A game-based intervention for the reduction of statistical cognitive biases

by

Cassandra Xia

S.B. Computer Science and Engineering, MIT 2012

Submitted to the Program in Media Arts and Sciences,
School of Architecture and Planning,
in Partial Fulfillment of the Requirements for the Degree of
Master of Science in the Media Arts and Sciences
at the
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
June 2014

© 2014 Massachusetts Institute of Technology. All rights reserved.

Signature of Author

Cassandra Xia
May 9, 2014

Certified by

Pattie Maes
Thesis Supervisor
Professor of Media Arts and Sciences

Accepted by

Pattie Maes
Associate Academic Head
Program in Media Arts and Sciences
Abstract
Probability and statistics is perhaps the area of mathematics education most directly applicable to everyday life. Yet, the methodologies traditionally used to cover these topics in school render the material formal and difficult to apply. In this thesis, I describe a game design that develops probabilistic concepts in real-life situations.

Psychologists have coined the term cognitive bias for instances in which the intuition of the average person disagrees with the formal mathematical analysis of the problem. This thesis examines if a one-hour game-based intervention can enact a change in the intuitive mental models people have for reasoning about probability and uncertainty in real-life. Two cognitive biases were selected for treatment: overconfidence effect and base rate neglect. These two biases represent instances of miscalibrated subjective probabilities and Bayesian inference, respectively.

Results of user tests suggest that it is possible to alter probabilistic intuitions, but that attention to the transitions from the current mental constructs must be carefully designed. Prototyping results suggest how some elements of game design may naturally lend themselves to deep learning objectives and heuristics.

Thesis supervisor: Pattie Maes
Title: Professor of Media Arts and Sciences, Fluid Interfaces Group, MIT Media Lab
A game-based intervention for the reduction of statistical cognitive biases

Cassandra Xia

Thesis Advisor:

________________________________________

Pattie Maes
Professor of Media Arts and Sciences
MIT Media Lab

Thesis Reader:

________________________________________

Mitch Resnick
Professor of Media Arts and Sciences
MIT Media Lab

Thesis Reader:

________________________________________

Dor Abrahamson
Professor of Cognition and Development
UC Berkeley Graduate School of Education
Acknowledgements

To Pattie Maes, who first introduced Engelbart’s 1962 paper on intelligence augmentation with a phrase I will always remember: “not AI but IA.” I would like to thank her for her remarkable ability to match students up to the ideas they need, as well as her wisdom, energy, and steady patience as I bounced from new idea to the next.

To my readers, Dor Abrahamson, who was the best mentor I could wish for in his patient and enthusiastic guidance through the math education space, and Mitch Resnick, whose uncannily accurate instinctual reactions should be included within a study of intuition itself. And to Linda Peterson, the most effective principal I have ever encountered.

To two world-distorting classes at MIT and the professors behind them: 9.660 and Josh Tenenbaum, for introducing the power and beauty of human intuition, and 6.437 and Lizhong Zheng, for motivationally teaching material so difficult that symbol-bashing no longer works and true intuition becomes necessary.

Thank you to my family and friends who served as dispensers of wisdom, good sense, and sanity: Tal Achituv, Ned Burnell, Sujay Chowdhury, Niaja Farve, Yasemin Gokce, Pragun Goyal, Scott Greenwald, Mark Hou, Sonia Jin, Erika Lee, Cory Li, Paul Medlock-Walton, Kelly Ran, Eric Shyu, Jacob Steinhardt, James Tao, Anette von Kapri, Jack Wang, Victoria Xia, Sophia Youn, Sherwin Yu, and Jan Zilinsky. To my mother and father, who always magically pick up the phone when I call. And to Bret Victor, whose belief in the human potential and methods for making tomorrow better are incredibly infectious.
# Table of Contents

**Introduction** .................................................................................................................. 6

Intuition .......................................................................................................................... 6

Intuition and immediacy .............................................................................................. 6

Intuition and vision ....................................................................................................... 7

Designing for intuition ................................................................................................. 7

Cognitive biases ........................................................................................................... 8

Debiasing ....................................................................................................................... 8

Overconfidence effect ................................................................................................. 9

Bayesian conservatism ............................................................................................... 10

Base rate neglect ........................................................................................................ 11

Prior work in statistics education research .................................................................. 12

**Discarded prototypes** ............................................................................................... 13

Discarded prototype #1: Bayes Eyes ........................................................................... 14

Problems associated with Bayes Eyes ....................................................................... 15

Discarded prototype #2: A Random Life .................................................................... 15

Module: Expected Value ............................................................................................. 16

Module: Bayesian Inference ......................................................................................... 18

Module: Hypothesis Testing ......................................................................................... 19

Problems associated with A Random Life .................................................................. 21

**Final prototype: Adventures in Cognitive Biases** ....................................................... 21

Design of Adventures in Cognitive Biases ................................................................. 23

Designing for the Overconfidence Bias .................................................................... 23

Designing for Intuitive Bayes ..................................................................................... 26

**Evaluation of final prototype** ..................................................................................... 32

Pretest results ............................................................................................................... 32

Posttest results ............................................................................................................. 35

Analysis ........................................................................................................................ 42

**Future Work** ............................................................................................................... 45

**Reflection**—On the biases of a novice designer ..................................................... 46

**Conclusion** ................................................................................................................ 47

**References** ................................................................................................................ 48
Introduction

In this thesis, I will describe my attempts to modify the way that people reason probabilistically. It might seem that since probability is directly applicable to everyday life, learners of probability would readily apply new ideas to out-of-classroom contexts. However, developing these probabilistic intuitions seems to be more difficult since learners arrive with strongly held preexisting beliefs.

Decades of psychology research have identified instances in which naïve probabilistic intuitions can cause us to reach conclusions that differ from the normatively accepted answers. Psychologists attribute the use of quick-use heuristics for the discrepancy between intuition and formal answers. If we wish to evolve primary intuitions into secondary intuitions that are more in line with the formal model, learners must achieve fluency with alternative methods of thought that provide the same level of immediacy. To this end, this thesis explores the use of diagrammatic models which offload some processing to the visual system, rendering solutions to probabilistic questions immediate by “seeing” the answer. This thesis combines research on intuition with game design elements to achieve changes in the models that learners use for probabilistic intuition.

The introductory material will begin by formalizing the definition of intuition, explore some of the ways in which intuitions can be developed, and finally explore some of the debiasing methods that have been attempted for correcting misguided intuitions. Synthesizing these prior findings, this thesis will describe a final game prototype developed by the author to teach probability and statistics in a more intuitive manner, as well as two unsuccessful prototypes developed during the design process.

Intuition

Intuition and immediacy

To clarify the definition of “intuition,” we turn to Fischbein’s comprehensive review on the subject, Intuition in Science and Mathematics (Fischbein, 1987). Fischbein observes that the word has been lauded by some as “the highest form of knowledge” while denigrated by others as “a global guess for which an individual is not able to offer a clear and complete justification.”

The unifying aspect that reconciles these two disparate concepts of intuition is the aspect of immediacy. Fischbein writes, “intuitive knowledge is immediate knowledge; that is, a form of cognition which seems to present itself to a person as being self-evident.” It is this aspect of immediate clarity that this research takes as a working definition of intuition.

From immediacy arises premature certainty. This level of certainty is posited by Fischbein to be requisite to journey down unknown paths when solving difficult problems. However, there are examples in which the immediate conclusions arising from primary intuitions – the intuitions we develop in everyday life – are wrong according to normative models. In these instances, it is common to discuss the development of secondary intuitions: the intuitions that arise after formal training in the specific domain that agree with normative models.

In this thesis, we are primarily concerned with problem solving intuitions, which (Fischbein, 1987) categorizes as either anticipatory or conclusive. Anticipatory intuitions help to anticipate a solution to the problem. Conclusive intuitions are the intuitions built after a problem has been solved. Fischbein defines conclusive intuitions to “summarize in a global, structured vision the basic ideas of the solution to a
problem,” giving an example from French mathematician Hadamard who describes it as “I do not feel like I have understood it as long as I do not succeed in grasping it in one global idea.” This type of intuition suggests an after-the-fact integrative consolidation of mental structures. Mathematician George Polya notes that the best students of mathematics tend to mull over a problem after it is solved and analyze it from different perspectives (Polya, 1973). Consolidating knowledge in this way may render the solution even more immediately obvious upon future encounters with similar problems.

**Intuition and vision**

A working definition of immediacy suggests that one important route towards intuition may be through visual mental models. The visual modality allows for flashes of insight that can guide the problem solving process by providing intermediate landmarks or by suggesting approaches via analogy to previously encountered problems. (Fischbein, 1987) elaborates on this notion by writing that “a visual image not only organizes the data at hand in meaningful structures but it is also an important factor guiding the analytical development of a solution; visual representations are an essential anticipatory device.”

In particular, one way to create intuition may be through diagrammatic models, which are “graphical representations of phenomena and relationships amongst them.” Fischbein gives examples of diagrammatic models as “Venn diagrams, tree diagrams, and histograms used for statistical representations.” Diagrammatic models may assist with the immediacy aspect of intuitions, rendering a visual picture that is quick to work with. This thesis uses diagrammatic models to enhance anticipatory intuitions. By familiarizing subjects with visual ways of thinking generically relevant to probabilities, it is hoped that subjects may begin to use these pictures to “see” the solution.

However, Fischbein also warns that the model needs to be properly assimilated before it can be used, writing that “a graph may become an intuitive device only after the system of conventions relating the original reality, the intervening conceptual system… and the graphical representation have been internalized and automatized.” Fischbein describes the research of Claude Janvier: a graph depicting the speed of a race car over time is intuitively interpreted by middle school students as the literal path of the racecar over time, even though the same students simultaneously describe it as a graph of the speed.

**Designing for intuition**

Research in the space of physics education lends additional insight into the process of designing for intuition. The formal models of Newtonian mechanics often contradict the pre-existing mental models that students acquire through interacting with the physical world. This suggests that it is important to design “bridging analogies” to help students’ primary intuitions transition over to formal representations (diSessa, 1988) (Clement, 1993).

