A multiple cue threshold learning model of selection and detection: balancing judgmental accuracy with threshold learning

April M. Roggio

University at Albany, State University of New York, aroggio@nycap.rr.com

The University at Albany community has made this article openly available. Please share how this access benefits you.

Follow this and additional works at: https://scholarsarchive.library.albany.edu/legacy-etd

Part of the Psychology Commons, and the Public Administration Commons

Recommended Citation

This Dissertation is brought to you for free and open access by the The Graduate School at Scholars Archive. It has been accepted for inclusion in Legacy Theses & Dissertations (2009 - 2024) by an authorized administrator of Scholars Archive. Please see Terms of Use. For more information, please contact scholarsarchive@albany.edu.
A MULTIPLE CUE THRESHOLD LEARNING MODEL OF SELECTION AND DETECTION: BALANCING JUDGMENTAL ACCURACY WITH THRESHOLD LEARNING

By

April M Roggio

A Dissertation

Submitted to the University at Albany, State University of New York in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

Rockefeller College of Public Affairs and Policy Department of Public Administration

2010
Abstract

Selection and detection problems represent some of the most challenging decision making tasks, especially in the fields of health and medicine. In a population of pregnant women, who is a candidate for a caesarean delivery? Does this mammogram indicate the presence of cancer? Should antibiotics be prescribed for this illness? We must judge cues, make decisions and evaluate feedback, in an uncertain, high stakes environment. This particular type of problem is relevant and timely, especially in the arena of public policy, yet we have limited understanding about how people make decisions in these circumstances.

The purpose of this investigation is multifold. In the following pages, I examine the literature, identifying principles and modeling rules based on the large volume of prior research. This review covers tool based analysis of judgment and decision making, explores the relationships between feedback and accuracy, feedback and confidence and feedback and implicit learning, while untangling the results of years of model building on cue judgment and threshold learning.

I identify the key elements of decision maker behavior, based on an analysis of recent experimental data, applying, wherever possible, those elements to a current problem in public policy and model decision making: the significant rise in caesarean deliveries. I present a simple model that combines error correction and hill-climbing principles, which provides a good match to empirical data, as well as demonstrating that it effectively seeks an optimal threshold in an uncertain environment.

I conclude by offering some general conclusions about what we know about this sizable intellectual landscape, and what additional insight has been drawn from this modeling investigation. In particular, I note that this research has shed some light on why the caesarean rate has risen so dramatically, and why it is a direct and very reasonable result of the decision making environment. And I will offer some prescriptions, based on model based exploration, and supported by the literature and empirical testing, about where we are at, and what we still need to uncover, in order to make better decisions.
# Table of Contents

Abstract ________________________________________________________________ ii

Table of Figures _________________________________________________________ v

1. Introduction ________________________________________________________  1

2. A review of the literature ______________________________________________  5

   Traditional approaches to understanding judgment and yes-no detection learning ____  5

      2.1. Judgment _____________________________________________________  6

          2.1.1. Understanding the landscape of learning: Brunswikian psychology ____  6

          2.1.2. Multiple cue probability learning _______________________________  9

          2.1.3. Lens model analysis ________________________________________  10

          2.1.4. Choice/decision _____________________________________________  11

              2.1.4.1. Signal detection theory ____________________________________  11

              2.1.4.2. Taylor Russell analysis ____________________________________  12

          2.1.5. Summary _________________________________________________  14

      2.2. Reviewing the literature: what do we know? _________________________  14

          2.2.1. Accuracy and the influence of feedback __________________________  15

              2.2.1.1. Judmental accuracy and feedback ___________________________  15

              2.2.1.2. Threshold placement accuracy and feedback ___________________  20

          2.2.1.3. Summary _______________________________________________  22

      2.2.2. Feedback effect on confidence in judgment and threshold learning _____  23

          2.2.2.1. Summary _______________________________________________  27

      2.2.3. Feedback and the role of implicit learning _________________________  28

          2.2.3.1. Summary _______________________________________________  31

      2.3. Modeling behavior in feedback-driven multiple cue threshold learning models __________  32

          2.3.1. Modeling judgment (cue learning models) _______________________  32

          2.3.2. Modeling threshold learning __________________________________  33

          2.3.3. Summary and modeling insights _______________________________  37

3. Key elements to include in a model ______________________________________  38

4. Understanding the data ____________________________________________________  40

   4.1. The task _______________________________________________________  40

   4.2. Results from base rate changes _____________________________________  43

   4.3. Results from changes in feedback ___________________________________  46

   4.4. Results from changes in uncertainty _________________________________  48

   4.5. Results from changes in values _____________________________________  51

   4.6. Summary ______________________________________________________  54

5. Model Construction _____________________________________________________  56

   5.1. Examining the whole _____________________________________________  57

   5.2. The environment sector ___________________________________________  60

   5.3. The payoff sector _________________________________________________  61

   5.4. The threshold adjustment sector ____________________________________  62

6. Model evaluation _________________________________________________________  65

   6.1. A simple, full feedback single cue model _____________________________  65

   6.2. Three cues, at .5 base rate, and with full feedback ______________________  67

   6.3. Changing the base rate _____________________________________________  68

   6.4. Conditional feedback ______________________________________________  72
Table of Figures

Figure 1: Multiple cue threshold learning _____________________________________ 5
Figure 2: Brunswik’s Lens Model, from Hammond and Summers (1972) __________ 8
Figure 3: Taylor and Russell (1939) identifies the selection ratio ________________ 12
Figure 4: Examining the differences in errors, as the selection rate changes. _______ 13
Figure 5: Feedback’s effect on accuracy, and the variables that shape the relationship __ 23
Figure 6: Sample of subject screen in experiment ______________________________ 41
Figure 7: Subjects had the option of viewing information about the cue weights ______ 42
Figure 8: Overall performance (points) for three base rates, across blocks of 100 ______ 44
Figure 9: False alarm rates, across of blocks of 25, indicating differences among base rates __ 44
Figure 10: Misses over all base rates, across blocks of 25 _______________________ 45
Figure 11: Hit rate for all base rates, across blocks of 25 _______________________ 45
Figure 12: Proportion of positive responses overall all base rates ________________ 46
Figure 13: Examination at the .5 base rate ________________________________ 47
Figure 14: Examination at the .8 base rate ________________________________ 48
Figure 15: Examination at the .1 base rate ________________________________ 48
Figure 16: Different levels of uncertainty (R² = .5, R² = .7, and R² = .9) at the .5 base rate ______ 49
Figure 17: Different levels of uncertainty (R² = .5, R² = .7, and R² = .9) at the .8 base rate ______ 50
Figure 18: Different levels of uncertainty (R² = .5, R² = .7, and R² = .9) at the .1 base rate ______ 50
Figure 19: Examining behavior at the .5 base rate with more uncertainty (R² = .5), for full and conditional feedback ________________________________ 50
Figure 20: Examining behavior at the .8 base rate with more uncertainty (R² = .5), for full and conditional feedback ________________________________ 51
Figure 21: Examining behavior at the .1 base rate with more uncertainty (R² = .5), for full and conditional feedback ________________________________ 51
Figure 22: Examining behavior at the .1 base rate, with full feedback, and differences in penalties ________________________________ 53
Figure 23: Examining behavior at the .1 base rate, with conditional feedback, and differences in penalties ________________________________ 53
Figure 24: Examining behavior at the .1 base rate, with full and conditional feedback, and penalties on false negatives, by rate of false negative errors ______ 53
Figure 25: Examining behavior at the .1 base rate, with full and conditional feedback, and penalties on false negatives, by points __________________________ 54
Figure 26: Signal detection graph __________________________________________ 57
Figure 27: Basic structure of threshold learning model _______________________ 58
Figure 28: Basic structure, expanded to include cues, cue weights, and judgment _____ 58
Figure 29: Basic structure, with influence of feedback on threshold adjustment _____ 59
Figure 30: Summarized view of the environment sector _______________________ 60
Figure 31: Summarized view of the payoff sector ______________________________ 61
Figure 32: Table of penalties ______________________________________________ 61
Figure 53: Comparison of simulation results for .1 base rate, conditional feedback condition with additional penalties on errors. ___________________________________________ 87
Figure 54: Comparison of experimental results for .1 base rate, conditional feedback condition with additional penalties on errors. ___________________________________________ 87
Figure 55: Threshold setting screen ___________________________________________ 91
Figure 56: Base runs for the medium base rate. The graph on the left indicates the **smoothed cutoff** over time, for 200 runs; the graph on the left is the **Indicated directed of next cutoff** ___________________________________________ 92
Figure 57: A series of graphs illustrating the number of true positives, true negatives, false positives and false negatives in the base run. ___________________________________________ 93
Figure 58: Medium base rate, with longer lag values on the recent cutoff and the recent payoff ___________________________________________ 93
Figure 59: Recent payoff - at medium base rate with longer recents ___________________________________________ 94
Figure 60: Smoothed cutoff and indicated threshold change at the medium base rate with a longer value on time to consider prior cutoffs ___________________________________________ 94
Figure 61: Smoothed cutoff (left) and indicated threshold change (right) at the medium base rate with longer value on the time over which to consider the next direction ___________________________________________ 95
Figure 62: Smoothed cutoff at the medium base rate with all smoothing times lengthened ___________________________________________ 96
Figure 63: Series indicating the frequency of true positives, true negatives, false positives and false negatives. ___________________________________________ 96
Figure 64: Low base rate, with longer smoothing times for all memory variables ___________________________________________ 97
Figure 65: Smoothed cutoff after Busemeyer/Myung weighting function is removed ___________________________________________ 98
Figure 66: Series indicating the frequency of true positives, true negatives, false positives and false negatives ___________________________________________ 98
Figure 67: High base rate, with improved structure ___________________________________________ 99
Figure 68: Series indicating frequencies of true positive, true negative, false positive and false negative ___________________________________________ 100
Figure 69: High base rate, with conditional feedback ___________________________________________ 101
Figure 70: Series indicating the frequency of true positives, true negatives, false positives and false negatives ___________________________________________ 101
Figure 71: Changes in uncertainty as represented by d’ values of the cues ___________________________________________ 102
Figure 72: Low base rate, with conditional feedback, with penalties on false negative errors (above, with smooth threshold on the left and the indicated change on the left) and false positive errors (below, with smooth threshold on the left and the indicated change on the left). ___________________________________________ 103
Figure 73: Low base rate, with conditional feedback and greater rewards for true positives and true negatives. Smoothed cutoff is shown on the left and the indicated direction of the cutoff is shown on the right. ___________________________________________ 104
Figure 74: Results, clockwise from top left: True positive, true negative, False negative, False positive. ___________________________________________ 104
Figure 75: Smoothed cutoff (on the left) and indicated direction of cutoff (on the right) with altered cue structure ___________________________________________ 105
Figure 76: Turning conditional feedback on its head: examining the selection rate when feedback is only on “no” ___________________________________________ 106
1. Introduction

Selection and detection problems represent an archetypal decision dilemma. Who should be admitted to a particular college? Who should undergo a more rigorous search at sensitive transportation hubs? How do we set the legal limit for certain criminal behaviors, such as the threshold for driving while intoxicated, or student behavior in this era of school shootings? This particular type of problem is relevant and timely, especially in the arena of public policy, yet we have limited understanding about how people make decisions in these circumstances.

