form of the LM gene is **43%**, and the probability that it has the M form is **57%**. If a turtle is a Freqosian, then the probability that it has the L form of the LM gene is **100%**, and the probability that it has the M form is **0%**.
 Posterior Probability	Consider a turtle picked at random from the **100** turtles on the island: The probability that it has the D form of the DE gene is **14%**. The probability that it has the L form of the LM gene is **60%**. If a turtle has the D form of the DE gene, then the probability that it is a Bayosian turtle is **20%**. If a turtle has the E form of the DE gene, then the probability that it is a Bayosian turtle is **78%**. If a turtle has the L form of the LM gene, then the probability that it is a Bayosian turtle is **50%**. If a turtle has the M form of the LM gene, then the probability that it is a Bayosian turtle is **100%**.

Posterior Probability with Complement

Consider a turtle picked at random from the **100** turtles on the island:
The probability that it has the D form of the DE gene is **14%**, and the probability that it has the E form is **86%**.
The probability that it has the L form of the LM gene is **60%**, and the probability that it has the M form is **40%**.

If a turtle has the D form of the DE gene, then the probability that it is a Bayosian turtle is **20%**, and the probability that it is a Freqosian turtle is **80%**.
If a turtle has the E form of the DE gene, then the probability that it is a Bayosian turtle is **78%**, and the probability that it is a Freqosian turtle is **22%**.

If a turtle has the L form of the LM gene, then the probability that it is a Bayosian turtle is **50%**, and the probability that it is a Freqosian turtle is **50%**.
If a turtle has the M form of the LM gene, then the probability that it is a Bayosian turtle is **100%**, and the probability that it is a Freqosian turtle is **0%**.

Natural Frequency

Out of the **100** turtles on the island, **70** are Bayosian turtles.

Out of the **70** Bayosian turtles, **3** turtles have the D form of the DE gene.
Out of the **30** Freqosian turtles, **11** turtles have the D form of the DE gene.

Out of the **70** Bayosian turtles, **30** turtles have the L form of the LM gene.
Out of the **30** Freqosian turtles, **30** turtles have the L form of the LM gene.

Natural Frequency with Complement

Out of the **100** turtles on the island, **70** are Bayosian turtles and **30** are Freqosian turtles.

Out of the **70** Bayosian turtles, **3** turtles have the D form of the DE gene, and **67** turtles have the E form of the gene.
Out of the **30** Freqosian turtles, **11** turtles have the D form of the DE gene, and **19** turtles have the E form of the gene.

Out of the **70** Bayosian turtles, **30** turtles have the L form of the LM gene, and **40** turtles have the M form of the gene.
Out of the **30** Freqosian turtles, **30** turtles have the L form of the LM gene, and **0** turtles have the M form of the gene.

Posterior Frequency

Out of the **100** turtles on the island, **14** have the D form of the DE gene.
Out of the **100** turtles on the island, **60** have the L form of the LM gene.

Out of the **14** turtles with the D form of the DE gene, **3** are Bayosian turtles.
Out of the **86** turtles with the E form of the DE gene, **67** are Bayosian turtles.

Out of the **60** turtles with the L form of the LM gene, **30** are Bayosian turtles.
Out of the **40** turtles with the M form of the LM gene, **40** are Bayosian turtles.

Posterior Frequency with Complement

Out of the **100** turtles on the island, **14** have the D form of the DE gene, and **86** turtles have the E form of the gene.
Out of the **100** turtles on the island, **60** have the L form of the LM gene, and **40** turtles have the M form of the gene.

Out of the **14** turtles with the D form of the DE gene, **3** are Bayosian turtles and **11** are Freqosian turtles.
Out of the **86** turtles with the E form of the DE gene, **67** are Bayosian turtles and **19** are Freqosian turtles.

Out of the **60** turtles with the L form of the LM gene, **30** are Bayosian turtles and **30** are Freqosian turtles.
Out of the **40** turtles with the M form of the LM gene, **40** are Bayosian turtles and **0** are Freqosian turtles.