In mathematics education, (Abrahamson, 2012) has advocated the creation of “auspicious conditions” that encourage the development of perceptual reasoning that agree with mathematical norms. One such tool is a physical scooper for taking samples of colored marbles from a bin. This takes learners from observed samples to visually thinking about the permutations of colored balls possible, which can subsequently be arranged into a combinatorial pyramid representing the frequencies of occurrence (Abrahamson, 2006).
To determine the degree to which intuitions have changed after physics instruction, (Hestenes, 1995) designed the Force Concept Inventory (FCI). The exam consists of 30 multiple choice questions to be completed in 30 minutes, describing real-world situations such as a ball rolling off a table. Rather than calculations, students are asked to make intuitive predictions about what would happen in the situation described. By design, the FCI is perceived as easy to all students, since among the multiple choice answers are predictions according to both naïve physics intuition and formal Newtonian physics. The design of the test allows for the diagnosis of misconceptions.

In the next section, I introduce the literature on cognitive biases as a useful tool to arouse misguided probabilistic intuitions. I seek realistic situations in which answers are immediately evident to all, but primary intuitions are often incorrect.

**Cognitive biases**

*Debiasing*

Research on cognitive biases has identified instances in which people primary intuitions are often incorrect, leading to instances in which people behave irrationally according to normative models (Tversky, 1974). The study of cognitive biases is also referred to as the heuristics and biases movement, as it is believed that people employ low effort heuristics which result in the observed bias. While most researchers believe that biases are detrimental and study methods for debiasing subjects, (Gigerenzer, 2009) argues that the use of biased heuristics may be optimal from the perspective of an accuracy-effort trade-off.

The detriment of cognitive biases is recognized by government organizations. In 2011, the Intelligence Advanced Research Projects Activity (IARPA) issued a request for proposals for a special program on debiasing, called the Sirius Program. The official statement from IARPA is:

“The goal of the Sirius Program is to create Serious Games to train participants and measure their proficiency in recognizing and mitigating the cognitive biases that commonly affect all types of intelligence analysis. The research objective is to experimentally manipulate variables in Virtual Learning Environments (VLE) to determine whether and how such variables might enable player-participant recognition and persistent mitigation of cognitive biases. The Program will provide a basis for experimental repeatability and independent validation of effects, and identify critical elements of design for effective analytic training in VLEs. The cognitive biases of interest that will be examined include: (1) Confirmation Bias, (2) Fundamental Attribution Error, (3) Bias Blind Spot, (4) Anchoring Bias, (5) Representativeness Bias, and (6) Projection Bias.”

At the time of this thesis, the Sirius Program was still in its infancy. One of the publications arising from this initiative is a discussion on how serious games can be used for critical thinking and debiasing (Flach, 2012). Some implementable suggestions from this panel discussion were that design in this area must include a description of the stages through which learners progress, a description of cognitive learning opportunities in each of these levels, as well as a theory of error to explain sources of bias.

Cognitive biases are often elicited through word problems discussing realistic inference scenarios. The biases have shown to be fairly consistently replicable even with variations to the problem, although there are exceptions that will be elaborated on in subsequent sections. Even mathematically inclined subjects
fall back on use of heuristics when asked to solve these realistically framed word problems, and these word problems may be an example of what (Prediger, 2008) describes as cognitive layers. (Prediger, 2008) hypothesizes that as students learn probability in school, they develop an additional cognitive layer to use in classroom settings, but students choose to fall back on their preexisting cognitive layer in out-of-school contexts. (Prediger, 2008) frames the development of learning arrangements such that probability is seen as a strategic tool inside and out of the classroom as a major design challenge.

The final prototype is designed to address the Overconfidence Effect (Mannes, 2013), the tendency to overestimate the accuracy of beliefs, as well as two biases associated with failure to intuitively apply Bayes Rule to problems as new evidence is given: Base Rate Neglect (Bar-Hillel, 1980) and Bayesian Conservatism (Wheeler, 1968). A background of the psychology research surrounding these biases is given to provide information on how the biases are measured as well as known methods of debiasing from which the final prototype takes inspiration.

Overconfidence Effect

The overconfidence effect is the tendency for people to be more confident about the correctness of their answers than their accuracy justifies. In the overconfidence literature, this is also referred to as overprecision (Mannes, 2013). The overconfidence effect has been demonstrated by asking a large number of subjects to indicate their level of driving skill relative to the other strangers in the room. 93% of U.S. drivers sampled believed themselves to be more skillful than the median driver; the effect was less pronounced, but still 69% for Swedish drivers (Svenson, 1981).

As summarized and critiqued by (Gigerenzer, 1991), one common method of eliciting the overconfidence effect involves providing subjects with binary questions such as this example:

Which city has more inhabitants?
(a) Hyderabad (b) Islamabad

After a subject has given an answer, the subject is asked to rate their confidence that their answer is correct:

How confident are you that your answer is correct?
(a) 50% (b) 60% (c) 70% (d) 80% (e) 90% (f) 100%

When subjects are asked to rate their confidence, they are being asked to assign a subjective probability that they got the answer right. The subject is considered to be well calibrated if over a number of these questions, their long-run accuracy matches their subjective probability on each question. For instance, subjects should average 60% questions right on the questions to which they assign 60% confidence (Mannes, 2013).

Most subjects, however, achieve an accuracy of much less than their indicated confidence. This effect is interpreted by many as evidence of the overconfidence effect. (Gigerenzer, 1991), however, has found that the overconfidence effect is absent when subjects are asked “How many of these 50 questions do you think you got right?” as opposed to their level of confidence after each individual question. Gigerenzer interpreted these results to call into question this method of measuring overconfidence, suggesting that humans are natural frequentists and should not be expected to assign calibrated subjective probabilities
for a single event. Although it seems possible that Gigerenzer’s results could also be evidence for the failure of subjects to conceptually understand subjective probabilities (i.e. the multitude of ways to interpret the meaning of “60%”). If this is the case, then it may be possible to achieve debiasing by helping learners associate a subjective probability with a different preexisting intuition, such as the idea of a fair bet as demonstrated in Figure 15.

(Mannes, 2013) summarizes generic methods and methods within management science to reduce the overconfidence effect. (Fischoff, 1982) informed subjects about the presence of the bias but it persisted. (Mannes, 2013) attempted to calibrate participants by providing feedback on the correctness of their estimates. Subjects who received feedback on their estimates reduced their level of confidence but still exhibited overconfidence; only those subjects who were manipulated into believing that their errors were 2.5 times greater than the truth were calibrated out of the bias.

(Haran, 2010) found that asking subjects to estimate the probabilities of several similar outcomes reduces overconfidence, possibly since subjects realize that there are many disjoint events that must collectively sum to a probability of 1. Similarly (Jain, 2013) found that asking subjects to “unpack” confidence intervals for the value of a random variable sometime far in the future, by additionally asking for predictions of the same confidence interval at intermediate points in the future, possibly since subjects realize that the widths of their confidence intervals should be monotonically increasing. Finally, a technique known as the premortem in management literature, in which decision-makers are asked to assume that the plan they chose has gone wrong, and to generate the most likely reasons for the failure, has also been shown to correct for overconfidence (Klein, 2007).

The next two cognitive biases are associated with failure to combine new information with preexisting beliefs in the correct proportion. The correction proportion is specified mathematically by Bayes Rule. In some instances, subjects stick too conservatively to pre-existing beliefs (Bayesian conservatism). In others, subjects neglect pre-existing beliefs and make judgments primarily on the new evidence (Base rate neglect). The debiasing techniques developed for base rate neglect were particularly useful in the development of the final game prototype.

Bayesian conservatism

The traditional problem to demonstrate Bayesian conservatism, a judgment error in which subjects fail to change their beliefs to as much of a degree as justified by new evidence, is the Two Urn Problem. Wheeler and Beach (1968) present the problem as follows:

There are two urns full of poker chips. One urn contains, say, .70 red chips and .30 blue chips and the other contains .30 red chips and .70 blue. Out of the subjects’ sight, the experimenter flips a coin to select an urn, draws a random sample of chips with replacement, and displays the sample.

Suppose that the sample contained eight red chips and four blue ones. Estimate the probability of each urn having been the source of observed sample.

Using Bayes Rule, the revised probability for each urn is .97 and .03 for the red urn and blue urn, respectively. However, Wheeler and Beach note that most subjects intuitively estimate .75 and .25, respectively.
The empirical results suggest that subjects intuitively perform an averaging operation to combine the observations with the prior belief, whereas the normatively correct model would require application of the Binomial formula:

\[
\text{Red urn likelihood} = \binom{12}{8} \times (.7)^8 \times (.3)^4 = 23.11\%
\]

\[
\text{Blue urn likelihood} = \binom{12}{8} \times (.3)^8 \times (.7)^4 = 0.78\%
\]

Relative likelihood: 96.7% (red urn) vs 3.3% (blue urn)

Wheeler and Beach attempted two interventions to address conservatism. Their first intervention to teach subjects about the theoretical distribution was not found to be effective. In their second intervention, the subjects made bets about which urn was the source of the samples, and were subsequently told which population was correct. Subjects would receive a low payoff or a high payoff based on their correctness and kept track of a running total of the points they had accumulated.

This intervention was more successful that the first and was able to correct some of the conservatism. While the use of incentive rewards in learning usually brings reason for concern (Kohn, 1999), in the particular case of developing statistical intuition, it seems justifiable in creating a motivated contextual environment for the proper modeling of uncertainty and information. The results of Wheeler and Beach suggest that both exposure to the normatively correct model rather and a method for motivating a shift in mental models is required to change intuition.

**Base rate neglect**

Of the two biases, Bayesian Conservatism and Base Rate Neglect, Base Rate Neglect has been the subject of considerably more research. Base rate neglect can be elicited with the Two Cab Problem (Kahneman, 1982):

Two cab companies operate in a given city, the Blue and the Green (according to the color of the cab they run). Eighty-five percent of the cabs in the city are Blue, and the remaining 15% are Green. A cab was involved in a hit-and-run accident at night. A witness later identified the cab as a Green cab. The court tested the witness’s ability to distinguish between Blue and Green cabs under nighttime visibility conditions. It found that the witness was able to identify each color correctly about 80% of the time, but confused it with the other color about 20% of the time. What do you think are the chances that the errant cab was indeed Green, as the witness claimed?