The study of selection and detection problems has interested scholars from a number of fields for many decades. Beginning with Brunswik (1935, 1939), scholars have examined the environment in which the problems are faced. They have examined how people judge the cues, or indicators, and how they can get more accurate (see Holzworth in Hammond and Stewart, 2001, for a review). Some have studied how people set thresholds in order to improve outcomes (for example, Green and Swets, 1966). Others have considered how feedback affects our cue judgments (Summers, 1962; Hammond and Summers, 1965 and 1972; Einhorn and Hogarth, 1978), and how it affects our placement of a threshold (Kubovy and Healy, 1977; Erev, 1998; Swets, Dawes and Monahan, 2000). They hypothesized about the internal learning mechanisms, such as confidence (Sharp, Cutler, Penrod, 1988; Pease and Sniezek, 1991; Subbotin, 1996) and intuitive reasoning (Hammond, 1996), and modeled the ways such mechanisms shape learning (Einhorn and Hogarth, 1978; Fischer, 2005). Scholars have tested people, created models, retested people and retooled models.

A review of this literature allows for some general conclusions. First, understanding human behavior in the context of selection and detection problems requires understanding how we learn to make judgments when presented with multiple cues, and how well we learn to find the best location for the decision threshold. Empirical work attempts to understand both skills. When explored as complementary research threads, significant overlap becomes obvious, specifically within the empirical work that studies the influence of feedback. To varying degrees, both the judgment and decision making literature have tried to uncover the influence of feedback and its effect on accuracy.

Currently, models exist that try to reproduce the behavior exhibited by people making judgments (for example, Kruschke and Johansen, 1999; Palmeri, 1999; and White and Koehler, 2004) and models exist that try to reproduce the behavior exhibited by people searching for the optimal threshold (Busemeyer and Myung, 1992; Erev, 1998; Camerer and Ho, 1997, 2000), but none exist that model learning in both situations simultaneously. In most cases, the process people undertake when faced with selection and detection problems involves learning both to make judgments and to set thresholds. Additionally, current modeling is often set in a completely different research track, which only rarely acknowledges the accumulation of empirical work from either the judgment or decision making
literature. Worse, few models since Einhorn and Hogarth (1978) have been willing to attempt to apply their modeling conclusions to real world problems; most have remained abstract to the degree that they offer little assistance to policy makers facing challenging decision tasks.

Based on these observations, I propose that the judgment and decision-making process under conditions of uncertainty be modeled using a methodology that explores the interplay of multiple variables over time. System dynamics modeling is constructed to model complex, interrelated causal feedback structures, especially those structures embedded in social systems. It is also a complementary approach to contemporary methods used to explore learning behavior in selection and detection environments, such as signal detection and Taylor Russell analysis.

I will first review the traditional approaches used to study decision-making behavior when faced with selection and detection problems. I will then review the empirical work associated with these approaches, with the following questions in mind:

1. Under what conditions does feedback improve judgmental accuracy and under what conditions does feedback improve accuracy of threshold placement?
2. How does feedback affect confidence in judgment and threshold placement?
3. What role does implicit learning play in improving judgmental accuracy and threshold placement?
4. How do we conceive of a multiple cue threshold learning model capable of reproducing human behavior when faced with multiple cues and a threshold placement requirement?

Based on a survey of the literature, I will identify areas where results converge on a common conclusion as well as areas of disagreement and apparent contradiction. Gaps in the literature will also be identified. Using the theory and empirical results from the literature, I will identify the key elements necessary to construct a multiple-cue threshold learning model, borrowing from a number of prior models.

Throughout this work, I will also rely on a current medical issue as a guiding example of the complexities of signal detection under complex circumstances. The caesarean rate in the US reached 27.5% of all births in 2003, the highest rate ever reported, and it has continued to increase since. The latest reports place the c-section rate at 31.8% in 2007, a rate which concerns activists as well as expectant mothers. Driven in part by the reluctance of hospitals and care providers to allow for VBAC (vaginal births after caesareans), the rate of caesarean delivers incurs huge medical costs as well as high rates of morbidity for women and babies.¹

Recent reports suggest that the risks of caesarean delivery include a much higher probability of return hospitalization as well much higher odds that infants will

¹ See, for example, Centers for Disease Control, Birth Statistics.
endure stays in neonatal intensive care.\(^2\) Zhang et al (2010) indicates that although there is much conjecture about the unnecessarily high rate of caesareans, including claims that high rates of insurance and reduced levels of prenatal care in some communities, we have very little reliable information about why the rate of surgical deliveries is so high, and the authors suggest that the upward trend is likely to continue. In fact, in spite of a decade of popular news coverage of the issue, and repeated calls for activism to reduce the caesarean rate, it has continued to increase.

This dilemma has prompted some research over the past several decades, and a lot of conjecture, by academics, activists and physicians:

Within broad upper and lower limits, rates of operative delivery in the United States are highly variable and suggest a pattern of almost random decision making. This reflects a lack of sufficient reliable, outcomes-based data to guide clinical decision making. (Clark et al 2007)

It has been estimated that approximately half the caesareans currently performed in the U.S. are medically unnecessary, resulting in considerable avoidable maternal mortality and morbidity, and a cost of over $1 billion each year. (El Sherer, 1993)

Additionally, although many decry the high and rising rate of caesareans, there is little consensus about what policy action might assist:

Attempts to define, or enforce, an “ideal” caesarean section rate are futile, and should be abandoned. The caesarean rate is a consequence of individual value-laden clinical decisions, and is not amenable to the methods of evidence-based medicine...Like other population health indices, the caesarean section rate is an indirect result of American public policy during the last century. Without major changes in the way health and maternity care are delivered in the US, the rate will continue to increase without improving population outcomes. (ACOG, 2006)

Despite this, government has attempted to intervene: the Centers for Disease Control (CDC) has an explicit (though incredibly unlikely) goal of achieving a caesarean rate of no more than 15% by 2010, as well as increasing the rate of VBACs to 37% from its current levels of around 8%. Recently, the American College of Obstetricians and Gynecologists has attempted to impact the increasing rate of caesareans by issuing new, more flexible guidelines regarding VBAC.\(^3\)

This issue has several benefits here. It is a contemporary problem and has clear threshold learning ramifications. Moreover, it represents the basic issues shared by many medical problems and the lessons learned from this example can be generalized to many other medical issues that involve making judgments and setting thresholds. As will be discussed in more detail later, judgments are based on several unequally weighted cues, or indicators. Physicians would use various tools to interpret symptoms, for example. This problem has a base rate, indicating where the

---

\(^2\) As reported in the New York Times, August 30, 2010 by Denise Grady

\(^3\) As reported in the New York Times on July 21, 2010 by Denise Grady. Full description of new guidelines are available at the ACOG’s website:

http://www.acog.org/from_home/publications/press_releases/nr07-21-10-1.cfm
caesarean rate should be, in an ideal world. There is a selection rate, which indicates how many women are currently undergoing caesareans. There is a high degree of uncertainty, indicating that the cues are ambiguous and difficult to judge, as well as conditional feedback, so information about results is restricted. In the case of caesareans, usually the only time a doctor knows for a fact that a caesarean delivery was the correct call was when he decided *not* to perform one, a false negative error. Lastly, there are values that can affect decision outcomes, such as when hospitals inflict heavy penalties on such false negatives errors, pressuring physicians to lower their thresholds. In addition to physician values and hospital values, patient and societal values have a role to play in the decision.

In the following pages, I will first examine the literature, identifying principles and modeling rules. Next I will identify the key elements of decision maker behavior, based on an analysis of recent experimental data. Finally, I will present a model, based on earlier attempts, and in a system dynamics framework, that explores the complex problem faced by the rising rate of caesareans. I conclude by offering some general conclusions about what we know about this sizable intellectual landscape, in light of the distantly related, voluminous literature. And I will offer some prescriptions, based on empirical and model based exploration, about where we are at, and what we still need to uncover, in order to make better decisions.
2. A review of the literature

Traditional approaches to understanding judgment and yes-no detection learning

As discussed above, the “puzzle” discussed herein has to do with uncovering how people learn to interpret multiple cues, or pieces of information, and translate their judgment about the state of the environment into an appropriate decision. As shown in Figure 1, our learning is based on a series of steps. How does a doctor learn to interpret the symptoms (cues) of a laboring woman and translate them into a judgment about the health and safety of her and her child? What prior experience does a physician bring to this particular laboring woman, and how does that shape his judgment? How does a doctor decide whether that judgment falls above or below his threshold for action? And once that decision has been made, how does a doctor interpret the outcome feedback about his decision – how does he alter his threshold or translation of the cues?