Icon Array

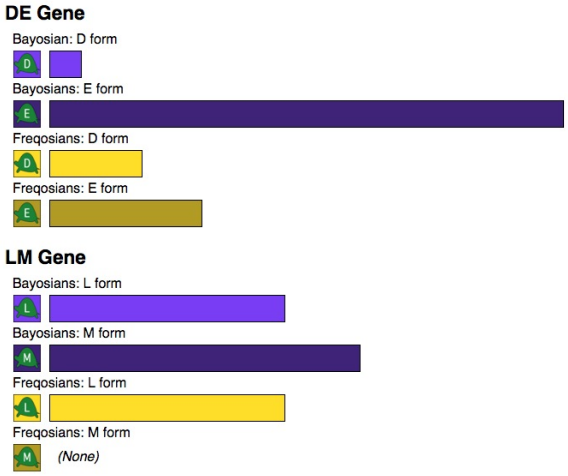
Below you see a graphical representation of the distribution of the DE gene and the LM gene in Bayosian and Freqosian turtles. Each icon represents **1** out of the total **100** turtles on the island.

Posterior Icon Array

Below you see a graphical representation of the distribution of Bayosian and Freqosian turtles for the DE gene and the LM gene. Each icon represents **1** out of the total **100** turtles on the island.

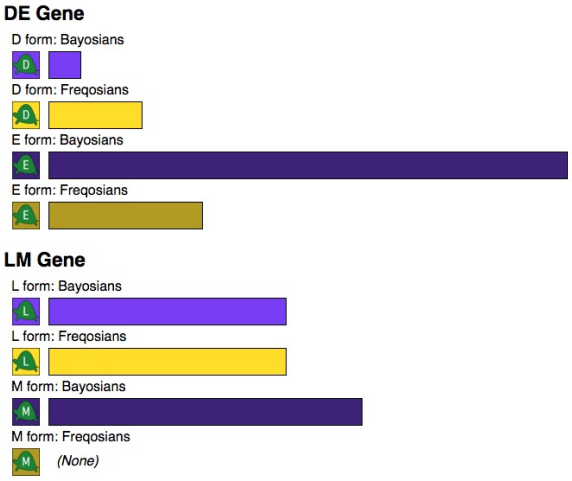
Bar Graph

Below you see a graphical representation of the distribution of the DE gene and the LM gene in Bayosian and Freqosian turtles. The length of each bar (relative to the span of the white space on the page) indicates the proportion of the total **100** turtles on the island.



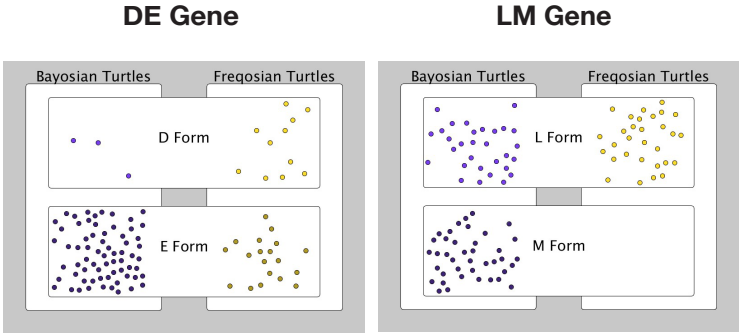
Posterior Bar Graph

Below you see a graphical representation of the distribution of Bayosian and Freqosian turtles for the DE gene and the LM gene. The length of each bar (relative to the span of the white space on the page) indicates the proportion of the total **100** turtles on the island.



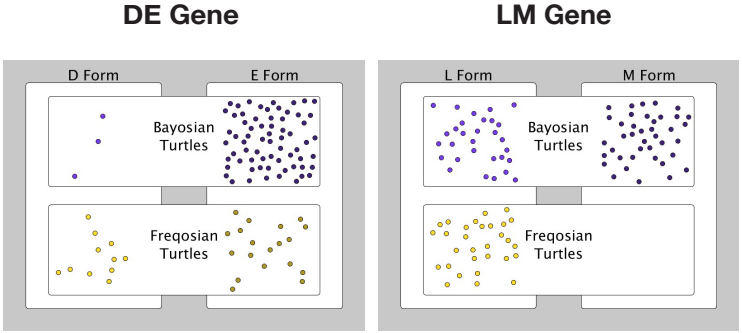
Dot Diagram

Below you see a graphical representation of the distribution of the DE gene and the LM gene in Bayosian and Freqosian turtles. Each dot represents **1** out of the total **100** turtles on the island.



Posterior Dot Diagram

Below you see a graphical representation of the distribution of Bayosian and Freqosian turtles for the DE gene and the LM gene. Each dot represents **1** out of the total **100** turtles on the island.