The Two Cab Problem is an example of a Bayesian inference problem in which a prior belief of the distribution of cabs in the city needs to be updated with new evidence of eyewitness testimony. Most people, however, ignore the base rate of cabs in the city and use the eyewitness account alone, and believe that the probability that it was a green cab is 80%. An answer near 80% is evidence of base rate neglect. The normative answer reached by combining both pieces of information using Bayes Rule is:

\[
\text{Prob green cab} = .15 \times .8 = .12
\]

\[
\text{Prob blue cab} = .85 \times .2 = .17
\]

Relative probability: 41% (green cab) vs 59% (blue cab)
(Gigerenzer, 1995) found that changing the problem to use wording in terms of frequency (“10 out of 1000”) instead of probabilities (“1%”) was able to reduce the strength of base rate neglect in an inference problem related to mammographies and false positives. His interpretation of this result was that the notion of a probability is a recent development in human history whereas our evolved statistical intuitions are better suited to interpreting frequencies. (Gigerenzer, 1991) also found that commitment to random sampling by having subjects draw physical samples from an urn was able to reduce base rate neglect.

A number of researchers (Tversky, 1977; Bar-Hillel, 1980; Krynski, 2007) have found that the effect of base-rate neglect is diminished when the base rates are causal in nature, for instance, when the base rate statistic can be interpreted as information relevant to accident rates. (Nisbett, 1976) hypothesized that this was due to “vividness”: the mind is interested in vivid causal relationships for explanatory power rather than abstract base rates.

Bar-Hillel (1980) developed variations of a similar word problem. Most variations replicated the base rate neglect found by Kahneman and Tversky, but some variations were able to induce subjects to perform more in accord with the normative model. These results induced Bar-Hillel to postulate that information is incorporated using a “relevance” heuristic. Causal base rates are perceived as more relevant to the situation at hand than descriptive base rates. Another method of achieving relevance offered by Bar-Hillel is specificity.

Motivated by the success of Gigerenzer’s frequency format representation, sampling methods, and the observation that causal base rates improve subject performance, this thesis postulates that the gap between subject intuition and the normative model may be narrowed by making the abstract more concrete and easy to visualize. (Gigerenzer, 1995) presents one method of interpreting false alarms and hits devised by one participant, known as the “beam cut” which visually compares the sizes of subgroups being sliced from a beam in order to guide and check a probability estimate.

Before describing the prototypes designed to develop probabilistic intuition, a review of similar technological interventions in the space of education research is reviewed.

Prior work in statistics education research

(Boyle, 2014) conducted large survey of papers on games, animations, and simulations for teaching statistics, finding positive effects on most approaches measured. A subset of the projects is summarized below. Two large projects described in the review were CHERMUG (Johnston, 2013) and Operation ARA (Halpern, 2012), both games for teaching statistics and research methods.

A number of these games motivate statistical concepts by finding realistic instantiations for the concept. For instance, (Chow, 2011) taught expected value through a digital version of “Deal or No Deal,” which was found to improve retention rates one week after the intervention.

(Ancker, 2011) built a game to convey information about health risks in such a way that players could experience the frequency of occurrence of an event, however, they found that this only improve players’ perceptions of risk for low numeracy participants. (Liu, 2010) found that students Simulation Assisted Learning (SALS) were less prone to misconceptions relating to correlations. (Wang, 2011) found that the most effective simulations are interactive ones that allow students to change variables and experience the
effects of their decisions. (Nte, 2008) found that explaining the normal distribution in terms of drinks consumed at a bar was able to decrease anxiety about statistics.

The education research communities have explored a number of systems for the contextualized active learning of statistics. Many of these systems follow a constructivist paradigm of exploration using real data (Papert, 1980). Indeed, guidelines for the design of Statistical Reasoning Learning Environments (SRLE) were set forth by (Garfield, 2008) that included goals to “engage students in making and testing conjectures” and “promotes classroom discourse that includes statistical arguments.”

Notable systems among this class include Tinkerplots (Konold, 2005) and Fathom (Finzer, 2006), designed with the intent that if students have agency in solving problems and modeling situations of interest, they will begin to develop a deep understanding for the normative representations used in the field. (Rubin, 2006) used Tinkerplots with middle and high school teachers as subjects. The teachers used the software to investigate an open-ended question with synthetic data about a factory break-in. Using Tinkerplots to slice and query the data, the teachers seemed to further their intuitions surrounding both the uncertainty of the underlying data as well as the uncertainty of the “correct” methodology.

Working with the NetLogo modeling-and-simulation environment (Wilensky, 1999), (Abrahamson, 2005) developed a suite of models collectively called ProbLab for understanding statistics. Concepts covered include the law or large numbers, stochastic processes, and the relationship between sample frequency and theoretical combination space. Another NetLogo based environment was S.A.M.P.L.E.R., which facilitates the discussion of classroom-driven statistical inquiry (Abrahamson, 2004).

However, these systems all require the presence of a knowledgeable tutor to guide the learning process. The role of the tutor is to notice logical consistencies and guide the learner to clarify misconceptions. I attempt the development of a system in which knowledge is contained within the game and does not require the assistance of a human mentor. A self-contained digital system would scale readily and be accessible to a more students.

**Discarded prototypes**

Chapter 3 describes the final prototype, Adventures in Cognitive Biases, a story with interactive widgets that attempts to reduce biases by providing visual methods for calculating subjective probabilities and performing Bayes inference. The final prototype came closest to achieving the goals outlined in the preceding chapter.

This current chapter describes two earlier prototypes that attempted to teach statistics and probability in a more intuitive way. The first was Bayes Eyes, a speed based game for judging sample representativeness from two alternative distributions. The second was A Random Life that phrased statistical concepts in terms of everyday scenarios. Both of these earlier prototypes were unsuccessful, but instructive in their failures. The prototypes are described here in order to illustrate possible pitfalls and lessons in designing a learning game.
Discarded game prototype #1: Bayes Eyes

The design of this first game was influenced by cognition tests in which a stimulus is presented and the subject is asked to select an answer among those given (Sternberg, 2001).

The game is a guessing game that asks the player to identify the generating distribution of a set of samples as samples rapidly appear on the screen. The game engine draws samples at a rate of one every 200ms and plots them in a histogram format. The player is asked to guess the secret distribution (A or B) as quickly as possible. After the player guesses by hitting keyboard “A” or keyboard “B,” the screen flashes green or red to signify the correctness of the player’s guess, and the game presents a new set of distributions and new samples.

![Screenshot from the Bayes Eyes game](image1)

Figure 1: Screenshot from the Bayes Eyes game

Players played the game for 9 minutes total in three sets of 3 minutes each with the opportunity to review mistakes between each of the three sets. Players were told that they would be scored with a scoring function chosen to incentivize answering as quickly as possible while maintaining accuracy (number_correct – number_incorrect).

Players were also able to review their errors and see the magnitude of their error as calculated using the normative model of independently drawn samples from each discrete distribution with equal prior probabilities. In the example shown below, the normative model strongly favors distribution B. It was hoped that through gameplay and review of performance, the player would adopt the same normative model that the computer was using to compute the likelihood ratio.

![Screenshot from review pane of Bayes Eyes](image2)

Figure 2: Screenshot from review pane of Bayes Eyes
Problems associated with Bayes Eyes prototype

There were a number of design flaws in this game:

- Timing aspect—the fact that it was a speed-based game interfered with learning. Time pressure can used in game design to increase difficulty, but use of time pressure should be used carefully. (Koster, 2013) observes that “Developing speed without precision is not all that useful. Going slow lets you practice the precision first, make it unconscious, and then work on speed.”

- Behaviorist approach was simplistic—the green/red flashing was a feedback mechanism taken from the behaviorist school of thought. Behavioral conditioning methods do work, but they are a primitive form of learning that is unlikely to engender complex higher levels of thought. Binary simplistic right or wrong feedback ignored issues of “why” people were doing what they were doing, so it failed to provide people with deeper mental models.

- Representation dependent—It is likely that any intuition built in a game like this will be tied to the representations used in the game: the histogram representation in this example. If the type of learning in the game is naïve shape-recognition with additional heuristics, it is unlikely that this learning would transfer to other domains. Transferable learning would have to include a model of multiplying independent relative likelihoods of samples to arrive at an overall conclusion. This way of thinking was not facilitated in the design of the game.

- Too abstracted—It is unclear how the game relates to real-life. One of the best motivators for probability and statistical inference is its direct applicability to real-life. The abstract nature made it difficult to see how the knowledge is directly transferrable.

Discarded Game Prototype #2: A Random Life

Another prototype, called A Random Life, was constructed as a game in which probability is directly introduced in the context of realistic life scenarios, thereby avoiding some of the problems that Bayes Eyes had with abstraction and representation dependence. In building a game that was connected to real-life the design attempted to help players make sense of the world around them. This type of connected learning allows for natural questions and extensions of learned material.

The game consists of multiple modules that are not directly connected except that they are all life events. The modules discussed below attempt to teach expected value (through the analogy of opening presents), Bayesian inference (through the analogy of leveling up in a sport with unknown talent), and hypothesis testing with Gaussian noise (through the analogy of inferring whether someone is in love with you or not). The remaining modules from A Random Life are not included in the discussion below because they suffer from similar design problems to the ones discussed below.

Figure 3: Modules from A Random Life
Figure 4: Game instructions for *A Random Life*

The game explored the use of blue buttons to give theoretical formal knowledge for modelling the real-life situations described and orange buttons to allow players to experience randomness in the situation described (to see alternative results from sampling from the distribution described for development of a gut reaction for reasonability).

*A Random Life Module: Expected Value*

One module of *A Random Life* introduced the idea of expected value (EV) and standard deviation (SD) through the physical sensation of opening presents with different random values. The module attempted to liken the wrapped present to the notion of a random variable drawn from a distribution with some known characteristics.

Figure 5: Screenshot of expected value lesson in *A Random Life*

Given the option of only one present to open, the optimal behavior is to open the present with the highest expected value regardless of the standard deviation. It was hoped that players would realize this as they interacted with this module.