Figure 1: Multiple cue threshold learning

This section is divided into two pieces – first we will examine how scholars have traditionally examined judgment, beginning with the work of Egon Brunswik and the consequent establishment of two crucial tools: multiple cue probability learning (MCPL) and lens models analyses. The second section examines threshold learning, including both signal detection and Taylor Russell analyses.
2.1. Judgment

2.1.1. Understanding the landscape of learning: Brunswikian psychology

The learning papers of Egon Brunswik provide an initial foundation, as his work represents the beginning of a philosophical tradition that would contribute to the study of learning. For the purpose of this review, it is the lens through which all later work will be interpreted. Brunswik’s general approach considered human judgment and learning under conditions of uncertainty using information about the environment provided by multiple ambiguous and fallible cues. He was concerned with representativeness, asserting that experimental conditions must be constructed so that they enable generalization to a particular condition. In other words, Brunswik’s experimentation carefully sampled the cases to be judged to represent the distribution of the real ones, so that his results were generalizable to other samples of cases from the same population.

Brunswik sought no less than a complete revamping of psychological understanding of learning. His philosophy is founded on a perception of the learning organism as one that must necessarily interpret a tangle of cues indicating (one or many) means of achieving (one or many) goals. Such a learning organism, be it lab rat or human politician, must venture guesses, based on invalid or ambiguous cues, about the nature of a particular means or mechanism, and about the probability of achieving the goal. The learner must judge how well any particular fallible indicator predicts progress towards a goal. In Brunswik’s (1935) words:

If, therefore, it is to be successful, the organism must eventually develop both cue systems and means-objects, which are, at one and the same time, both wide and inclusive and yet full of very fine discriminations. (From *The Organism ad the Causal Texture of the Environment*, in Hammond and Stewart, 2001, p. 31)

Brunswik completed two papers (1939, 1951) that specifically reference “learning”. *Probability as a Determiner of Rat Behavior* (1939) represents his very earliest attempt to shape research questions into recognizable theory. The particular paradigm used – rats in mazes searching for rewards and avoiding potential “punishment” – is worth noting. Brunswik, unlike many of his peers, was interested in the ability of an organism to learn to discriminate between probabilistic rewards, and so his rats faced a task qualitatively different from those used in traditional psychology. In this particular series of experiments, he was interested in finding out whether these organisms would be able to learn to identify which side of a maze is the more frequently rewarded one – “or, in short, to distinguish a better chance of reward from a less good one.” (Brunswik, 1939, in Hammond and Stewart, 2001, pg 196) He established that there is a clear “threshold of certainty” below which, learning, at least in the case of these rats, was unclear. In these particular experiments, probabilities of both 100:0 (or 100 percent chance of finding food on side A and 0 percent chance of finding food on side B) and 75:25 exhibited learning
behavior, whereas 67:33 (or two thirds of the time side A would be rewarded and one third of the time side B would be rewarded) failed to produce evidence that the rats were learning which side was more profitable.

Discrimination of the profitable side is correlated not just with the difference of the probabilities, but also with the increasing ratio of probabilities between the sides, as in the increasing discriminability between type 75:25 and 100:50. In other words, rats were increasingly able to discern the more profitable side as the difference between the probabilities increased, as would be expected. They also, however, tended to learn less well as the probability ratio declined. For example, even though the probability difference between the three conditions (50:0, 75:25 and 100:50) were all 50, the ratio decreases from infinity to 3:1 to 2:1, which dampened learning.

Later, Brunswik and Herma (1951) tackled a similar problem, this time in a series of perceptual experiments testing subjects’ ability to correctly discriminate between objects of differing weights, where the relative frequency of positive trials (representing heavy on one hand, light on the other) is manipulated. Brunswik and Herma (1951) find that this experiment seems to emphasize a pattern of rapid “learning” followed by gradual erosion to a “compromise” position. In this particular case, the test was an evaluation of a subject’s ability to use an ambiguous cue as a partial predictor. They found that testing in early trials resulted in behavior that was closer to the true probabilities than the behavior in later trials. Brunswik and Herma suggested that this could either be a paradoxical aspect of multiple cue probability learning, as compared with traditional psychological procedures, or it could be an illusion created by their particular experiment.

---

4 A partial predictor, according to Brunswik, refers to the probabilistic, rather than absolute, relationship between the cue and the outcome. In other words, evidence of a particular cue, may suggest that there is a greater chance at an outcome, but its ambiguous nature does not necessarily predict one
Both studies offer insight of relevance to this review. Their specific empirical questions are driven by an interest in our ability to learn about the nature of a probabilistic environment. According to Brunswik, we continuously translate the information we get from proximal variables, which are often uncertain and ambiguous, to make inferences about distal variables. Indeed, his elaboration of a “lens model” (see Figure 2) that portrays symmetry between the organism and the environment and a “feedback loop” that altered or reinforced behavior, represents one of the most valuable contributions of his work. He concisely articulated his theory of the use of fallible cues to infer a criterion value within a probabilistic and uncertain environment:

The total pattern involved, when viewed as a composite picture of numerous cases of individual mediation from initial to terminal focus bears resemblance to a bundle of rays scattering from a light source and brought back to convergence in a distant second point by a convex lens. (Brunswik, 1952)

In summary, Brunswik’s work is important for at least two reasons. He was concerned about the development of empirical work that was representative, and could therefore be generalized to wider research questions. The judgment and decision making literature is rife with examples of research that has been conducted in a way that makes it difficult to generalize, such as through the use of strictly perceptual experimentation. The degree to which we can generalize from that empirical work obviously limits which insights should be built into a simulation.
Secondly, Brunswik provided an initial map of human judgment through his introduction of a lens model. Through his work, we gain a better grasp of the entire process, from observation of cues, to judgment, and corrective feedback; we now have a philosophical framework for understanding. Those scholars who followed in his footsteps would further quantify this philosophy, into multiple cue probability learning, which provides a tool for experimentation, and lens model analysis, which provides a tool for analyzing the results of those experiments.

2.1.2. Multiple cue probability learning

Multiple cue probability learning (MCPL) is primarily about how well subjects can learn to make inferences about a criterion based on cues that have varying degrees of validity. Traditionally, several cues (usually between 2 and 7 cues) are presented during each trial, with feedback provided either immediately after the trial or after a series, or block, of trials. Feedback can take two forms. Outcome feedback indicates the correct judgment. Cognitive feedback (Hammond and Summers, 1965, 1972) provides additional guidance about the reasons for errors and the possible considerations for improvement. MCPL is a useful addition to any scholar’s toolbox, and especially for those interested in uncovering what’s inside of the black box of human judgment.

The work of Hammond and his colleagues (see, for example, Hammond, 1966, 1971; Hammond and Summers, 1965 and 1972), represents one of the first fruitful explorations of multiple cue probability learning. In a representative early work, Hammond and Summers (1972) used multiple cue probability learning because they found that such tasks were amenable to a great deal of manipulation. Specifically, complexity and uncertainty could be varied in several different ways. MCPL allows for variation in the number of cues, variation in the uncertainty associated with each particular cue, and variation in the functional form of each cue relationship with the criterion judgment of interest.

The usefulness of the MCPL paradigm is intrinsically linked to lens model analysis, as the Lens Model Equation was the breakthrough needed to finalize the difference between the traditional stimulus-response psychologists and Brunswik’s approach to learning in a probabilistic environment. Multiple regression statistics, employed using the lens model, would allow experimenters to employ tasks where there was irreducible uncertainty, and therefore, a “limit to achievement”.

Later MCPL studies considered such complex questions as how judges handle missing information (White and Koehler, 2004) and how we discriminate between the implicit and explicit processes that shape cue learning (Evans, et. al. 2003). Others have moved the use of MCPL from a strictly empirical tool to a modeling framework (Kruschke, 1999). Holzworth (2001) summed up the usefulness of MCPL nicely:

The MCPL research paradigm served its usefulness by providing a means for
demonstrating that Brunswik’s probabilistic theory about cognition and his explicat
of it in terms of the lens model could be directly translated into useful research.

In summary, MCPL allowed for an ambitious research agenda identifying how
judges identify weights and function forms, and how those indicators change over
time, based on feedback. For the purposes of this review, the insights gained based
on MCPL empirical work provides substantial flesh to a still skeletal view of multi-
cue threshold learning.

2.1.3. Lens model analysis

Brunswik offered the beginnings: we should consider the entire spectrum of
influence – from the cues appearing from the environment to our interpretation of
them. A judgment, in this context, is therefore a product of cues and relationships,
and it is only through a careful (and quantitative) assessment of those cues and cue
relationships with each other and with the criterion of interest that we can
understand judgment.

Post-Brunswik, there have been a number of attempts to utilize the power of lens
model assessment. This is obvious, for example, in the rich literature of the 1960’s,
70’s and early 80’s devoted to multiple cue probability learning (Holzworth, 2001).
From a more theoretical angle, scholars have attempted to build on the
underpinnings to this methodological instrument in order to extract more detail, with
the result being the Lens Model Equation (LME) (for an overview, see Stewart,
2001).

The LME is an attempt to extract the statistical properties, through multiple
regression, of the environment and human response process in judgment tasks. It
would aid, for example, in understanding learning in the context of MCPL studies.
Tucker’s (1964) paper, a review of an earlier version of the lens model equation (see
Hammond, Hursch and Todd, 1964), proposed a formulation that would become the
standard. Essentially, it allows for a general assessment of performance, while also
providing a way to analyze the particular contributions of the judges, the
environment, and the relationship between the two.