Note. Likelihood formats are shown as empty boxes, whereas posterior formats are shown as filled in boxes. "+" denotes that a format includes the complement information.

Experiment 4

Table B4: Each Type of Format used in Experiment 4

Format	Example
 Standard Probability	Consider a turtle picked at random from the **100** turtles on the island: The probability that it is a Bayosian turtle is **72%**. If a turtle is a Bayosian, then the probability that it has the D form of the DE gene is **3%**. If a turtle is a Freqosian, then the probability that it has the D form of the DE gene is **83%**. If a turtle is a Bayosian, then the probability that it has the L form of the LM gene is **39%**. If a turtle is a Freqosian, then the probability that it has the L form of the LM gene is **100%**.
 Probability with Complement	Consider a turtle picked at random from the **100** turtles on the island: The probability that it is a Bayosian turtle is **72%**, and the probability that it is a Freqosian turtle is **28%**. If a turtle is a Bayosian, then the probability that it has the D form of the DE gene is **3%**, and the probability that it has the E form is **97%**. If a turtle is a Freqosian, then the probability that it has the D form of the DE gene is **83%**, and the probability that it has the E form is **17%**. If a turtle is a Bayosian, then the probability that it has the L form of the LM gene is **39%**, and the probability that it has the M form is **61%**. If a turtle is a Freqosian, then the probability that it has the L form of the LM gene is **100%**, and the probability that it has the M form is **0%**.
 Posterior Probability	Consider a turtle picked at random from the **100** turtles on the island: The probability that it has the D form of the DE gene is **25%**. The probability that it has the L form of the LM gene is **56%**. If a turtle has the D form of the DE gene, then the probability that it is a Bayosian turtle is **9%**. If a turtle has the E form of the DE gene, then the probability that it is a Bayosian turtle is **94%**. If a turtle has the L form of the LM gene, then the probability that it is a Bayosian turtle is **50%**. If a turtle has the M form of the LM gene, then the probability that it is a Bayosian turtle is **100%**.

Posterior Probability with Complement

Consider a turtle picked at random from the **100** turtles on the island:
The probability that it has the D form of the DE gene is **25%**, and the probability that it has the E form is **75%**.
The probability that it has the L form of the LM gene is **56%**, and the probability that it has the M form is **44%**.

If a turtle has the D form of the DE gene, then the probability that it is a Bayosian turtle is **9%**, and the probability that it is a Freqosian turtle is **91%**.
If a turtle has the E form of the DE gene, then the probability that it is a Bayosian turtle is **94%**, and the probability that it is a Freqosian turtle is **6%**.

If a turtle has the L form of the LM gene, then the probability that it is a Bayosian turtle is **50%**, and the probability that it is a Freqosian turtle is **50%**.
If a turtle has the M form of the LM gene, then the probability that it is a Bayosian turtle is **100%**, and the probability that it is a Freqosian turtle is **0%**.

Natural Frequency

Out of the **100** turtles on the island, **72** are Bayosian turtles.

Out of the **72** Bayosian turtles, **2** turtles have the D form of the DE gene.
Out of the **28** Freqosian turtles, **23** turtles have the D form of the DE gene.

Out of the **72** Bayosian turtles, **28** turtles have the L form of the LM gene.
Out of the **28** Freqosian turtles, **28** turtles have the L form of the LM gene.

Natural Frequency with Complement

Out of the **100** turtles on the island, **72** are Bayosian turtles and **28** are Freqosian turtles.

Out of the **72** Bayosian turtles, **2** turtles have the D form of the DE gene, and **70** turtles have the E form of the gene.
Out of the **28** Freqosian turtles, **23** turtles have the D form of the DE gene, and **5** turtles have the E form of the gene.

Out of the **72** Bayosian turtles, **28** turtles have the L form of the LM gene, and **44** turtles have the M form of the gene.
Out of the **28** Freqosian turtles, **28** turtles have the L form of the LM gene, and **0** turtles have the M form of the gene.

Posterior Frequency

Out of the **100** turtles on the island, **25** have the D form of the DE gene.
Out of the **100** turtles on the island, **56** have the L form of the LM gene.

Out of the **25** turtles with the D form of the DE gene, **2** are Bayosian turtles.
Out of the **75** turtles with the E form of the DE gene, **70** are Bayosian turtles.

Out of the **56** turtles with the L form of the LM gene, **28** are Bayosian turtles.
Out of the **44** turtles with the M form of the LM gene, **44** are Bayosian turtles.