Hovering over an orange box caused the computer to randomly generate ten samples from the distribution described and make it available to the player in a side box. In this case, the button hovered over was a uniform distribution with expected value of 4 and a standard deviation of 3. The sample statistics after
drawing 10 random samples differ considerably from the theoretical distribution, hopefully imparting upon players a gut feeling for reasonable deviation from the expected value.

Figure 6: Samples drawn on hover of button in expected value lesson on *A Random Life*. The samples drawn have an average of 5.5 and standard deviation of 1.75, while the theoretical distribution has an average of 4 and standard deviation of 3.

The module also tried to show that very different distributions can give the same type of statistic. For instance, uniform distributions, normal distributions, and even a distribution in which most of the mass is on a single event can all have the same expected value and standard deviation.

Figure 7: Distributions with the same sample statistics
A Random Life Module: Bayesian Inference

Another module from the game introduced Bayesian inference and posterior beliefs. In this example, the player attempted to infer his natural talent for a sport (the weight of a biased coin) as he spends time units attempting to level up each sport. Spending a time unit resulted in a “1” (level successfully increased) or in a “0” (level not successfully increased). Those who have had statistics training would recognize this weighted coin flip as a sample drawn from a Bernoulli distribution.

Figure 8: Leveling up a sport (Bayesian inference) in A Random Life

Here, there were four failed attempts to level up football, so the player likely does not have that much of a natural talent for football. When hovering over the blue football button we see that the posterior belief for Football is more skewed towards unskilled (a low value of theta) than its initial setting. The visual presented in this game of changing beliefs is the same image used in Bayesian inference of a single unknown parameter.
Figure 9: Posterior belief for skill in football after 4 failures and 0 successes (skew towards unskilled)

Here, we have not tried to increase our level in tennis yet, so our belief about our skill in tennis is the same as the given prior Uniform(0.1, 0.4), as we see when we hover over the blue tennis button.

Figure 10: Posterior belief for skill in tennis after 0 failures and 0 successes (skewed neither towards skilled or unskilled)

The near-optimal behavior in this module would be to level up on the skill that has the maximum expected value in the posterior belief. This module was designed to give exposure to the graphical representation of beliefs in such a way to provide intuition for how the shape of the posterior belief should change as new information arrives.

*A Random Life Module: Hypothesis Testing*

In another module from *A Random Life*, the prospect of finding someone to love is modelled as a signal detection problem. Based off of the warmth of interactions as observed data, the player makes an inference on whether the person in question is in love, not in love, or if more information is needed to make the decision.

The situation being modeled is that a person who loves you reciprocally will exhibit more warmth than someone who does not. However, like in real life, most signals are contaminated with some amount of additive Gaussian noise, so that even if the person is sending you a high warmth signal, due to noise it
may be lower than expected. Through multiple independent interactions with the person, the player can make a well-informed estimate for whether the person is in love or not.

Figure 11: Inferring if someone loves you (hypothesis testing with Gaussian noise) in *A Random Life*

Figure 12: Visual representation of the inference model in *A Random Life*

This is a contextualized remake similar to *Bayes Eyes*. Rather than judging representativeness of samples to two abstract distributions, players judge with respect to two signals of different strengths with Gaussian noise. The game makes the analogy between classical signal detection problems and determining whether someone in love with you.
**Problems associated with A Random Life**

There were a number of design flaws in this game that were discovered after building this prototype. Several colleagues played with the prototype, but no user study was conducted. The design flaws included:

- **Concepts were not introduced well**—in this prototype, although the situations used were realistic, insufficient care was made to introduce foreign concepts to a novice learner. This was in part caused by the heuristic of reducing text. Blocks of text were hypothesized to be not game-like, and they intimidated players by being too evocative of traditional textbooks on the subject.

- **Lack of bridging analogies**—although the prototype does try to demonstrate how probabilistic concepts can be used for modelling in everyday scenarios, there was no transition between how players currently think about the scenarios to the formal abstraction. This was not a design shortcoming because diagrammatic models cannot be used until they are internalized.

- **Difficulty did not ramp**—it is important for the game to ramp in difficulty after players achieve a level of mastery. This contributes to the “fun” factor of the game, so the game does not become too repetitive. This also contributes to the learning aspect of the game, because the player is able to perform more difficult actions when the progression is well-designed and ordered by the game designer.

- **Lacked gating**—the game design should set up clearer objectives of what is possible and what can be achieved. A level should not let a player pass unless they are able to achieve X. This guarantees that a player has learned X by the end of the level. Without a gating aspect, there was no clear goal so each player’s improvement got stuck at local optimum.

**Final Prototype**

The *Bayes Eyes* prototype suffered from excessive abstraction and insufficiently directed interaction. The prototype of *A Random Life* moved in the direction of introducing the concepts in realistic scenarios, but fell short in providing bridging analogies to transition players to the more formal way of thinking.

The final design of the game focused on the design considerations given by research on intuitions. In particular, it emphasized the idea of externalization of useful mental models in the game, in attempt to make them more readily adoptable by players.

It explored the use of using diagrammatic models—pictures that allow problems to be solved more quickly. These pictures employ the use of visual processing capabilities and exercise the relationship between sight and intuition, in an attempt to move some of the workload of computation from symbolic processing to the visual (Thurston, 1994). A purposeful focus on visual methods may facilitate the advent of paperless problem solving methodologies, and allow for a greater diversity in mental representations if each individual representation can be called forth, used, and evaluated quickly.

The primary design challenge is to thoughtfully identify useful visual concepts and package them into in-game widgets. The design process is described below for visual models introduced within the game to correct the Overconfidence Bias and Bayes Rule Errors (Base Rate Neglect and Bayesian Conservatism).

The game is modelled after an adventure game with quests and tasks to achieve. The story unfolds as you answer questions correctly and remains stuck if you do not. The story is useful in creating a real-world
setting and keeping the player engaged in a way that *A Random Life* was unable to. The game begins with the player at a monastery.

You have found yourself in front of a monastery....

![Knock on the door](image)

A monk greets you.

![Greet the monk](image)

MONK: Hello, young explorer....

MONK: I am training explorers like yourself to rid themselves of cognitive biases. Cognitive biases are instances in which humans consistently make irrational decisions. Would you like a lesson?

![Start training now](image)

Figure 13: Introductory story to *Adventures in Cognitive Biases*

The monk offers to train the player, but asks that the player complete a pretest (Appendix A) of abilities before enrolling in monk school. The pretest is used to measure the starting point of the player before the intervention. After the pretest is completed, the monk guides the player through a set of trivia questions to reduce the overconfidence bias. The trivia questions require use of an interactive widget for specifying a confidence interval and belief spread and a second interactive widget that links subjective probabilities with fair bets. When players have demonstrated that they can provide sufficiently wide confidence intervals and sufficiently conservative bets, the monk allows them to graduate from monk school. The monk suggests that they visit the oracle on top of the mountain.
The oracle turns out to be the Bayesian Oracle, who suggests that the player seek out the location of Kahneman, a researcher of cognitive biases. However, the location of Kahneman is unknown and the player must infer the location of Kahneman. The oracle gives the player a lesson on visually representing beliefs as graphs and a visual method for doing Bayes Rule using an interactive widget. Performing Bayes Rule correctly involves incorporating new information the correct amount, and serves to treat the two cognitive biases of Base Rate Neglect (belief changed too much in light of new evidence) and Bayesian conservatism (belief changed insufficiently in light of new evidence). After leaving the oracle, the player encounters a number of town folk with inference problems that can be solved with the Bayes Rule widget.

The player must solve the problems correctly to gain additional knowledge of Kahneman’s likely location as well as Kahneman’s favorite food, which the story suggests that the player ought to bring to Kahneman’s likely location. Kahneman’s likely location and favorite food are also set up as Bayesian inference problems in the story as the player gradually gains evidence to update the beliefs. After the player gains enough evidence to infer that Kahneman is probably at the Cave of Uncertainty and would like some cookies, the player adventures to the cave and encounters a floating scroll. The floating scroll contains the posttest questionnaire (Appendix B) and is used to measure if the player has improved after the intervention. After completing the posttest, the story stops at a cliffhanger with no Kahneman in sight and the player unsure if he is truly at the Cave of Uncertainty. The cliffhanger was written in to allow for future installments of Adventures in Cognitive Biases.

The following section describes the Overconfidence Bias and Bayes Rule interventions in depth and illustrates the user experience.

**Design of Adventures in Cognitive Biases**

(a) Designing for the Overconfidence Bias

As described previously, the overprecision variant of the overconfidence interval is typically measured by asking for a subjective probability of confidence or by asking subjects to give an X% confidence interval. In both of these cases, subjects are asked for their subjective beliefs on an abstract notion of X%. If the bias is not due to an intrinsic overconfidence, but rather a misunderstanding on the abstract notion of a probability, then a concrete visual image may help.

If subjects truly are overprecise with their confidences (and the effect is not a misunderstanding of subjective probabilities), then several mental models may help. One mental model is the Bayesian idea of spreading beliefs. In a valid probability density, the area must sum to 1. If a human or a machine is highly confident about the accuracy of the estimate, this is represented with the majority of the probability mass located over a small area of possible estimates. Conversely, if the modeler is highly uncertain about the accuracy of the estimate, this is represented with the probability mass thinly spread over a large area. One visual model for reducing confidence may be to imagine spreading out the probability mass over a larger area.

Concretely, when one is asked to make an estimate, usually a single point estimate comes to mind. It would help to reduce overconfidence if instead a range of reasonable values come to mind. Visually this would be represented as a probability density with mass spread over many values. The shape and the height of the belief over the values are flexible. This more elaborate model allows for a number of more
complex thoughts such as the visual ability to calculate the expected value of the belief by guessing the center of mass, as well as the ability to visually calculate confidence intervals of different sizes.

If there is a misunderstanding of subjective probabilities, that is subjects do not have a miscalibrated understanding of an abstract probability like 90%, then a visual image to make the meaning concrete will help. To make a number like 90% concrete, we can explain a subjective confidence in terms of a fair bet. The game links the subjective probability with the fair bet and attempts to make them equivalent in the eyes of the player.

The game format attempts to realistically mirror a common situation that occurs in the real world. A question is asked, and your first reaction is a point estimate. From the point estimate, you are encouraged to move to a probability spread. After generating a reasonable probability spread, you are then asked the confidence around your probability spread.