Stewart (2001) maintained that the elegance and utility of the LME is not simply
due to the nature of the mathematics used, but due to its remarkable ability to
replicate the symmetrical properties of Brunswik’s lens model. In this way, scholars
seem to have come full circle, from Brunswik’s initial proposal of the lens model, to
careful examination and years of empirical exploration, to a return to the simple, yet
elegantly descriptive model of a lens. And it is this picture of a lens which will form
the foundation of a multiple cue threshold learning model.
2.1.4. Choice/decision

2.1.4.1. Signal detection theory

Signal detection theory was born from engineering, specifically the study of electronic signals (Peterson, Birdsall, and Fox, 1954), and eventually reached the field of psychology, psychophysics and diagnostics (Swets, 1996). Briefly summarized, the study of signal detection is concerned with the identification of signals that may or may not be present along with random background noise (Green and Swets, 1966). When faced with a selection or detection decision, we usually consider information provided by cues, make an observation and respond with a decision that indicates whether a signal, or event, is believed to be present or not. Physicians, for example, are often asked to decide whether their judgment of a pregnancy as high risk necessitates a caesarean or not. Errors are inevitable: the decision maker is assumed to be fallible, and the nature of the uncertain environment ensures that even a good decision maker will occasionally be wrong. Recall one physician’s remarks, cited above, that the caesarean rate is “a consequence of individual value-laden clinical decisions.” Although it is difficult to admit, even skilled and very experienced obstetricians make mistakes.

Swets (1991; see also Swets, Dawes, Monahan 2000) considered the identification of “signal” in various diagnostic situations, suggesting that there are several ways that information can aid in the accuracy of decision-making. His thesis, throughout much of his work, is simple: the applicability of various techniques designed to reduce error in diagnosis is unquestionable, as those strategies have undergone rigorous testing and have proven their value – their lack of use in policy is simply a lack of will.

Swets (1991) argued that there are two different types of decision support systems needed in selection and detection tasks: we need to know how to set a threshold and we need to know how to be accurate judges of the information presented through cues. To set a threshold, diagnosticians must assign a value not only to costs and benefits to correct and incorrect identification, but must consider the frequency of occurrence within a population as well; in policy, these considerations are rarely understood and integrated into decision-making. Again recall physician remarks cited above, suggesting that the rise in the c-section rate illustrates a “pattern of almost random decision making”. Moreover, Swets (1991) asserted that accuracy can be enhanced substantially by structuring decision aids that assist a judge in repetitive tasks, such as when radiologists determine whether there is evidence of cancer.

To complicate attempts to understand a decision maker’s behavior in this particular environment, much of the empirical work is based on a single cue. Little research exists that tests a decision maker’s ability to make a find an optimal threshold, based on the judgment of multiple cues. The absence of this type of investigation represents an important gap in the literature, and an interesting springboard for the
research proposed herein.

## 2.1.4.2. Taylor Russell analysis

Another tool that deserves mention, as it is often used in a similar context as signal detection, and due to its unique ability to illustrate decision-making tradeoffs, traces its beginnings back to Taylor and Russell in 1939. The purpose of their brief paper was straightforward: the correlation coefficient as the statistic of choice to measure the strength of a relationship between two variables was insufficient. In its place, there were already several statistical constructs, all, according to Taylor and Russell “relatively simple functions of the Pearson r”.

Moreover, the traditional methodology for interpreting the correlation coefficients is ineffective at drawing a true picture of the interplay between the environment and the judge’s decision – it was ineffective at portraying the trade-offs. Taylor and Russell proposed an alternative that is more explicative, and deeply intuitive.

In Taylor and Russell (1939), the authors explored their proposal, using employment as an example. As shown by their diagram (see Figure 2), those individuals who are above the SS’ line, area A and D, are considered satisfactory, while those below it, in areas B and C, are deemed unsatisfactory. The TT” line represents the selection ratio, or threshold, which divides the applicants who are hired from those who are not.

![Figure 3: Taylor and Russell (1939) identifies the selection ratio](image)

Of critical importance is Taylor and Russell’s (1939) construction of a framework to assess the duality of error – how an uncertain environment necessarily leaves a decision maker in the position of having to choose between four exhaustive and mutually exclusive categories. As shown in Figure 3, there are two incorrect categories, either a false positive (B) or a false negative (D), and two correct
choices, true positive (A) or true negative (C). To illustrate, Stewart, Nath and Mumpower (2002) used the Taylor Russell framework in the context of affirmative action. The authors concluded that when explored in this fashion, it becomes clear that there is a certain degree of unavoidable injustice (Hammond, 1996), regardless of which particular policy is enacted.

Figure 4: Examining the differences in errors, as the selection rate changes.

Figure 4, above, illustrates the resulting distribution of choices when the selection rate is at .3 (shown on the left) and .15 (shown on the right). As mentioned above, although application of Swets-like aids would improve accuracy (and many activists accuse doctors of simply refusing to use those aids), a shift in the threshold to avoid false positives necessarily incurs more false negative errors. In Figure 4, this results in 22 additional women who would have been candidates for cesarean delivery. Moreover, the lower the threshold, the more intractable the situation; pushing the threshold back up implies that you are risking more life-threatening caesareans for the small group concerned with the burgeoning rate of unnecessary false alarms. Also note that even if physicians were to choose the optimal threshold – in Figure 4 above, the selection rate would match the base rate of .1 – there would still be 56 false positive errors and 56 false negative errors. And, in reality, physicians are moving in the opposite direction, and lowering their threshold for action, in some US hospitals, to nearly 40% of all pregnant women. This strategy results in a high overall error rate, but at a .4 selection rate, only 12 women per thousand were considered misses, or should have received a caesarean section and did not.

For the purposes of this review, however, it is also useful to consider Taylor Russell analysis from a less political, and more structural, perspective. In a situation where the decision is based on a judgment of a continuous variable, replicating a Taylor

---

5 The graphs below are borrowed from an interactive spreadsheet devised by Dr. Thomas Stewart. It allows one to examine the resulting differences in errors (and in correct responses) as the selection rate deviates from the base rate.
Russell framework provides an accurate representation of actual decision maker behavior. Additionally, it does represent the critical point of decision, as indicated by the “decision” line in Figure 3. Just for the simple fact that it neutralizes the continuous – discrete modeling tension, it represents a useful guide for the construction of simulation models.

Even more importantly, the Taylor Russell framework is a philosophical construct that quantifies the inevitable error that results from judgment in an uncertain environment. It allows us to take one step past detection theory: with Taylor Russell analysis, we are able to explore a decision maker’s strategy for finding the optimal threshold in a continuous, rather than discrete, learning environment, and understand the consequences inherent in the inevitable trade-offs.

### 2.1.5. Summary

These tools have a two-fold importance to this particular quest. A simulation model constructed to reproduce multiple cue threshold learning behavior is, first and foremost, not starting from scratch. Although the particular tools described here are designed to support empirical testing, they also provide a roadmap for model construction. And as such, these tools help to specify a model that is representative of an uncertain environment, with ambiguous cues and cue relationships. The simulated decision maker, in such a model, must have a shifting threshold that is sensitive to feedback about payoff and about information regarding the task’s implicit uncertainty.

Moreover, as will be shown in the following pages, these tools have been used in research for decades. Laboratory investigations have accumulated insights about how human decision makers learn to navigate our uncertain environment. So in addition to constructing a model that can mimic the environment this study should be able to reproduce the human behavior observed when the tools described above have been applied in the lab.

### 2.2. Reviewing the literature: what do we know?

Brunswik’s theory and the various tools he used to study judgment and decision making have shaped both subsequent research questions and the style researchers used to investigate them. MCPL research, for example, is primarily concerned with the ability of the judge to learn to use cues to make accurate judgments – studies based on research questions that investigate the influence of various types of feedback abound. As mentioned earlier, MCPL is also a tool built to support exploration of judgment changes over time. Signal detection theory, and Taylor Russell analysis, are used to describe the ability of a decision maker to find an optimal threshold, and so how and why people adjust their threshold remains a key focus of investigations that utilize SDT and Taylor Russell. Although all add additional substance to a multiple cue threshold learning model, those tools that can provide details about the behavior over time are most critical.
In general, the empirical work tends to be somewhat disconnected, with scholarship focused on either judgment or decision making, but not both judgment and decision-making. There are, however, some compelling areas of overlap: both are obviously concerned with accuracy – either one’s ability to correctly interpret cues, or our ability to correctly set thresholds for optimal performance. Both are interested in how feedback affects performance, and similarly, how information about performance affects a judge’s confidence. And scholars from both perspectives have become increasingly interested in understanding the mechanisms that shape learning, including those that are a product of clear and explicit manipulation of variables, and those that are subtler indicators of an implicit process. Each of these areas will be examined in turn, with careful attention paid to the similarities and differences in empirical findings produced by the different traditions.

2.2.1. Accuracy and the influence of feedback

2.2.1.1. Judgmental accuracy and feedback

The work of Hammond and his colleagues represents some of the most exciting, and, for this purpose, applicable empirical work. Hammond (1971), Hammond and Adelman (1976), Summers (1962) Hammond and Summers (1965; 1972) represented a series of works committed to answering some of the crucial questions posed by Brunswik. As a collection, they provide a fascinating view of the academic landscape, post-Brunswik.

Several representative works will be explored below. For example, Summers (1962) delved deeper into the relationship between ecological validity and functional validity. In other words, one might assume that “learning” would, in this environment, mean that the relationship between the cue and the judge’s response (functional validity) would tend, over time, to mirror the relationship between the cue and the environment (ecological validity). Summers (1962) devised an experiment using 9th grade students, testing them on their ability to learn to use three different cues – color, angle and hue of a triangle – to determine the correct length of a line. Most of the experiment offered feedback. He concluded that the results indicate that subjects will respond to multiple cues at the same time – or, in other words, be capable of learning from multiple cues, and adjusting their cue weighting to approximately match the weight placed on cue’s provided by the environment. This is especially relevant to the construction of a multiple cue threshold learning model, as it utilizes the framework described above – Brunswik’s theory and MCPL – and provides some formal basis for understanding the behavior of judges in this environment.