Posterior Frequency with Complement

Out of the **100** turtles on the island, **25** have the D form of the DE gene, and **75** turtles have the E form of the gene.
Out of the **100** turtles on the island, **56** have the L form of the LM gene, and **44** turtles have the M form of the gene.

Out of the **25** turtles with the D form of the DE gene, **2** are Bayosian turtles and **23** are Freqosian turtles.
Out of the **75** turtles with the E form of the DE gene, **70** are Bayosian turtles and **5** are Freqosian turtles.

Out of the **56** turtles with the L form of the LM gene, **28** are Bayosian turtles and **28** are Freqosian turtles.
Out of the **44** turtles with the M form of the LM gene, **44** are Bayosian turtles and **0** are Freqosian turtles.

Icon Array

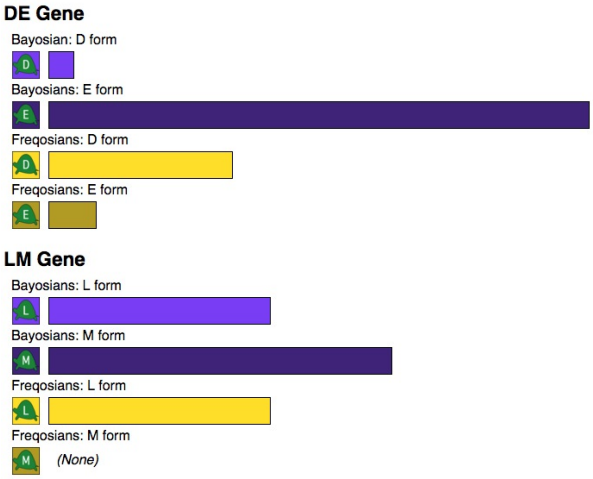
Below you see a graphical representation of the distribution of the DE gene and the LM gene in Bayosian and Freqosian turtles. Each icon represents **1** out of the total **100** turtles on the island.

Posterior Icon Array

Below you see a graphical representation of the distribution of Bayosian and Freqosian turtles for the DE gene and the LM gene. Each icon represents **1** out of the total **100** turtles on the island.

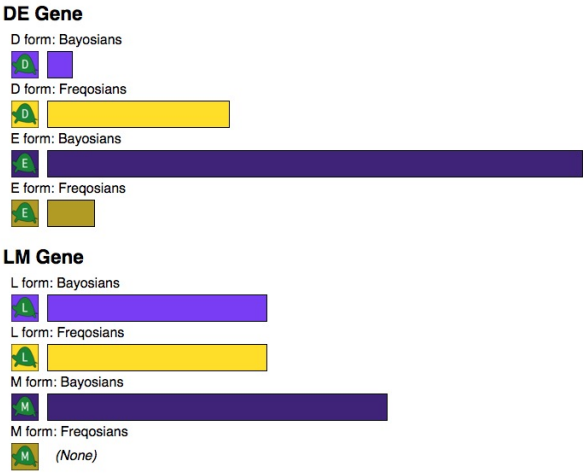
Bar Graph

Below you see a graphical representation of the distribution of the DE gene and the LM gene in Bayosian and Freqosian turtles. The length of each bar (relative to the span of the white space on the page) indicates the proportion of the total **100** turtles on the island.



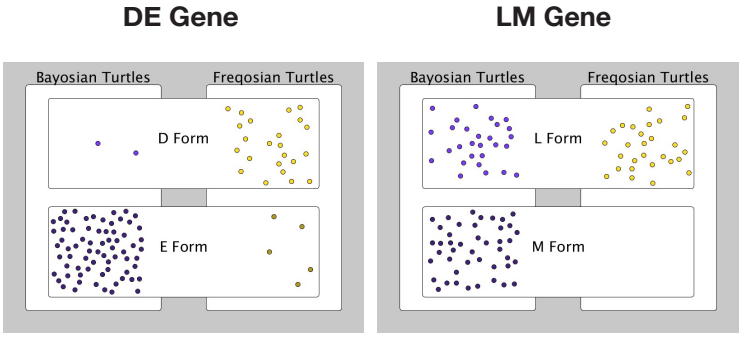
Posterior Bar Graph

Below you see a graphical representation of the distribution of Bayosian and Freqosian turtles for the DE gene and the LM gene. The length of each bar (relative to the span of the white space on the page) indicates the proportion of the total **100** turtles on the island.