First, the player enters a point estimate, since a point estimate is our normal instinctual response when asked to make an estimate.

![Image](image1.png)

**How many years does it take a plastic bag to degrade?**

100 [Guess]

Next, the player uses widget to drag down and spread his belief over a spread of answers.

![Image](image2.png)

**Between 100 and 100**

... 94 95 96 97 98 99 100 101 102 103 104 105 106 ...

Now drag down the circle so that your belief covers all answers that you think are reasonable. This indicates your uncertainty.

Submit confidence interval

Figure 14: Bayesian beliefs spreader in *Adventure in Cognitive Biases*

If the player thinks anywhere between 0 to 200 years is reasonable, the final state of the widget representing the player’s confidence interval might look like this.

![Image](image3.png)

**Between 0 and 200**

0 21 42 63 84 105 126 147 168 189

The player is then asked to make a bet on their confidence interval by stating a level of confidence.
The monk takes the player on the bet and the player gains or loses the number of moneybags indicated by the bet.

MONK: I'll take you on that bet.

The answer is 1000. More importantly, this was not in your confidence interval of 0 to 200. So you owe me 9 moneybags as per our 90% confidence bet.

After losing nine moneybags to the monk, the player might realize that he should both specify wider confidence intervals and have less confidence about his interval.
The player graduates from monk school after meeting the graduation requirements, and the game continues. The monk shows the player what would be a fair bet if the player is truly 99% confident.

![99% confidence and fair bet visualization](image)

Figure 16: 99% confidence and fair bet visualization in *Adventure in Cognitive Biases*

The monk parts ways with the player, giving the player the item of a “queasy stomach” as a reward—a visual reminder of the overconfidence bias.

![Queasy stomach memento](image)

Figure 17: Queasy stomach memento in *Adventure in Cognitive Biases*

The practice of naming important concepts and having them double as game items is yet another attempt to influence player thinking, and may merit a separate user study. It is an attempt to augment intelligence through symbols or language, one of the components of the H-LAM/T system (humans using language, artifacts and methodology in which they are trained) (Engelbart, 1962). This naming methodology is practiced throughout the game with phrases like “overconfidence monk,” “Bayesian Oracle,” and “belief goggles”—phrases that, hopefully, evoke visual representations that serve to embody ideas from the game.

(b) Designing for Intuitive Bayes Rule

After graduating from monk school, the player goes to see the Oracle who teaches the player about Bayesian inference. The Oracle gives the player a widget for performing intuitive Bayes Rule calculations. This widget provides a visual method of performing Bayes calculations using the odds ratio method.

The approach of the widget is to interpret Bayes Rule as saying the “posterior belief is proportional to the prior belief * likelihood of evidence” as is commonly done in machine learning. Bayes Rule calculations can be performed using odds ratios for the prior belief and the evidence by doing visual multiplication.

It also facilitates the development of intuitive gut reactions. For instance, intuitively, if the evidence and prior belief both favor a particular hypothesis, then it is clear that Bayes Rule favors that hypothesis. The
resulting extremeness by which the hypothesis is favored should be more extreme than any component alone. If the evidence and prior belief favor different hypotheses, then the resulting judgment will be in the direction of the stronger two components, but the extremeness of the decision will be weaker than the component itself.

This approach differs markedly from the way that Bayes Rule is traditionally taught in schools as an abstract relationship between two conditional probabilities: event A conditioned on event B and event B conditioned on event A.

\[
P(A \mid B) = \frac{P(B \mid A) \times P(A)}{P(B)}
\]

Figure 18: Traditional abstract presentation of Bayes Rule

It is easy to miss application opportunities of Bayes Rule when it is expressed in its general form. More suggestive is a special case of Bayes Rule as it applies to inference problems.

\[
P(H \mid D) = \frac{P(D \mid H) \times P(H)}{P(D)}
\]

Figure 19: Presentation of Bayes Rule as hypotheses and data

Let D denote “data” and H denote “hypothesis.” This formulation of Bayes Rule says that your belief of a hypothesis after seeing data, \(P(H\mid D)\), is proportional to the likelihood of observing the data under the hypothesis, \(P(D\mid H)\), times your prior belief of the hypothesis \(P(H)\).

When comparing the relative probabilities of multiple hypotheses \(H\), it is sufficient to compare the relative sizes of the product of their \(P(D\mid H)\) (evidence) and \(P(H)\) (prior belief) terms. This computation can be performed visually symbolically and is facilitated by the Oracle’s widget. This is known as the odds ratio version of Bayes Rule.

The odds ratio version of Bayes Rule was chosen for use in the game due to its pencil-less nature. Bayes Rule will likely not be used in casual situations such as walking down the street if it requires use of a pencil (i.e. artifact augmentation). The odds ratio version also continues the pattern of visualizing beliefs as graphs. Visualizing uncertain beliefs as graphs is probably one of the most important mental tools for players to take away from the game.
In the context of finding Kahneman’s location, the oracle introduces the idea of the prior belief and shows it represented as a graph.

**ORACLE:** Well, Kahneman goes to the tavern to study people's decision-making abilities. He goes to the cave to process his data and write papers. In the past Kahneman has split his time between the cave and the tavern in a roughly 2 to 3 ratio (40% vs 60%). ORACLE: You can use this information to create a visual picture of your knowledge. Drag up the bars until their relative heights reflect what you currently know about Kahneman's location.

![Prior belief graph](image)

Next, the Oracle gives evidence to use to modify the prior belief.

Right. Say that when the weather is nice, Kahneman favors going to the cave over the tavern six times more than normal. Add this as evidence by dragging up the evidence bars. Then click and look to see how this evidence affects your new belief.

![Evidence graph](image)

Figure 20: Bayes Rule widget in *Adventure in Cognitive Biases*

The widget calculates the posterior and the player is asked to read a result from the posterior belief.

![Posterior belief calculation](image)

What is the probability that Kahneman is in the cave if the weather is nice?

At this point, the game is asking a difficult question. The tool provides the answer and shows the calculation, and it is assumed that the player would not be able to do this type of calculation unassisted.
yet. It is hoped that through different examples, the player would eventually be able to perform the calculation unassisted.

The player is then tasked with using Bayesian inference to infer whether Kahneman is located in a cave or a tavern. A widget for this inference task appears in a side panel.

The Kahneman inference problem is not immediately solvable. The player first has to journey through neighboring villages solving Bayes Rule problems and gather more information.

For instance, the player first has to help a farmer with an inference problem relating to the type of animal that is leaving marks on a chicken coop.

And there is a somewhat distressed looking farmer on the side of the road.

Ask her what’s wrong

FARMER: There have been thefts from my chicken coop. The local animal population consists of 7 raccoons to every 3 foxes (70% raccoons, 30% foxes). Judging by the type of marks on the door alone, the marks look five times more likely to be caused by a fox than a raccoon.

FARMER: I want to trap the animal, but I only have traps for foxes. If a raccoon goes in, it will get hurt. If I set up the trap, what is the probability that the animal I will catch will be a fox?

Show the farmer the appropriate belief.

The player is tasked with parsing the word problem for the prior belief and evidence information and calculating an explicit probability for the farmer.
The player is presented with more inference problems of this nature, and a shadow figure tells the player that in addition to inferring Kahneman’s location, it is also important to infer whether Kahneman likes cookies or beer.

**SHADOW FIGURE:** Well, he doesn't take visitors easily. You want to take the test of your abilities, I assume. SHADOW FIGURE: But usually you can't get him to see you just by declaring that. I would suggest either saying that you are selling cookies or that you are selling beer. You need to infer which type of food he is interested in before you see him.

The player has to perform five different Bayesian inference tasks before progressing to the final location where a posttest (Appendix B) attempts to measure if players have been debiased since the pretest. Some of the inference tasks involve incorporating new information into existing beliefs of Kahneman’s location.
An adventurer approaches you.

ADVENTURER: I hear you are looking for Kahneman.
ADVENTURER: I am 90% sure that Kahneman is at the cave of uncertainty!

You talk to the adventurer in attempt to find out his intentions. He seems to have no ill wishes to you.

Discounting for his questionable reliability, you think that it is more reasonable to assume that the man is correct with 75% probability (3:1) instead of 90% (9:1).

Use the adventurer's information to update your belief on Kahneman’s location.

You shift course and proceed directly to the Cave of Uncertainty.

After almost a week, you make it to the cave with cookies in hand...

“Hello, Kahneman?” you venture tentatively.

You just hear your echo. But in the cave of uncertainty, lies a fluttering scroll with a test.
Evaluation of Final Prototype

The debiasing performance of Adventures in Cognitive Biases was evaluated with a pilot study. The participants were unpaid and recruited over social media. The game takes one hour to complete so many players dropped out without completing the game; 76 players completed the pretest and 21 players completed the posttest. Of these 76 subjects, all had taken a class in statistics and probability (13 at a high school level, 49 at a college level, and 14 at a graduate level). Thus the participants in this pilot study most likely have had preexisting formal exposure to the concepts in the game. The 21 players who completed the game did not seem to constitute a significantly different population from the 55 players who did not (Appendix C).

Given that the participants in the evaluation of this prototype were self-selected, high-performing, mathematically inclined volunteers, their failures offer much stronger evidence than their successes. The areas in which they were unsuccessful offer strong evidence for a weakness in design, whereas the areas in which they succeed offer hints that may or may not generalize to broader audiences.

Pretest results

The pretest consisted of five binary overconfidence questions as well as two Bayes Rule questions. The pretest responses of the subjects to the overconfidence questions are shown next to the responses from Amazon Mechanical Turk (MTurk) workers. At this point in time, neither the subjects nor the MTurk workers have gone through training. Differences in the shapes of the frequency plots are likely due to differences in the subject pools. On average, the subjects participating in the intervention seem to be less overconfident than the MTurk workers. The question of “which animal has the longer tail: panda or brown bear” ought to be discarded since for most other questions, the confidences exhibit a linear staircase pattern, whereas nobody seems sure about the answer to this question.
Figure 23: MTurk confidence of “Which country produced more cars in 2012: Italy or France?”

Figure 24: Subject pretest confidence of “Which country produced more cars in 2012: Italy or France?”