Hammond and Summers (1965) continued work in the same vein. Here they scrutinized the proposition that nonlinear cue learning follows the same learning
path as linear cue learning, noting that few others have actually investigated whether judges are capable of utilizing nonlinear cues in making inductive inferences. They insisted that it remains to be seen whether judges who ordinarily utilize linear cues in a linear fashion would also utilize nonlinear cues in a linear fashion as well. And, even more importantly, can they use both linear and nonlinear cues simultaneously?

Hammond and Summers (1965) devised an experiment to test the ability of judges to use both types of cues, using a two-cue task, and a pool of subjects split into three groups. Group 1 was told to predict based on the two cues; group 2 was told to predict based on the two cues, and that there were linear and non-linear cues. Group 3 was provided information about which cue was linear and which was nonlinear, and information about the form of the relationship. Hammond and Summers (1965) concluded that, as suspected, human decision makers will utilize both types of cues. Moreover, depending on the level of information provided about the task, they would effectively utilize each type of cue and garner almost equal accuracy.

Finally, Hammond and Summers (1972) proposed the concept of cognitive control, asserting that not only is knowledge acquisition essential to learning, but the ability to apply knowledge is also a critical component in cognitive tasks. The authors evaluated some of the diverse empirical work, including MCPL experiments, clinical judgment trials and within the interpersonal conflict thread, and suggest that it is possible to “disentangle” cognitive control and knowledge, and gain a better understanding of why failures in accuracy occur.

Furthermore, Hammond and Summers (1972) discussed the possibility that outcome feedback might actually hinder performance. In its place, and much more in line with the spirit of Brunswikian theory, it might be better to provide the sort of feedback that might help judges do better – feedback that allows for greater cognitive control. Taken as a whole, these three works provide important contributions to our collection of important modeling principles: judges are capable of scrutinizing multiple cues, and extracting some of the key information, such as the different cue weights that linearly predict the criterion, and the functional relationships between the cues and the criterion. Importantly, this collection of works suggests that given appropriate forms of feedback, judges can make inferences that will assist with greater accuracy over time. It is also important to the construction of a simulation model of multiple cue judgment, as it suggests that a model should show different behaviors, depending on the degree of cognitive control offered.

Lindell (1976) suggested that although the work of Hammond and Summers (1972) uncovered important insights, there had been insufficient testing of those hypotheses. Specifically, earlier works (see Summers, Tallioferro and Fletcher, 1970 and Deane et al., 1972) may have supplemented outcome feedback, but not provided the critical pieces that would allow judges to accurately compare their judgmental system to the task environment. Lindell (1976) argued that to “get through”, judges need the comparison between their predictions and the outcomes
fed back by the task environment, and so tasks must be designed in a way that allows the judge to make one to one comparisons - the judge’s weights to the tasks weights and the judge’s function form to the task function form.

Lindell (1976) hypothesized that cognitive feedback (information about the cue weights and their function form) is superior to outcome feedback (sometimes referred to as knowledge of results), even in task situations with cues of unequal weight and cues that are intercorrelated. An MCPL experimental task was designed to explore performance differences. Students were randomly assigned to receive one of two different types of feedback (cognitive or outcome), with different 3-cue correlations, and with tasks that either allowed for equally weighted cues, differentially weighted cues, or differentially weighted cues with different signs.

Lindell’s (1976) general hypotheses were upheld – performance was clearly superior when cognitive feedback was offered. He offered some additional insights as well: acquisition of information and its application is enhanced by cognitive feedback. Knowledge is acquired at a much greater rate when cognitive feedback is provided. In addition, as noise is eliminated, it becomes much easier to apply task information, which improves consistency.

Einhorn and Hogarth (1978) represented another early study of judgmental accuracy and the effect of feedback. In contrast to their contemporaries utilizing MCPL, Einhorn and Hogarth (1978) replicate an earlier stimulus response experiment, and build a model of confidence based, in part, on those results. In the tradition of Brunswik, among others, Einhorn and Hogarth (1978) maintained that the characteristics of the environment within which an observer makes a judgment is critical to understanding the judge’s behavior. To further the point, they proposed that to understand how people learn about their own judgments, scholars must investigate judgments, actions and outcome feedback as fragmented, but important, pieces of a whole.

Einhorn and Hogarth (1978) questioned why there is such a discrepancy between our fallible judgments and our confidence. Specifically, why do the results of so many empirical studies suggest that decision makers only respond to feedback that indicates they have made a correct (and positive) judgment? (Incidentally, this is of particular relevance to later modeling scholars, as some emphatically believe that the action not taken is of almost as much value as the decision itself.) Einhorn and Hogarth also wondered why it is that the frequency of hits, not the probability or some sort of relative frequency, determines how decision makers shape strategies. They argued that the answers to both lie in our seeming inability to interpret and use so-called disconfirming information.

Of great interest to those retracing the history of human decision making models is what Einhorn and Hogarth (1978) propose as basic modeling assumptions. Confidence is defined by the strength of belief in one’s ability to make accurate decisions. Relative weights for positive and negative feedback act to reinforce
actions and boost (or reduce) confidence. So with weights that sum to one, when
the weight on the positive reinforcing value is greater than .5, positive reinforcing
value will have more influence. Also keep in mind that total feedback is a function
of how many hits and false alarms occur, or the amount of feedback, as well as the
type of feedback.

Einhorn and Hogarth (1978) offered several compelling conclusions. Their model
suggests that in order for feedback to have a negative effect on confidence, it would
need to have greater reinforcing power than its positive counterpart. And although
that can happen, empirical work provides evidence to the contrary – that in most
cases, positive reinforcement is more influential than negative reinforcement. They
also hypothesize about the relationship between increasing number of judgments
made and confidence, suggesting that as judges gain more experience, their
confidence rises. Specifically, Hogarth and Einhorn (1978) suggested that the
functional relationship may look something like an “S” curve, with judges gaining
confidence slowly with the first few judgments, then more rapidly, and then slowing
again to a compromise position.

Einhorn and Hogarth (1978) further found that a confident learner is also deeply
resistant to change. So even in the face of negative reinforcement, confidence will
likely reduce the speed of response to feedback, slowing correction in judgment
strategy or threshold placement, and accuracy will suffer as a result.

Almost twenty years later, Harvey (1995) examined accuracy from a similar
perspective. In fact, Harvey addressed a question that also perplexed Einhorn and
Hogarth (1978) – what accounts for the unsystematic, or “noisy” error in judgment?
Does the level of predictability within the task itself predict the level of judgmental
inconsistency? Harvey (1995) examined, in turn, some of the more popular
theories. For example, Hammond and Summers (1972) propose their theory of
cognitive control – that incorrect judgment are not errors in reasoning, per se, but an
expected result of using one’s own faulty memory of what is the best strategy.
Another proposal suggests that the sheer volume of information required to be
maintained within short term memory overloads the ability of most judges, and
hence they fail to predictably make correct judgments. These memory theories of
error or noise in judgment have been proposed, despite the fact that investigators are
unclear about the actual psychological process involved and do not assume that
individuals use short term memory to statistically assess weights over time. Another
hypothesis suggests that the variation is due to recursive weighting, where weights
are assessed, and then adjusted over time. Harvey (1995) finds that none of these
offer a sufficient explanation for judgmental inconsistency.

There is also the possibility that people are trying to understand the correct cue-
criterion correlation, and are using an anchor and adjustment strategy to get there,
although where the anchor is initially set remains a question. Lastly, Harvey offers
Brehmer’s (1978) rule switching hypothesis, where strategies are tested and either
found to be acceptable or not, and consequently chosen or rejected. Judgmental
inconsistency will result from the rejection of appropriate strategies, due primarily to the uncertainty in the task environment.

To identify which theory best fits actual behavior, Harvey (1995) used time series data to explore the relationship between task predictability and judgmental inconsistency, proposing that the two are proportional. He also suggested that a simple anchor and adjustment heuristic is used, but it alone cannot account for the variation in judgment. Harvey (1995) concludes that people add noise to their judgment in an attempt to simulate the pattern they perceive in the stimuli. He proposed that this attempt to explicitly represent the randomness in the data with every judgment results in inaccurate judgments. The identification of judgment strategies such as anchor-and-adjustment behavior in response to feedback, and attempts to replicate a pattern that includes some amount of perceived randomness are crucial to this review because they represent transparent psychological mechanisms that can be realistically modeled.

Balzer et al. (1992) reconsidered Hammond and Summers (1971) cognitive feedback proposition, and decomposed cognitive feedback still further. Specifically, they proposed that cognitive feedback can be decomposed into task information, cognitive information and functional validity information. Task information alone can provide a range of valuable cognitive feedback, including such as the multiple correlation between the cues and the criterion, the individual cue-criterion correlations, and the intercorrelations between the cues. Cognitive information provides information about the relationship between the judge’s strategy and the cues, indicating, for example, the degree of judgmental consistency. Functional validity can also be decomposed further by providing information about the correlation between the subject’s judgment and the actual criterion.

Balzer et al. (1992) tested 133 undergraduate students, requiring them to predict baseball team wins based on team statistics. Participants completed the initial questionnaire without any feedback, then returned one week later to retake a reordered task, this time with some form of cognitive feedback provided. The authors concluded that feedback about task information, such as the correct relative weights, was the component most likely to increase overall accuracy. This simple, yet important conclusion cannot be overstated – as has been suggested in earlier works, the construction of a model that can successfully replicate human behavior, and for the right reasons, must integrate these principles of behavior. Accuracy, it appears, is a function of feedback. And the type of feedback provided is clearly relevant.