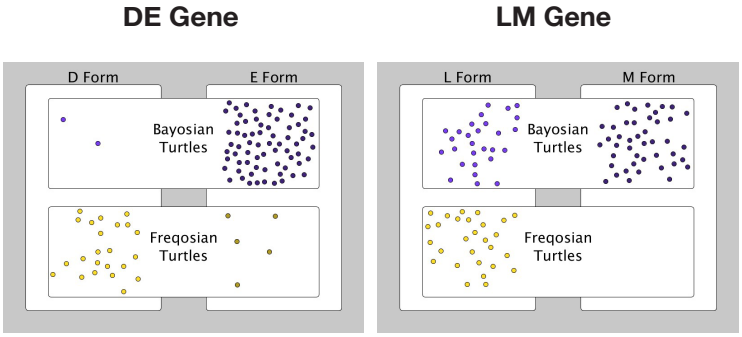
Dot Diagram

Below you see a graphical representation of the distribution of the DE gene and the LM gene in Bayosian and Freqosian turtles. Each dot represents **1** out of the total **100** turtles on the island.



Posterior Dot Diagram

Below you see a graphical representation of the distribution of Bayosian and Freqosian turtles for the DE gene and the LM gene. Each dot represents **1** out of the total **100** turtles on the island.



Note. Likelihood formats are shown as empty boxes, whereas posterior formats are shown as filled in boxes. "+" denotes that a format includes the complement information.

Appendix C: Screenshots

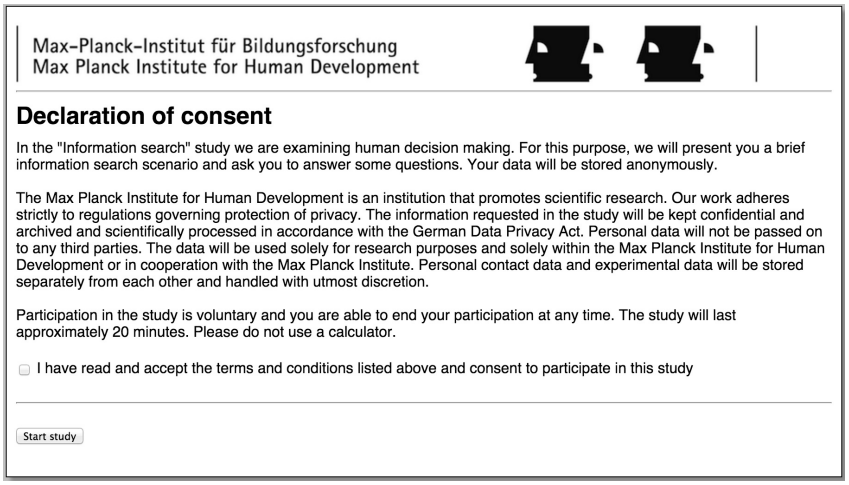


FIGURE C1: The consent form provided to subjects before starting the study.

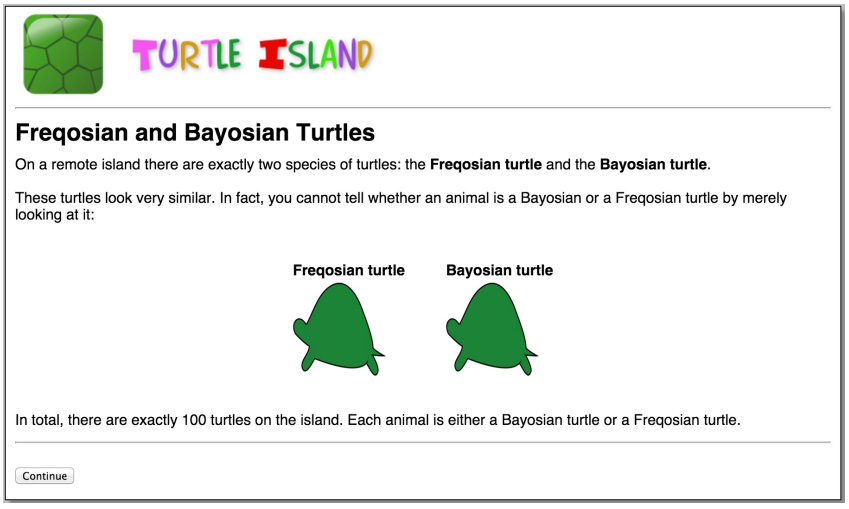


FIGURE C2: Introduction to the experiment.