Figure 25: MTurk confidence of “Which country has worse air pollution: Turkey or Mongolia?”

Figure 26: Subject pretest confidence of “Which country has worse air pollution: Turkey or Mongolia?”

Figure 27: MTurk confidence of “Do more people fear flying (aerophobia) or social situations (sociophobia)?”

Figure 28: Subject pretest confidence of “Do more people fear flying (aerophobia) or social situations (sociophobia)?”
The plots of subjects’ responses to two Bayes Rule questions are shown alongside the results of these questions when originally poised by (Bar-Hillel, 1980).

The subjects in this study exhibited a higher level of Bayesian sophistication than (Bar-Hillel, 1980) subjects. Half the subjects arrived near the normatively correct solution (38 out of 76 subjects). As practiced by (Gigerenzer, 1995), answers were considered to be “near” the normatively correct solution and thus not exhibiting a bias if they were within +/- 3% of the correct solution.

5 out of 37 subjects answered correctly (14%).

38 out of 76 pretest subjects answered correctly (50%).
In the second Bayes Rule question posed on the pretest, the subjects in this study again exhibited a higher level of Bayesian sophistication than in (Bar-Hillel, 1980). 36 out of 76 subjects arrived near the normatively correct answer of 0.41. Only 5 out of 76 subjects gave the answer expected by base rate neglect of 0.8.

Figure 33: (Bar-Hillel, 1980) intercom problem. Normatively correct answer: 0.41.

3 out of 35 subjects answered correctly (9%).

Figure 34: Pretest result of intercom problem. Normatively correct answer: 0.41.

36 out of 76 pretest subjects answered correctly (47%).

In general, from the pretest results, the subjects participating in the study are more sophisticated than the MTurk workers (exhibited lower confidence) and the subjects in the (Bar-Hillel, 1980) study (exhibiting greater frequency of normatively correct answers to Bayes Rule problems).

**Posttest results**

After the intervention, subjects took the posttest. The posttest consisted of five binary overconfidence questions and three Bayes Rule questions. The posttest results discussed below are hopeful but not conclusive for several reasons. The questions were not randomized between pretest and posttest. An attempt to mitigate this effect was taken by comparing against MTurk results to check for uniformity among the certainty of the questions posed. From the MTurk results, we see that the questions posed evoke similar levels of uncertainty and confidence with the exception of the bear tail question which should be ignored because of the high level of uncertainty it evokes.

An additional problem affecting interpretation of these results is the possible confounding factor of selection bias among the completers of the game. The completers of the game may exhibit higher levels of persistence, obedience, or interest than those who did not complete the game. Some effort was made to demonstrate similar performance on the pretest among both groups to suggest that this explanation is not the case (Appendix C).

The frequency plots of the posttest confidence exhibit suggestive differences from the pretest plots. The plots are no longer shaped like linear staircases, and are now crowded on 0.5 and 0.6. These are the levels
of confidence trained by the game to be reasonable bets (1 moneybag for 1 moneybag and 1.5 moneybags for 1 moneybag, respectively).

Table 2: Posttest confidence results

Figure 35: MTurk confidence of “Which state has the larger territory: Alabama or Virginia?”

Figure 36: Subject posttest confidence of “Which state has the larger territory: Alabama or Virginia?”

Figure 37: MTurk confidence of “Which city has more inhabitants? Hyderabad or Islamabad?”

Figure 38: Subject posttest confidence of “Which city has more inhabitants? Hyderabad or Islamabad?”
Figure 39: MTurk confidence of “Are more people on the organ donor waiting list for heart transplants or liver transplants?”

Figure 40: Subject posttest confidence of “Are more people on the organ donor waiting list for heart transplants or liver transplants?”

Figure 41: MTurk confidence of “Which country is more popular as a tourist destination: United States or Spain?”

Figure 42: Subject posttest confidence “Which country is more popular as a tourist destination: United States or Spain?”

Figure 43: MTurk confidence of “What does the average American spend more water using each day: showers or toilets?”

Figure 44: Subject posttest confidence of “What does the average American spend more water using each day: showers or toilets?”
It seems that going through the game has changed the shape of the frequency histograms compared to the MTurk responses, even beyond the differences that were observed in the pretest. The evidence for the effectiveness of the Overconfidence Effect intervention seems fairly strong. The effect of the Bayes Rule intervention is more ambiguous.

Three Bayes Rule questions were used on the posttest (Appendix B):

- The urn problem (Bayesian conservatism and mathematical sophistication)
- The trick cab problem (test for operator-schemata)
- The motor problem (causal base rate)

**Urn problem**

The urn problem was included in the posttest as a measure of mathematical sophistication. Solving the urn problem requires application of the Binomial formula, a topic not covered within the game. 10 out of 21 subjects in the posttest correctly solved the urn problem.

> Imagine ten urns full of red and blue beads. Eight of these urns contain a majority of blue beads, and will be referred to hereafter as the Blue urns. The other two urns contain a majority of red beads, and will be referred to hereafter as the Red urns. The proportion of the majority color in each urn is 75%. Suppose someone first selects an urn on a random basis, and then blindly draws four beads from the urn. Three of the beads turn out to be blue and one red. What do you think is the probability that the beads were drawn from a Blue urn? (Bar-Hillel, 1980)

<table>
<thead>
<tr>
<th>Normative Bayesian assessment</th>
<th>Modal assessment</th>
<th>Freq. of modal assessment</th>
<th>No. of subjects</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.97B</td>
<td>0.75B</td>
<td>14</td>
<td>54</td>
</tr>
</tbody>
</table>

Figure 45: (Bar-Hillel, 1980) urn problem. Normatively correct answer: 0.97. 15 out of 54 subjects gave the incorrect modal estimate of 0.75. No information was provided about the other subjects.

It is interesting to see in the frequency histogram that some subjects reached answers that were both below the odds ratio implied by the base rate and evidence. The fact that answers were in this range may indicate a number of things such as fatigue or the inability to do intuitive gut sanity checks. Both of these factors could be eliminated in future versions of the game through more careful design.
Trick cab question

The following question was included the posttest as a trick question. In this problem, common sense dictates that the second, more specific piece of location-based evidence should replace the first piece of location-based evidence entirely:

*Two cab companies operate in a given city, the Blue and the Green (according to the color of cab they run). Eighty-five percent of the cabs in the city are Blue, and the remaining 15% are Green. A cab was involved in a hit-and-run accident at night. The police investigation discovered that in the neighborhood in which the accident occurred, which is nearer to the Green Cab company headquarters than to the Blue Cab company, 80% of all taxis are Green, and 20% are Blue. What, do you think, are the chances that the errant cab was green? (Bar-Hillel, 1980)*

However, if subjects are blindly following the method of likelihood and priors, they will arrive at the incorrect answer of 0.41. In the results, we see that even subjects with high levels of sophistication arrived at the answer from blindly following the method in the game. This may be evidence for use of operator-schemata methods, in which students “learn to explicitly or implicitly associate a cue in a problem with the strategy for solving the problem (Ben-Zeev, 2001).

(Bar-Hillel, 1980) did not provide a chart, but reported that almost 60% of the 37 subjects gave the correct estimate of 80%

As observed in Figure 47, the performance of subjects who have completed the game declined on a trick question in which only one out of two pieces of information presented is relevant. More subjects from (Bar-Hillel, 1980) answered the question correctly. This decrease in performance may be due to operator-schemata, in which trained learners recognize the form of the problem and blindly apply the taught method. There is anecdotal evidence of this occurring, with one subject writing to me after the study:

*I used "my" method to answer the first set of word problems, and I am fairly certain those were correct. I used "your" method to answer the second set of word problems (because I'm assuming that's what you want participants to do).*

Figure 47: Posttest result trick cab problem. Normatively correct answer: 0.8.
7 out of 21 subjects answered correctly (33%)
If the Bayes Rule game teaches a sequence of operations rather than a method of thinking, this is a rather unfortunate setback. A more charitable explanation may be that untrained subjects exposed to the problem do not have the skill to apply Bayes Rule and thus with fewer strategies at their disposal, and are more likely to guess the correct solution in this instantiation of the problem. When trained subjects encounter the problem, they have more strategies at their disposal and are less likely to hit upon the correct solution by guessing alone.

The design lesson to take away from reduced performance after training is that it is important to design for connections between the new knowledge and existing knowledge. Future iterations a game such as *Adventures in Cognitive Biases* should encourage subjects to reflect “does this reconcile with gut sense” in order to better integrate new knowledge with old, and build up conclusive intuitions.

Additionally, it may be instructive to show learners examples of similar looking problems that have different solutions. This practice is known as near-miss learning and was developed in artificial intelligence for training machines based on a small number of examples (Winston, 1970).

*Motor problem*

The motor problem was identified by (Bar-Hillel, 1980) as an example in which a causal base rate allows the problem to be solved more easily. Despite the complicated wording, the causal nature of the problem renders this an easier Bayes Rule problem to reason about, comparable to the suicide problem in the pretest.

*A large water-pumping facility is operated simultaneously by two giant motors. The motors are virtually identical (in terms of model, age, etc.), except that a long history of breakdowns in the facility has shown that one motor, call it A, was responsible for 85% of the breakdowns, whereas the other, B, caused 15% of the breakdowns only. To mend a motor, it must be idled and taken apart, an expensive and drawn out affair. Therefore, several tests are usually done to get some prior notion of which motor to tackle. One of these tests employs a mechanical device which operates, roughly, by pointing at the motor whose magnetic field is weaker. In 4 cases out of 5, a faulty motor creates a weaker field, but in 1 case out of 5, this effect may be accidentally caused. Suppose a breakdown has just occurred. The device is pointed at motor B. What do you think are the chances that motor B is responsible for this breakdown? (Bar-Hillel, 1980)*

The motor problem was solved correctly by 12 out of 21 subjects (57%) who completed the posttest. This result is only a small improvement over the success rates of the similar problems on the pretest.
Figure 48: (Bar-Hillel, 1980) motor problem. Normatively correct answer: 0.41.

3 out of 39 subjects answered correctly (7%).

Figure 49: Posttest result of (Bar-Hillel, 1980) motor problem. Normatively correct answer: 0.41.

12 out of 21 subjects answered correctly (57%).