Allan, Siegel and Tangen (2005) have examined accuracy in the context of a “contingency assessment task”, with the goal of merging the interests of two separate literatures: signal detection theory and contingency learning. They rightly point out that contingency assessment is remarkably similar to the typical signal detection task: cues are observed and then an outcome is presented. The goal is to judge to what degree the outcome is dependent on the (uncertain) cues. Allan,
Siegel and Tangen (2005) were interested in a particular behavioral tendency often linked with problems with accuracy, the outcome density effect.

The outcome density effect suggests that subjects will show improved accuracy in assessing the relationship between cues and outcomes as the strength of the correlation between the cues and the outcomes increases. However, they are also prone to overestimating the strength of those relationships as they are presented with more trials that indicate a predicted outcome, regardless of the presence of a cue. And so accuracy is compromised and decision makers adopt strategies that are sub-optimal. This appears to be identical to the base rate effect observed by judgment scholars. To use a medical example, doctors must often decide whether a symptom, or series of symptoms, indicates the presence of a disease. A high base rate condition, in this example, might occur during the height of winter flu season, and doctors, according to the base rate or outcome density effect, would overestimate the relationship between the cues and the outcome.

This is an interesting phenomenon for at least two reasons. First, it is not unlike the Einhorn and Hogarth (1978) finding that judges are unable to consistently utilize disconfirmatory feedback. Second, when questioned, Allan, Siegel and Tangen’s (2005) subjects were remarkably good at predicting the actual likelihood of an event, which is an interesting complement to both the results found by Harvey (1995) and Einhorn and Hogarth (1978), that judges were typically unable to articulate their strategies.

Allen, Siegel and Tangen (2005) successfully replicated the outcome density effect in their experiment involving undergraduates who judged the effect of fictitious chemicals on bacterial strains: these judges perceived an increasingly strong relationship between the cue and the outcome as the probability of an outcome increased even as the contingency between cues and the criterion did not change. They find that signal detection offers greater insight as it allows investigators to discern between the inherent uncertainty in the task ($d'$) and the location of the threshold ($\beta$). In other words, the outcome density effect can be reinterpreted to mean that a judge’s sensitivity to the task uncertainty is not changing, but as the base rate, or outcome density increases, a judge is more willing to change the location of the threshold. In other words, people raise and lower their thresholds as they perceive a change in the thresholds.

### 2.2.1.2. Threshold placement accuracy and feedback

Among the empirical studies of threshold learning, Green and Swets (1966) conducted one of the earliest, and still most influential, works. They suggest that an individual will adjust threshold in an attempt to respond to accumulating information about the base rate, but will not adjust quite enough. For example, given some outcome feedback, subjects will act conservatively. They will respond by raising the threshold when the base rate appears low, but will fail to raise it high enough; likewise, given some knowledge of results, subjects will respond by
lowering the threshold when the base rate appears high, but will fail to drop it low enough. In this vein, later work by Kubovy and Healy (1981) find that such conservative cutoff placement is fairly common in these types of decision tasks. Defined relative to the optimal decision rule, conservative cutoff placement implies that decision makers will tend to choose a conservative threshold placed higher than optimal when the base rate is low. In fact, conservative cutoff placement is one of a number of sub-optimal behavioral tendencies have been recorded in the literature.

Tanner and Rauk (1970) continued the work of Kinchla (1966) and Tanner, Haller and Atkinson (1967), reviewing the important conclusions related to feedback in an uncertain task environment. This experiment tested the ability of students to discriminate between two auditory signals, given some degree of feedback. Generally, Tanner and Rauk (1970) concluded that base rate matters, in that both the hit rates and the false alarm rates increase as the probability of a signal increases. Of greater interest are the variations in response elucidated by the differences in the type of feedback, which the author proposes fall along a continuum. At one end, those subjects who receive trial-by-trial feedback and knowledge about the signal presentation probabilities (or base rate), result in the best outcomes. At the other end, those who receive no information about the probabilities of signal, and no trial-by-trial feedback, generate the worse outcomes. Between these two points, those subjects who receive no trial-by-trial feedback, but do receive information about the base rate, improve slightly over time. Also between the poles, with results only slightly better than the no feedback condition, are those who receive no information about the base rate, and no trial-by-trial feedback, but do receive a round of “training”, or practice (and so can be considered “experienced”).

Williges and North (1972) concluded from their visual experiment involving brightness testing on 24 undergraduate students that outcome feedback tends to improve discrimination between signals and noise (as measured by d’), but not (at least not directly) the decision threshold (as measured by β). Outcome feedback seemed to make the subject more sensitive to the signal, however, which helped move the decision threshold toward optimal. Interestingly, this effect was seen over the longer term – in earlier trials, knowledge of results had a detrimental effect, with fewer correct identifications and more false alarms. These insights are relevant to any threshold model – we can tentatively assume that decision makers in a probabilistic environment are adjusting thresholds in an attempt to become more accurate, but that they are probably influenced by the type of feedback and its frequency.

Kubovy and Healy (1977) is another one of the key pieces of literature contributing to our current investigation of threshold shifts. This study examined a relatively small group of people (12, over several different sessions, and hundreds of trials per session) in an attempt to uncover evidence for a dynamic cutoff rule. Prior research (see Lee’s micromatching model in Lee 1963 and 1966) had failed to describe a decision rule that predicted learning behavior in probabilistic tasks; this experiment suggested that both earlier modeling attempts fail to account for the sub-optimal
behavior exhibited by subjects. In their place, the authors proposed a dynamic
decision rule: when examining a stimulus in a category discrimination task (such as
whether a particular height indicator more likely represents a man or a women, in
the Kubovy and Healy (1977) task), people will adopt an adjustable decision cutoff,
which will move in response to trial-by-trial outcome feedback about the
correctness of judgment. Much like the earlier cue judgment research, these
particular insights are essential ingredients for a simulation model: they propose the
actual psychological mechanisms that likely shape learning behavior.

Erev (1998) provided a comprehensive review of the empirical studies related to
threshold learning. He also replicated the work of Kubovy and Healy (1977), and
borrows many of their measures, including the use of static cutoff violations and
measures of cutoff shifts before and after correct responses and errors. Nineteen
phenomena observed in empirical investigation are identified and replicated, to one
degree or another, in the cutoff reinforcement learning (CRL) model. According to
Erev (1998), decision makers tend to commit static cutoff violations and fail to find
the optimal cutoff, among many others. Specifically with regard to accuracy,
decision makers exhibit a complicated pattern of toward and away shifts of the
threshold, as feedback indicates whether the previous trial’s response was correct or
not. Furthermore, this pattern was dampened over time. Such behavior naturally
shapes accuracy, but was missing from earlier error correction or additive models of
learning, which allowed only limited threshold shifting (only after errors, not after
correct responses) and ignored the effect of experience on learning.

2.2.1.3. Summary

As shown in Figure 4, perceptions of noise, uncertainty, and the nature of the
feedback itself, all contribute to how well we can use feedback to improve accuracy.

Scholars have found:

- That we can integrate multiple cues into a comprehensive judgment strategy.
  Specifically, we can adjust cue weights (Summers, 1962) and cue function
  forms (Hammond and Summers, 1965). Hammond and Summers (1972)
  and, later Lindell (1976) discovered that the type of feedback and how it is
  presented could predict accuracy.

- Uncertainty in the task environment will affect both judgment of cues and
  threshold placement (Williges and North, 1972).

- There is evidence to suggest that various psychological mechanisms can be
  clearly identified: an anchor-and-adjustment strategy as suggested by
  judgment scholars (Harvey, 1995) supplemented with a desire to match the
  noise perceived in the environment is not unlike the dynamic decision rule
  proposed by threshold learning scholars Kubovy and Healy (1977). Einhorn
  and Hogarth (1978) offer a caveat: adjustment of cue weighting is heavily
  dependent on receiving positive feedback; judges are notoriously unable to
  utilize disconfirming feedback.

- There is also evidence to suggest that decision makers in a probabilistic
  environment are adjusting thresholds in an attempt to become more accurate,
but that the accuracy they achieve depends on the type of feedback and its frequency (Green and Swets, 1966; Tanner and Rauk, 1970).

Figure 5: Feedback's effect on accuracy, and the variables that shape the relationship

### 2.2.2. Feedback effect on confidence in judgment and threshold learning

The effect of feedback on confidence in judgment and threshold learning is at once compelling and inconclusive. The methods used to test hypotheses in this research area do not fall neatly into our tool categories, and so their applicability is not as clear cut as earlier MCPL and signal detection work. Confidence, and the nature of the confidence – accuracy relationship, is a critical component to understanding learning in uncertain task environments, as it is not nearly as intuitive as one might think. In preceding pages, I have reviewed empirical work that attempted to understand under what conditions judges and decision makers will achieve greatest accuracy. In this section, empirical work that focuses on confidence will be reviewed, with the explicit purpose of understanding how it affects a judge’s ability to form strategies. The confidence-accuracy relationship, as I will show, is complex and there are few conclusions available. Most studies address only the effect of over- and under-confidence on accuracy. Over-confidence, for our purposes here, refers to those who believe they are more accurate than they are. Under-confident decision makers believe they are less accurate, or correct fewer times than they actually are.

Sharp, Cutler, Penrod (1988) examined the influence of feedback on the confidence-accuracy relationship, in an attempt to understand to what degree confidence affects accuracy. This question is particularly useful in light of the caesarean dilemma: are do very confident doctors tend to make more errors (and more errors that go
unnoticed, such as false positive errors), and are they more reticent to alter their strategies when faced with errors? The authors proposed decomposition of the relationship, specifically utilizing two elements of the Brier skill score—calibration and resolution, following earlier work (Lichtenstein and Fischhoff, 1977). Calibration refers to the ability of the judge to correctly match probability levels to judgments, in other words, to accurately say that 40% of my binary judgments are likely to be correct. They point out that calibration is related to confidence in that, for example, overconfidence implies that a judge is assigning a probability to a category of judgments that is more extreme than the correct level of probability. Resolution refers to the ability of the judge to respond to correct or incorrect responses, by adjusting confidence levels. In other words, the best possible resolution “score” is attained by correctly sorting all of the correct judgments into one category, and all of the incorrect ones into another.