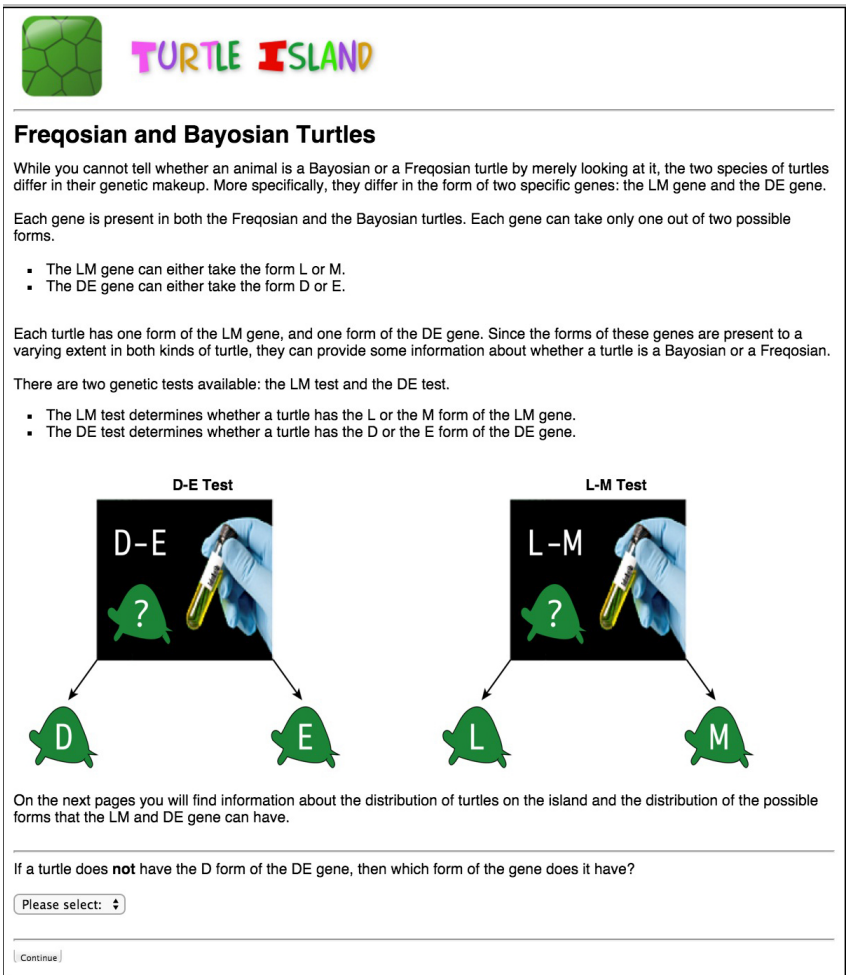
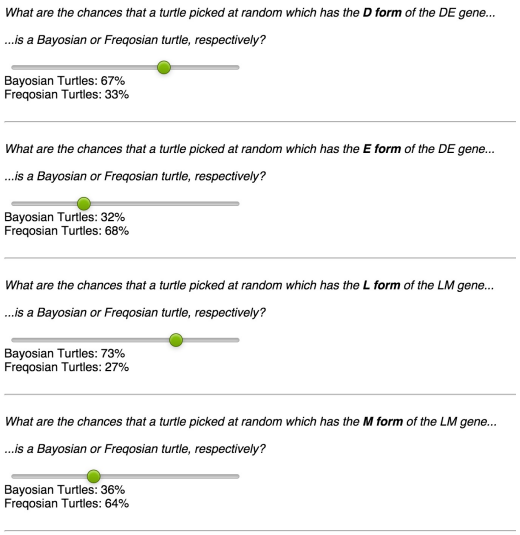


Figure C3: Information about the DNA tests.