If a subject was not able to solve this problem after the intervention, it is worth exploring their train of thought. 3 out of 21 subjects gave the answer of 12% presumably obtained by $15\% \times 4/5 = 12\%$ which reflects a failure to normalize the probability between the two hypotheses (motor A and motor B). The other errors were of unknown cause as participants did not indicate their process.

The overall performance difference between players in the pretest and posttest is summarized in the heat map below. The majority of participants maintained their pretest performance levels with the trick question included and with it removed (Figure 50 and 51).

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pretest # correct (out of 2)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>4</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>2</td>
</tr>
</tbody>
</table>

Figure 50: Comparison of Bayes Rule pretest and posttest results (all questions)
With the trick question removed, the performance exhibited a similar pattern (i.e. no significant improvement between the scores of the pretest and posttest):

<table>
<thead>
<tr>
<th>Pretest # correct (out of 2)</th>
<th>Posttest (# correct out of 2)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

Figure 51: Comparison of Bayes Rule pretest and posttest results (no trick question)

Despite the high level of mathematical and statistical literacy among the subjects, only 2 out of 21 subjects receive a perfect score among the Bayes section of the posttest. This suggests that the game was likely incomplete in showing the space of problems and thinking methods, or perhaps that giving a difficult posttest to subjects who have just completed an hour-long exercise was non-ideal.

Although the quantitative results of the Bayes Rule intervention were disappointing, the qualitative feedback given by the players for this particular section were particularly strong. Multiple players said that they liked Bayes Rule visualization and that it changed the way they saw things. The qualitative feedback from players is discussed in depth in the following section.

Analysis

In addition to the quantitative data given by players pretest and posttest differences, players who completed the game also had feedback and suggestions about the game. Since only those who completed the game provided feedback, there is some selection bias in these perspectives. There were several consistent themes among the posttest feedback.

Overconfidence reduced

A subset of players explicitly said that their overconfidence had been reduced and they

- "Became more conservative in my estimates”
- "Became more risk averse”
- "I do seem to be susceptible to overconfidence when given the chance to define an interval as opposed to just making a binary choice.”

Learned intuitive Bayes

Although most participants have likely been exposed to Bayes Rule before, some seemed to appreciate the more intuitive approach

- "Most in remembering myself to always consider the prior information with the Bayes theorem.”
- "Before, I didn't know how to incorporate new evidence with prior belief to adjust new belief.”
- "One change worth mentioning is that a few of the examples of Bayesian updating were a little different than what I was used to. For instance, the story where Kahneman likes cookies twice as
much as the normal person, whereas he likes beer exactly as much as the normal person, did not sound like a Bayes problem - this tells me about how *much* Kahneman likes each of these things, but I need to know *which* of them Kahneman likes better. But I realized that the whole point of Bayesian learning is to think of belief itself in terms of *much* rather than *which*: not which choice is correct, but how likely each of them is. So those examples helped me think about probabilities in a more Bayesian (and less frequentist) way.”

More willing to do math

Others seemed to become less calculation averse.

“Yes. I relied on math to make my interventions after going through the training rather than just guessing.”

“I have become more careful about the mental calculations and corrected it learning from my previous mistakes in at least one case.”

“Yeah I actually bothered whipping out pencil and paper and doing the math the second time around. I had actually learned the appropriate formulas beforehand, just didn't really feel as motivated to number-crunch during the pretest before doing the exercises.”

Monk overconfidence game has problems

But some of the mathematically more sophisticated players disliked that the optimal strategy for the monk game was too easy. Since the monk in the game is always programmed to bet that the answer to the trivia question is outside of the player’s specified interval, the player should always make intervals that are infinitely wide. Other players observed that there is no incentive in the game to be more than 50% confident about any interval.

“I feel like it's too easy to "game" the monk by picking a big distribution and being 50% confident about it.”

One way to correct these issues raised by the players is to randomize whether the monk bets that the answer is outside of the player’s confidence interval or within. This would require that the player be perfectly calibrated—neither overconfident or underconfident, and make it difficult for the player to game the monk.

Other players felt that the overconfidence widget was not smooth. Indeed it was too discrete and possibly visually unsatisfying due to area conservation.

“The controls were a little clunky, dragging the confidence intervals up and down, for example. liked the prior/evidence visualization”

Liked Bayes visualization

A number of people liked the interactivity of the Bayes widget and found it helpful in understanding the intuition behind the formula in terms of belief updating.

“I thought the bar graphs were very effective. Having to enter your prior belief and your evidence in separate steps made things very clear.”
“I really like the bars to visualize the prior beliefs, etc. It wasn't the way I was visualizing them before.”

“The interactive charts about the belief based on priors makes the probability changes more visual and clear, showing the importance of the prior information.”

“The visuals were awesome, especially for the Bayes' ratios!”

However, one person felt that the widget was too abstract and could have been improved perhaps using frequencies or some indication of the reasoning behind the visualization.

“Graphics showing how a new belief is updated can be better. Straight up giving calculation doesn't really help intuition.”

**Realistic applications**

Players seemed to feel the narrative aspect of the game helped learning. Other players explicitly commented on enjoying the story.

“Applying the concepts learned to problems enforces the learning”

“The metaphors were effective.”

Another player remarked that although the game presented Bayesian inference in situations that seemed familiar, he still felt that it would be difficult to recognize Bayesian inference problems in everyday life.

“I got a tool to apply (Bayesian inference) in situations that seem familiar. Although sometimes it's hard to recognize a familiar situation, mostly because the questions are asked in different words.”

**Not prepared for posttest questions**

One of the common complaints was that the intuitive wording during the story made it hard to transition to the formal wording of the posttest questions.

“I liked the games! I think it's somewhat difficult to make the jump from the simple examples in the games to the more complex tasks in the free response, but maybe that's the point-- does a simple intervention lead to more complex gains?”

“Found it difficult in some cases to extrapolate method to some of the debrief questions.”

**Too long**

Even among the players who finished the game, many found the one-hour game to be too long. This was also a colloquially cited reason that many people did not complete the game. This can be addressed in future iterations by allowing players to jump to smaller module-like sections of the concept and to see a roadmap of how much left there is to accomplish.

“However, the long task made me, sometimes, hard to keep concentrated across intervention logic line.”
“The game is really long. If I didn't know the author, I would have exited the game early.”

**Future Work**

Another method that might be helpful in connecting the new knowledge with existing knowledge is a method proposed by (Smith, 1993) of paradox presentation and resolution. Particularly in case of cognitive biases and real-life situations that subjects have encountered before, subjects likely have a preexisting repertoire of heuristics for use in these situations. In order to override these heuristics, it is required to demonstrate why these heuristics are insufficient. In future iterations of the game, it may be worth experimenting with allowing players to make irrational choices and showing the logical repercussions of those decisions.

The current version of the game was carefully designed to ensure that players were modularly introduced to the concepts required for solving more advanced topics in the game, a concept known in game design as *level design*. Some effort was made to introduce a *gating mechanism* such that a player could not advance until mastery of a concept was achieved as demonstrated by completion of a particular task. This was implemented in the monk overconfidence game as graduation requirements (if the player meets any of the criteria, they have proven that they understand overconfidence). It was implemented in the Bayes oracle game as widget-guided questions that must be answered before the story continues, with the difficulty level of these questions slowly increasing.

However, due to programming costs, there were some elements of the game that bordered too much on the repetition side. For instance, in the monk overconfidence game, the format remained the same even as the player advanced in skill. One subject remarked profoundly, “I think visualizing things in multiple ways is effective. Also, repetition in slight variation is effective…” Indeed, there are many different ways in which a confidence interval can be intuitively explained even though only a limited subset is considered safe by statistics teachers (Foster, 2014), and it seems well worth the effort in terms of richness in mental models developed and fun factor of the game to explore these variations.

On the Bayesian inference half of the game, more repetition with variation would have been helpful as well. Although feedback for the interactive Bayes Widget was generally positive, subjects felt that there was a difficulty gap between the questions encountered in the game and the questions on the posttest. Indeed the Bayesian inference questions encountered during the game were designed by the author to use real-world examples and intuitive language to motivate the wide applicability of Bayes Rule to frequently encountered situations. The questions on the posttest were rather convolutedly written by psychologists in hopes of demonstrating human irrationality.

The path to take as the designer of the game-based intervention depends on the objective. If the intent of the game is to promote use of formal mathematical thinking in realistic situations, then a measurement method more representative of those realistically encountered, such as that of (Hestenes, 1995) ought to be developed. If the intent of the game is both to promote formal thinking in the real world as well as to improve performance on traditional testing methods, then some method of generating questions between these two different points in the spectrum must be used in order to account for bias generated by using unrealistic measurement methods.
Reflections—On the biases of a novice designer

Prototyping as Bayesian Inference

As I was designing games to correct cognitive biases, writing about Overconfidence and Bayes Rule errors of failing to incorporate new evidence, I failed to update my beliefs on the success of my own intervention.

At the prototype-level, I never stopped to quantify the probability that an individual prototype would work—until the prototype was almost done. At the time, it felt like the only way I could "know" if the prototype was good or not, would be to build it out in its entirety and test it. However, as thesis reader Dor Abrahamson had me reflect afterwards, I realized that I could have viewed the situation from a Bayesian inference perspective: that I could build small parts and get additional "information" on whether the prototype would succeed or not. Even though I would not know for sure, I could model the problem in this way and get an estimate.

In fact, this exact situation of a Bayesian inference problem in which each additional unit of information requires a small cost was modelled by both the Bayes Eyes game and the finding love module within A Random Life. Viewing the prototyping process as a Bayesian inference problem allows early termination of the expensive process and greater awareness of the parts of a successful prototype.

At the thesis level, I also didn't quantify the success probability but just relied on gut estimates. Over the course of my thesis, my initial gut feeling of "it'll probably work" (which was almost definitely overconfidence) morphed into the gut feeling of "it might work; it might not work." However, I never bothered to quantify the gut feelings by assigning them a concrete number, even though I had developed methods to do so for Adventures in Cognitive Biases.