The experiments involved general knowledge testing of undergraduate psychology students, over the course of four weeks, in a style that may best be compared to a two alternative forced choice task. Feedback was provided to half of the students, indicating performance levels of corresponding confidence ratings. The authors find that their hypothesis that feedback reduces confidence and improves accuracy was generally not supported. They did find support for the claim that the resolution of confidence improved with feedback.

Echoing the earlier work of Einhorn and Hogarth (1978), Pease and Sniezek (1991) were also primarily concerned with understanding the role of confidence in judgment. They devise a judgment task in which college males were asked to determine the earned run average of particular baseball players based on various cues. Three groups were used—a baseline group was given the number of innings pitched, the number of hits per game, and the number of walks given up per game; another group used these cues plus several additional relevant cues; and a third used the baseline cues and several additional irrelevant cues. Each subject was asked to make a judgment regarding the earned run average and provide a rating of his personal confidence regarding that judgment.

The authors offered compelling conclusions. In addition to proposing that confidence is unrelated to accuracy, they suggested that confidence and consistency increase over time, even in the absence of feedback. In other words, even in an environment where we have no knowledge of the results of our actions, we will tend to become more confident and more consistent. Pease and Sniezek (1991) suggested that this is especially relevant in policy environments, where feedback is often delayed or absent.

Subbotin (1996) investigated the effect of outcome feedback on over- and under-confidence in judgment tasks. He proposed that in tasks where the items (trials) are non-independent and related, task information can be accumulated and learning in

---

6 From Brier (1950), the Brier Skill Score represents measure of the adequacy of performance in a probabilistic task.
possible. These types of tasks can be compared to the types of tasks in MCPL and in two alternative forced choice tasks. The authors proposed that outcome feedback will have a corrective effect on underconfidence judgments; feedback will boost confidence and improve calibration (accuracy). However, in the case of overconfidence judgments, outcome feedback will do neither because the two forces that respond to feedback (confidence and bias) will effectively work against each other. In the case of tasks that are unrelated and independent, feedback will improve calibration and confidence when the subject indicates both over and under-confidence.

The particular experimental task involved asking college students to respond to a binary choice question about the size of European cities. Specifically, the capital cities of two nation states would appear on a screen, and students were asked to identify which has the largest population, then make a judgment regarding his or her personal confidence level. An additional task condition split up the subject pool by degree of question difficulty.

Subbotin (1996) concluded that feedback, in tasks where subjects believe they can learn to improve through knowledge of results, is asymmetrical. When subjects are under-confident, feedback will boost confidence and improve calibration. When subjects are over confident, feedback would theoretically diminish their overconfidence (by telling them when they are wrong), but this is muted by the fact that feedback in and of itself will boost confidence. So even when feedback indicates an error, overconfident decision makers will not respond with less confidence, and consequently, there will less improvement in calibration.

Bornstein and Zickafoose (1999) offered similar evidence, although their research was developed from an interest in judicial issues. The authors conducted a series of experiments to gauge the confidence-accuracy relationship in two different contexts—general knowledge (GM) and eyewitness memory (EM). Like earlier authors (i.e. Sharp, Cutler, Penrod, 1988), they measured performance with regard to calibration and resolution.

The specific task involved two speakers addressing classrooms of students. The first speaker made an announcement and introduced the second, who administered the GK questionnaire. The EM questionnaire was administered 2, 5 or 7 days later. The authors hypothesized that their judges would be overconfident in both domains, they would be more confident when they answered correctly than when they answered incorrectly, and that there would be a consistent degree of overconfidence when measured with both absolute and relative measures. Bornstein and Zickafoose (1999) found that their predictions were correct—judges were over confident, and their confidence were consistent across domains, which the authors interpret to mean that there is an underlying mechanism that is affecting both.

A second experiment is more relevant to our investigation. The authors proposed that feedback might temper confidence, so they altered their task to provide
feedback for the general knowledge questionnaire, which would then be followed by the EM questionnaire. They predicted, based on the earlier finding of a correlation between their judge’s absolute and relative measures on the two tests that a provision for corrective feedback on the GK test might also affect the overconfidence exhibited on the EM test. They were specifically interested in identifying whether resolution improved – whether there were correct shifts toward or away, in response to feedback. Unlike the feedback provided in other settings, no trial-by-trial feedback was provided. Instead, each student was placed into one of three feedback groups. Those receiving specific feedback were told that their test had been evaluated and they tended to be overconfident in their responses. The general feedback group was told that their specific test had not been evaluated, but that in general, students tended to be overconfident. A final group did not receive feedback at all.

Bornstein and Zickaloo (1999) concluded, based on this second experiment, that outcome feedback of either variety does reduce the average confidence ratings. It did not, however, help to improve calibration or resolution. So, in general, although we can conclude that there is consistency across domains regarding overconfidence, and that feedback regarding overconfidence will reduce confidence, there is still little far less knowledge about the specific relationship between confidence and accuracy.

Additionally, scholars do not have a great deal of knowledge about how cognitive feedback might affect confidence. If we can generalize earlier conclusions (Hammond, 1971; Summers, 1962; Hammond and Summers, 1965 and 1972), one might propose that specific detailed information may improve accuracy, and with it, confidence will also rise. The question remains: but how do judges identify the appropriate confidence levels?

Fischer (2005) conducted one of the most recent investigations, and considers this issue directly. Unlike earlier work, where the goal is to identify principles that make a complicated relationship produce particular behavior, Fischer (2005) is interested in developing a “model” of the confidence-accuracy relationship. The authors are specifically interested in understanding which types of tasks, and what type of feedback, are likely to induce appropriate confidence ratings in judges, noting that earlier work was specifically interested in understanding the context for such behavior.7

Fischer (2005) specified three tasks. One task, a screening task, is essentially a signal detection task, where judges are asked to place potential job applicants into one of two categories – in this case either the acceptable or unacceptable category. Judges are provided with conditional feedback – when they place a candidate into the acceptable category, they are told whether they are correct (a hit) or incorrect (a false alarm). As the author explains, this task forces a tradeoff between the judge’s expertise goals (which would encourage a strategy of more feedback, and thus more

7 Note that Fischer (2005) cites both Brunswik and Einhorn and Hogarth (1977) in this paper.
potential false alarms) and the payoff goals (which would encourage a strategy that would avoid false alarms). A second task, discrimination, is like the first except that it provides full feedback. In other words, feedback is provided, in Fischer’s example, whether a financial forecaster predicts that the market will go up or down during the following week. Consistent with signal detection theory, our decision maker will find out whether her response is a hit, false alarm, miss or correct rejection. A third task, a classification task, adds another layer of complexity. Fischer (2005) offers an example of a physician faced with a potential serious health issue that could also be two other, less serious, health concerns. The authors assign the potentially serious health issue as the target category, and create a signal detection table with all nine possibilities.

Fischer (2005) argues that decision makers probably re-frame complicated tasks into more simple ones, from classification, to discrimination to screening. He suggests that in the process of simplifying, valuable feedback information is lost, and accuracy and confidence suffer as a result. Performance and accuracy, in this context, degrade as the task’s difficulty increases, but improve with the experience of the judge (number of trials) and the feedback structure (as outlined above).

The author concludes that confidence develops gradually at a diminishingly increasing rate, among all base rates. (Notably, this is contrary to Einhorn (1978), who suggested an “S” shape.) Also, the task matters. The screening task resulted in the least learning, because of the limited feedback, and yet confidence was not impeded. This is consistent with Einhorn and Hogarth’s (1978) illusion of validity hypothesis. As predicted, discrimination tasks led to good performance, confidence, and correspondence between the two. The classification task was predicted to be slower to learn, but the existence of informative feedback would result in overall learning after a considerable amount of training. So, to conclude, when the task is easy, the feedback structure complete, and the decision makers sufficiently experienced, we will witness the best correspondence between confidence and performance. When the task is more difficult, the feedback structure incomplete, and decision makers are comparatively inexperienced, then confidence and performance there is far less correspondence.

2.2.2.1. Summary

In general, we know from the scholarship that has been completed that:

- The confidence-accuracy relationship is complex – and few conclusions have been drawn regarding the effect of feedback. Sharp, Cutler, Penrod (1988) found that their proposition that feedback will reduce confidence and improve accuracy, while intuitive, is not generally supported.
- Even in the absence of feedback, confidence, and its corollary, consistency, will rise (Pease and Sniezek, 1991). Fischer (2005) supported this finding with a screening task with limited feedback, where learning was seriously inhibited, yet confidence was unimpeded.
• Subbutin (1996) found an added complexity: underconfident decision makers will be aided by feedback, and their accuracy will improve. Overconfident decision makers would theoretically find their confidence dampened, but this is muted by the fact that feedback in itself will boost confidence. So those who are overconfident (and most of us are) will not gain accuracy improvements via feedback.

Both of these insights pertain, at least in the empirical work, to judgmental accuracy in multiple cue probability learning tasks. An important gap remains: we can only assume that in a threshold learning task, behavior will be affected by confidence in similar ways. This gap represents an opportunity for future empirical research, and for testing in a simulation model.

2.2.3. Feedback and the role of implicit learning

Another part of the “puzzle” may be that as we are learning (in our experimental settings and out in the real world) how to become more accurate judges and decision makers, and become more confident about our choices, we are establishing helpful rules and guides. Many of those rules may be “learned” without clear knowledge that we are even creating them. In other words, cognitive strategies may fall into two categories – while many of the tasks we engage in on a daily basis compel us to follow a set of explicit rules, other tasks are considered “implicit learning”, and they may be learned below the radar of cognitive awareness. So we improve as we receive feedback, but we may not know why we improve or how to explain the strategy or the rules. In medicine, as well as many other fields, doctors may refer to a gut instinct, an intuition that guides them.