Appendix D: Probability Judgment Questions

Table D1: All Probability Judgment Questions

Block	Type	Question
1	Base Rate and Marginals	*What are the chances that a turtle picked at random from the island...* *... is a Bayosian or Freqosian turtle, respectively?* Bayosian Turtles: 36% Freqosian Turtles: 64% *...has the D or E form, respectively, of the DE gene?* D form: 72% E form: 28% *...has the L or M form, respectively, of the LM gene?* L form: 37% M form: 63%
2	Likelihoods	*What are the chances that a **Bayosian** turtle...* *...has the D or E form, respectively, of the DE gene?* D form: 75% E form: 25% *... has the L or M form, respectively, of the LM gene?* L form: 29% M form: 71% *What are the chances that a **Freqosian** turtle...* *...has the D or E form, respectively, of the DE gene?* D form: 74% E form: 26% *... has the L or M form, respectively, of the LM gene?* L form: 31% M form: 69%

3	Posteriors	

Note. Sliders are shown at random values to illustrate how the text below dynamically updates with both complementary values. Each block of questions was presented to participants in random order.

Appendix E: Numeracy Test

Table E1: Numeracy Questions

Test	#	Question
Schwartz et al.	1	Imagine that we flip a fair coin 1,000 times. What is your best guess about how many times the coin would come up heads in 1,000 flips?
	2	In the BIG BUCKS LOTTERY, the chance of winning a $10 prize is 1%. What is your best guess about how many people would win a $10 prize if 1000 people each buy a single ticket to BIG BUCKS?
	3	In ACME PUBLISHING SWEEPSTAKES, the chance of winning a car is 1 in 1,000. What percent of tickets to ACME PUBLISHING SWEEPSTAKES win a car?
Berlin Numeracy Test	1	Out of 1,000 people in a small town 500 are members of a choir. Out of these 500 members in the choir 100 are men. Out of the 500 inhabitants that are not in the choir 300 are men. What is the probability that a randomly drawn man is a member of the choir?
	2a	Imagine we are throwing a five-sided die 50 times. On average, out of these 50 throws how many times would this five-sided die show an odd number (1, 3 or 5)?
	2b	Imagine we are throwing a loaded die (6 sides). The probability that the die shows a 6 is twice as high as the probability of each of the other numbers. On average, out of 70 throws how many times would the die show the number 6?
	4	In a forest 20% of mushrooms are red, 50% brown and 30% white. A red mushroom is poisonous with a probability of 20%. A mushroom that is not red is poisonous with a probability of 5%. What is the probability that a poisonous mushroom in the forest is red?

Note. The Berlin Numeracy test is an adaptive test. See Fig E1 below. The Schwartz et al. test yields a score between 0 and 3, while the Berlin Numeracy Test gives an adaptive score between 1 and 4. The total score of the combined numeracy test ranges between 1-7.

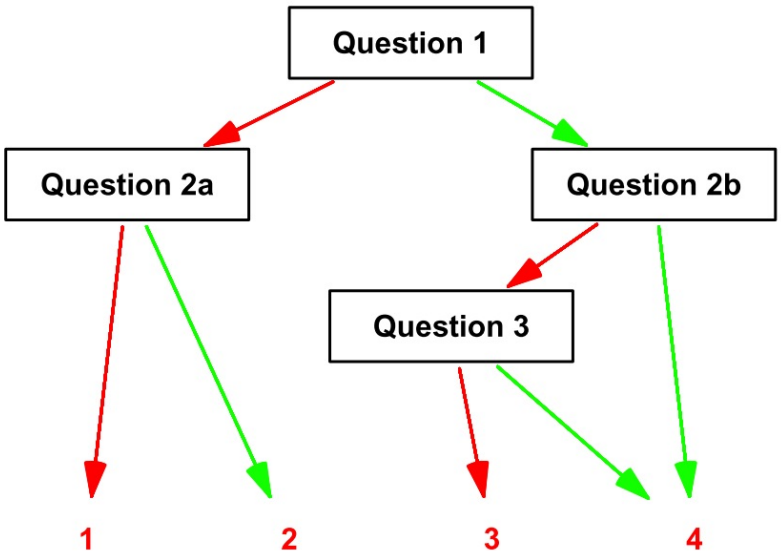


Figure E1: The adaptive design of the Berlin Numeracy test from Cokely et al. (2012). Red arrows denote a wrong answer, while green arrows denote a correct answer. The final scores are shown at the bottom.