Why did I not apply the methods I knew to this context? At the time, it simply did not occur to me to do so. I should have had an alert that triggers when embarking on expensive processes, that my gut instinct for the optimal thing to do should be double checked with all of the engineering best practices for analyzing systems. This heuristic should serve as a reminder to be wary of our natural intuitions when the costs are high. While learned heuristics and alerts do not have much elegance, they might be a necessary form of training wheels to apply new knowledge as mental models are reconfiguring.

Related work and overconfidence

Coming to the field of education design for the first time, many of the ideas in the literature appeared to be nice-to-haves or features. I was overconfident about the probability of success, and lax in following the best practices and ideas within the literature. I was more tempted to try things that I had not seen before.

This was definitely an overconfident approach to the thesis and I would likely have benefited from some of the calibration strategies developed to treat overconfidence. On the long run, it is unclear whether this overconfidence was overall detrimental—building prototypes that did not work and reflecting on the reasons they did not, provided strong emotionally-weighted justification for many literature ideas that initially seemed abstract. Now I have visceral reactions to the importance for some ideas grounded in personal experience.
While it is the hope of this research that some forms of knowledge can be made more intuitive through better visual representations, this prototyping experience suggests that perhaps other forms of knowledge can only be learned through personal experience. Emotional judgments on the goodness of certain ideas may fall into this category. Research by (Borgida, 1977) indeed find that information that is presented vividly is weighted much more heavily than what is presented abstractly. At least for me personally, the abstract nature of much related literature seemed to affect its incorporation into my mental models.

One surprising lesson from this experience was that my research on the Overconfidence Effect and Bayes Rule errors for incorporating new information is more widely applicable than I realized. That although I thought I personally was immune to biases of out-of-classroom usages of academic material, there seems to have been at least one important and expensive instance that escaped me.

**Conclusion**

This concludes a foray into the design process and multiple game prototypes that attempt to alter fundamental probabilistic intuitions. The most important lessons from the failed prototypes were the importance of creating bridging analogies and ordering lessons in such a way that repetition with variation is achieved—helping learners transition smoothly from their existing mental models to formal representations. The final prototype was the closest to fulfilling the design goals of this thesis of presenting probabilistic concepts in an intuitive and applicable manner. The results were suggestively mixed—offering both hope for the underlying ideas and warning of the executional difficulties in educational design. Although there were limitations in the study due to both the special population and design, I tentatively offer the following results.

By definition, intuition is knowledge that is immediately accessible. Educational materials today that emphasize formulas and procedures seem to lack the immediacy necessary for intuitive application in out-of-classroom contexts. The conscious design of new representations that can be quickly utilized may help with out-of-classroom applications of ideas. Visual diagrammatic representations to supplant traditional text-based or symbolic-representations may add richness to the mental models that people apply. New encodings potentially offer more creative encodings that challenge both learners and designers to think in different ways.

Designing for intuition change requires careful consideration of existing intuitions. And the ideal learning process likely begins with the diagnosis of existing intuitions and a carefully arranged ordering of experiences that help learners transition towards formally correct intuitions. The creation of analogies between intuitive concepts from one domain to another may be one way to efficiently encourage this transition, and one that comes with the advantageous side effect of fostering creative connections between disparate ideas. Yet, in the meta-story of writing this thesis, I have personally experienced that some forms of learning cannot be designed for, and may require the traditional process of trial and error before internalizing the lessons that older and wiser mentors have tried to impart.
References


Appendix A: Monk’s Pre-test

How familiar are you with cognitive biases?
(a) I have not heard the term used before
(b) I have heard the term used before but cannot give an example
(c) I can give an example of a cognitive bias
(d) I can give an example of a word question used to test for a cognitive bias

Describe your background in probability/statistics (if any).
(a) No formal training
(b) High school class in statistics/probability
(c) College level class in statistics/probability
(d) Graduate level class in statistics/probability

Which animal has the longer tail?
(a) brown bear (b) panda

And how confident are you that your previous answer is correct?
(a) 50% (b) 60% (c) 70% (d) 80% (e) 90% (f) 100%

Which country produced more cars in 2012?
(a) France (b) Italy

And how confident are you that your previous answer is correct?
(a) 50% (b) 60% (c) 70% (d) 80% (e) 90% (f) 100%

Which country has worse air pollution?
(a) Mongolia (b) Turkey

And how confident are you that your previous answer is correct?
(a) 50% (b) 60% (c) 70% (d) 80% (e) 90% (f) 100%

Do more people fear flying (aerophobia) or social situations (sociophobia)?
(a) aerophobia (b) sociophobia
And how confident are you that your previous answer is correct?
(a) 50% (b) 60% (c) 70% (d) 80% (e) 90% (f) 100%

Does cheese consumption increase or decrease with education level?
(a) increase (b) decrease
And how confident are you that your previous answer is correct?
(a) 50% (b) 60% (c) 70% (d) 80% (e) 90% (f) 100%

----

You are welcome to use pencil, paper, and calculator on these questions. If you are uncertain of your answer or are having difficulty arriving at a number, it would be very helpful to me if you would use the space to type in your thought process instead.

A study was done on causes of suicide among young adults (aged 25 to 35). It was found that the percentage of suicides is three times larger among single people than among married people. In this age group, 80% are married and 20% are single. Of 100 cases of suicide among people aged 25 to 35, how many would you estimate were single? (Bar-Hillel, 1980)

Two cab companies operate in a given city, the Blue and the Green (according to the color of cab they run). Eighty-five percent of the cabs in the city are Blue, and 15% are Green. A cab was involved in a hit-and-run accident at night, in which a pedestrian was run down. The wounded pedestrian later testified that though he did not see the color of the cab due to the bad visibility conditions that night, he remembers hearing the sound of an intercom coming through the cab window. The police investigation discovered that intercoms are installed in 80% of the Green cabs, and in 20% of the Blue cabs. What do you think are the chances that the errant cab was Green? (Bar-Hillel, 1980)
Appendix B: Fluttering Scroll Posttest

Which state has the larger territory?
(a) Alabama (b) Virginia
And how confident are you that your previous answer is correct?
(a) 50% (b) 60% (c) 70% (d) 80% (e) 90% (f) 100%

Which city has more inhabitants?
(a) Hyderabad (b) Islamabad
And how confident are you that your previous answer is correct?
(a) 50% (b) 60% (c) 70% (d) 80% (e) 90% (f) 100%

Are more people on the organ donor waiting list for heart transplants or liver transplants?
(a) Heart (b) Liver
And how confident are you that your previous answer is correct?
(a) 50% (b) 60% (c) 70% (d) 80% (e) 90% (f) 100%

Which country is more popular as a tourist destination?
(a) Spain (b) United States
And how confident are you that your previous answer is correct?
(a) 50% (b) 60% (c) 70% (d) 80% (e) 90% (f) 100%

What does the average American spend more water using each day?
(a) showers (b) toilets
And how confident are you that your previous answer is correct?
(a) 50% (b) 60% (c) 70% (d) 80% (e) 90% (f) 100%

----
You are welcome to use pencil, paper, and calculator on these questions. If you are uncertain of your answer or are having difficulty arriving at a number, it would be very helpful to me if you would use the space to type in your thought process instead.

Imagine ten urns full of red and blue beads. Eight of these urns contain a majority of blue beads, and will be referred to hereafter as the Blue urns. The other two urns contain a majority of red beads, and will be referred to hereafter as the Red urns. The proportion of the majority color in each urn is 75%. Suppose someone first selects an urn on a random basis, and then blindly draws four beads from the urn. Three of
the beads turn out to be blue, and one red. What do you think is the probability that the beads were drawn from a Blue urn? (Bar-Hillel, 1980)

Two cab companies operate in a given city, the Blue and the Green (according to the color of cab they run). Eighty-five percent of the cabs in the city are Blue, and the remaining 15% are Green. A cab was involved in a hit-and-run accident at night. The police investigation discovered that in the neighborhood in which the accident occurred, which is nearer to the Green Cab company headquarters than to the Blue Cab company, 80% of all taxis are Green, and 20% are Blue. What, do you think, are the chances that the errant cab was green? (Bar-Hillel, 1980)

A large water-pumping facility is operated simultaneously by two giant motors. The motors are virtually identical (in terms of model, age, etc.), except that a long history of breakdowns in the facility has shown that one motor, call it A, was responsible for 85% of the breakdowns, whereas the other, B, caused 15% of the breakdowns only. To mend a motor, it must be idled and taken apart, an expensive and drawn out affair. Therefore, several tests are usually done to get some prior notion of which motor to tackle. One of these tests employs a mechanical device which operates, roughly, by pointing at the motor whose magnetic field is weaker. In 4 cases out of 5, a faulty motor creates a weaker field, but in 1 case out of 5, this effect may be accidentally caused. Suppose a breakdown has just occurred. The device is pointed at motor B. What do you think are the chances that motor B is responsible for this breakdown? (Bar-Hillel, 1980)

---

A panel of psychologists have interviewed and administered personality tests to 30 engineers and 70 lawyers, all successful in their respective fields. On the basis of this information, thumbnail descriptions of the 30 engineers and 70 lawyers have been written. Here are descriptions that have been chosen at random from the 100 available descriptions. For each description, please indicate your probability that the person described is an engineer, on a scale from 0 to 100. (Tversky, 1974)

Dick is a 30 year old man. He is married with no children. A man of high ability and high motivation, he promises to be quite successful in his field. He is well liked by his colleagues. The probability that Dick is an engineer is __%. Describe your reasoning.

Jack is a 45-year-old man. He is married and has four children. He is generally conservative, careful, and ambitious. He shows no interest in political and social issues and spends most of his free time on his many hobbies which include home carpentry, sailing, and mathematical puzzles. The probability that Jack is an engineer is __%. Describe your reasoning.

---

What did you feel was effective vs ineffective about the game as a cognitive bias intervention? Were there changes in how you thought about things after going through the game? If so, explain.
Appendix C: Comparison of pretest statistics for players who completed the game versus those who did not

Figure 52: Pretest results for (Bar-Hillel, 1980) suicide problem for players who completed the game

Figure 53: Pretest results for (Bar-Hillel, 1980) suicide problem for all players

Figure 54: Pretest results for (Bar-Hillel, 1980) two-cab problem for players who completed the game

Figure 55: Pretest results for (Bar-Hillel, 1980) two-cab problem for all players