Index

Charley Mingshuo Wu

Curriculum Vitae

General Information:

A. Personal Data:

Email: charleymswu [at] gmail.com
Phone: +49 151 1000 3753
Address: Roseggerstr. 38, 12059 Berlin, Germany
Citizenship: Canadian
Languages: English, Chinese (Mandarin, Shanghainese, Classical Chinese), and some German, French, and Spanish

B. Areas:

Areas of Specialization: Information Search, Mathematical Models of Uncertainty, Judgment and Decision Making, Natural Language Processing, Machine Learning, and Semantic Embedding
Areas of Competence: Neural Networks, Cognitive Linguistics, Ecological Rationality, Neuroscience, Philosophy of Mind, and Evolutionary Psychology
Areas of Interest: Learning of Structured Environments, Learning and Application of Heuristics, and Ecologically Valid Theories of Language Acquisition

C. Current Position:

Masters Student, MEi: Cognitive Science, Universität Wien (2013 - Present)
Guest Student Researcher, Center for Adaptive Behavior and Cognition, Max Planck Institute for Human Development, Berlin (2015)

D. Previous Positions:

Student Research Assistant, Center for Adaptive Behavior and Cognition, Max Planck Institute for Human Development, Berlin (2014)
Natural Language Processing and Machine Learning Consultant, Sportsmanias, USA (2014 - 2015)
Student Researcher, OFAI (Austrian Institute for Artificial Intelligence), Intelligent Software Agents and New Media Group (2013 - 2014)
Owner, Skylantern Recording Studios, Vancouver (2011 - 2013)
Business Intelligence Analyst, Teligence (2010 - 2013)
Music Director, Harbourfront Centre, Toronto (2010)
Research Assistant, Center for Human Evolution, Cognition, and Culture (HECC), University of British Columbia (2009)

Academic Information:

E. Education:

2015 (Expected) - M.Sc. Cognitive Science, Universität Wien
2009 – B.A. in Philosophy, University of British Columbia (Dean's List)

F. Grants, Honours, and Awards:

· Interdisciplinary College (IK) Stipend (2014)
· Joseph-Armand Bombardier Canada Graduate Scholarship, Social Sciences and Humanities Research Council of Canada (SSHRC), declined due to personal reasons (2011-2012)
· President's Entrance Scholarship, University of British Columbia (2004)

G. Research:

· Research on Information Search supported by the DFG grant "Frameworks of Rationality", the Center for Adaptive Behavior and Cognition, Max Planck Institute for Human Development, Berlin (2014 - 2015)
· Semantic Embedding and Neural Networks for Classification of News Content, Sportsmanias (2014 - 2015)
· Coupling Latent Dirichlet Allocation with expert knowledge for hierarchical topic generation from academic abstracts, University of Vienna Visualization and Data Analysis research group (2014)
· Creation of a Compositional Vector Space Model of Language using Semantic Role Labeling with (OFAI), the Austrian Research Institute for Artificial Intelligence (2013-2014)
· Contributing member of the Centre for Human Evolution, Cognition, and Culture (HECC), under the direction of Joseph Henrich, UBC (2009)
· Research Assistant, Oxford Project, Edward Slingerland, UBC (2009)

H. Conferences:

· SPUDM, Budapest (2015, Paper Presentation)
· MEi: CogSci Conference, Ljubljana (2014, Talk)
· NIPS, Semantics Workshop, Montreal (2014)
· MEi: CogSci Conference, Krakow (2014, Poster)
· Interdisciplinary College (IK), Günne (2014, Poster)
· International Conference on Cognitive Modeling (ICCM), Berlin (2012)

I. Additional Abilities and Interests:

· Fluency in Python, Octave, Matlab, R and Java, specifically in tasks related to Data Analysis, Visualization, Machine Learning, Natural Language Processing, Neural Network GPU learning, Algorithm Optimization, Social Network Analysis, Cross-server communication, and Data Modeling.

· Programming of human psychological experiments on the web using Amazon's MTurk platform, implemented with Javascript, HTML, CSS, jQuery, AJAX, and PHP

· 4-5 years of experience with database architectures, including MongoDB, MySQL, and Redis (memcache)

· Classically trained Pianist (Royal Conservatory of Music) and multi-instrumentalist (Guitar, Bass, Drums, Mandolin, Accordion, and Saxophone)

· Independent musician, sound engineer, and recording artist (5 published albums) with experience booking tours across Canada and the USA

J. Volunteer Experience:

· Vancouver Food Bank (2013)
· Habitat For Humanity (2012)
· Public Dreams Foundation, Vancouver (2009-2012)
· Megaphone Magazine, Vancouver's Street Paper (2012)
· Media Democracy Days, Vancouver (2011)
· PALS Autism School, Vancouver (2010)
· Food Not Bombs, Vancouver (2008-2009)