Implicit learning has provoked a great deal of interest in recent years, much of it focused on why we have little understanding of what we are learning, in certain task environments. Unfortunately, much of the work is inconclusive. In Green and Shanks (1993), the authors set out to replicate a prior study (Hayes and Broadbent, 1988), suggesting that different tasks will produce different types of learning, and find little to support that finding. They maintained that there are not different systems of learning, only levels of task difficulty. Lagnado et.al. (2006) derived their hypothesis from an understanding of Brunswikian psychology and utilize multiple cue probability learning as a methodology. They examined the research question: why, even during times of accelerated learning, do people have little understanding of why they are improving. Authors concluded, among other things, that those subjects with the best grasp of their explicit weights are the best performers, but are not much closer to answering their own question.

Dienes and Berry (1997), however, contributed a particularly useful review of the implicit learning literature, and attempts to compile a working definition of what we can consider “implicit knowledge”. Additionally, they were interested in understanding the “features” of implicit learning that distinguish it from explicit learning. Dienes and Berry (1997) identifies three features: there is limited ability
to transfer knowledge from one context to another; it tends to be focused on tasks that involve observation or memorization, not rule-based or hypothesis testing conditions; and implicit learning should be robust, as it should remain more intact over time and it is relatively unaffected by individual differences.

Dienes and Berry (1997) specifically considered the research devoted to the control of complex systems, most closely associated with the sugar-production and person-interaction games devised by Berry and Broadbent and their colleagues (see, for example, Berry & Broadbent, 1984, 1987, 1988; Broadbent, 1977; Broadbent & Aston, 1978; Broadbent, FitzGerald, & Broadbent, 1986). These particular games required subjects to optimize sugar production in a changing environment. Of particular relevance is the fact that these particular games normally induced behavior that suggested learning was occurring implicitly. The several important insights derived from these particular experiments are all reminiscent of earlier empirical work within cognitive psychology. First, when end-of-game reports are structured in a way to “cue up” the type of information used during play (such as giving examples of particular scenarios and asking the subject how she would respond), results are good. This is not unlike Hammond and Summers (1971) discovery that cognitive feedback is normally superior to simple outcome feedback. Generally, subjects will also tend to rely on memorization of particular cue-outcomes in the beginning, and shift to a more general knowledge of how the system works as experience accumulates.8

Later, Hammond (1996) offered a theory of the cognitive continuum that takes this insight further. His definition of analysis and intuition are unmistakably similar to how earlier authors describe explicit and implicit learning, despite the fact that Hammond was not studying human learning, but human judgment:

The meaning of analysis or analytical thought in ordinary language is clear; it signifies a step-by-step conscious, logically defensible process. The ordinary meaning of intuition signifies the opposite – a cognitive process that somehow produces an answer, solution or idea without the use of a conscious, logically defensible, step-by-step process.” (Hammond, 1996, p. 60.)

Moreover, Hammond (1996) argued that this type of intuitive judgment is exactly what Brunswik was articulating with his representation of a lens model; we use multiple fallible indicators to infer information about the environment, often while we are unaware, and make remarkably accurate judgments. In this way, intuition and analysis become more than conflicting constructs – they represent poles on a continuum. Again, Hammond is clear: he is describing ways of judging, not learning. Still, his theory contributes to the current discussion.

Hammond (1996) also proposed that quasi-rationality is the “cognitive compromise” between analysis and intuition, and that it is the cognitive activities and task conditions that induce movement from one pole to the other. He identified several factors that might induce a movement from analysis to intuition: many,

---

8
ambiguous cues, presented simultaneously, not sequentially; a unfulfilled need to define, label and measure the cue values oneself; the lack of a ready organizing principle in order to create rules; and time pressure. Hammond (1996) suggested that errors represent one of the significant reasons for moving from intuition towards analysis.

Hammond (1996) also suggested that the cognitive continuum is mirrored by a correlated task continuum. He identified six characteristics that guide cognitive activity along the cognitive continuum which are not unlike the principles identified by Dienes and Berry (1997) in their study of implicit learning. Analysis will necessarily require a high degree of cognitive control and a high degree of awareness of the cognitive activity. Analysis will encourage less shifting of judgment, and, as would be expected of more analytical judgment, the speed of cognition is slower. Tasks that induce this sort of cognition tend to be verbal or quantitative. On the other hand, intuitive judgment will have both low cognitive control, and a low awareness of cognitive activity. There will be a great deal of shifting across cues, and with great speed. And tasks will often be visually stimulating.

Hammond’s line of research contributed in two ways. First, it is an attempt to uncover the actual psychological constructs of judging humans, which is undeniably helpful. Second, and even more importantly, Hammond (1996) adds detail to the lens model. Based on this expansion of Hammond’s theory, I propose that a multiple cue threshold learning model should clearly replicate the correlation between the judge’s cognitive activity and the task structure, as the basis for understanding learning behavior over time, in a probabilistic task environment.

Unfortunately, due to the apparent gap between literatures, later work has taken insufficient notice of Hammond (1996). Evans and Clibbens (2003) investigated the implicit and explicit learning processes taking place during multi-cue judgment tasks. The authors concluded that both implicit and explicit processes influence multi-cue judgments. They also found that what researchers have until now referred to as "self-insight" may actually just be a measure of how explicit processes influence judgment. Additionally, they found that explicit knowledge acquisition is negatively influenced to a much greater degree than implicit learning processes as task difficulty increases. Finally, they found that explicit processes continue to have a partial influence on judgment even where there is little explicit learning occurring. In fact, one might propose that what they are actually identifying is the quasi-rational learning proposed by Hammond (1996).

Some of the most interesting research, and most applicable to this study, is found in Dalgleish (2006) who, while not focused specifically on implicit learning, theorized that a type of implicit feedback mechanism comes into play when explicit feedback is missing. In other words, when direct knowledge of results is missing, such as when a decision maker is only getting selective feedback, the decision maker will “fill in the blanks” based on experience and prior knowledge of the environment.
Such an implicit value structure would embody many of the principles already uncovered by learning psychologists, such as the propensity to avoid punishment, and the effect of confidence on judgment. Dalgleish (2006) contributed directly to the development of a multiple cue threshold learning model by demonstrating, to some degree, of how confidence might fit with implicit (or intuitive) learning.

2.2.3.1. Summary

The literature involving implicit learning is far from conclusive. It remains, in some important ways, an interesting digression from some of the larger issues within cognitive psychology. For the purposes of this review, some of the tentative conclusions deserve emphasis.

- Although there is (somewhat unfortunately) no clear indication that the current research has been drawn from an understanding of Brunswikian thought, much of it is reminiscent. Dienes and Berry’s (1997) working definition of implicit learning as the type of learning that is, for the most part, inexpressible because of the reliance on complex systems:

  “In terms of controlling complex systems, for example, people can learn to reach and maintain specified levels of target variables without being able to freely describe to others how the decisions should be made.” (Dienes, 1997)

Dienes and Berry (1997) finds that a common strategy is employed by judges in situations where there is an expressed inability to explain how learning is happening. In the beginning, judges will rely on certain cue-outcome combinations, and then later they generalize them to a broader strategy.

- As the purpose of this review is to identify a foundation for the development of a multiple cue threshold learning model, whether classification tasks induce implicit or explicit learning becomes an unavoidable question. As has been suggested here, rule-based explicit learning produces very different behavior than implicit, more intuitive forms of learning, with the more optimal strategy being very conditional on the type of task. (Green and Shank, 1993)

For the purposes of a simulation model, the literature on implicit learning is intriguing. On one hand, we can be tempted to conclude that perhaps modeling psychological mechanisms are unnecessary, as we have yet to uncover exactly what happens to induce implicit learning. Luckily, there have been enough attempts to unlock the black box of quasi-rational thinking (i.e. Hammond, 1996) that we can conclude that judgment, as it moves along a cognitive continuum, is not just acquiring rules and accumulating information. Indeed, a simulation model that provides the most accurate representation of actual behavior would also need to be
constructed on the premise that the type of learning – from analytical (or explicit) to quasi-rational to intuitive (or implicit) – shifts based on how the task characteristics are interpreted.

2.3. Modeling behavior in feedback-driven multiple cue threshold learning models

2.3.1. Modeling judgment (cue learning models)

The empirical insights of the studies reviewed above have inspired a substantial amount of modeling. For the most part, the simulation models born from judgmental accuracy studies are an attempt, first, to replicate empirical findings, but also to accumulate evidence to support certain principles of cue learning. The purpose here is to discuss how simulation modelers have proceeded in the past to extract the principles that regularly occur in the literature.

Kruschke and Johansen (1999) addressed the question of "irrational learning" through a series of simulation models. Their approach to simulation, cleverly names RASHNL (Rapid Attention Shifts "N" Learning), accounts for many of the behavioral tendencies identified in the learning literature. RASHNL is based on their empirical work, which uncovered a number of principles. They find that when cues are deemed increasingly valid, for example, subjects respond with increased utilization; likewise, if a cue is deemed less valid than another source, its utilization will decline. The salience of a particular cue is found to have similar effects. Notably, though unsurprisingly, they find that there is evidence of base rate neglect.

RASHNL is a model that "learns" to associate input values with output categories by computing attention weights. Expressed simply, the outcome feedback shapes how the model treats the cues, and it adjusts to create a better “strategy”. The authors added a mechanism for limited capacity attention and the assumption that learners make large, rapid shifts. A third assumption is that all learning rates are assumed to slow over time: learners adapt to a background level of unavoidable error and start to discount errors. These assumptions provide RASHNL with a sound foundation for comparison: it is able to accurately predict the results of experiments, indicating that the outlined principles seem to correctly capture learning behavior. The authors specifically account for two concepts that prior models do not capture as well - the effect of irrelevant cues and the effect of base rate neglect. Interestingly, RASHNL embodies, to a great degree, what Hammond and Summers (1972) proposed: the type of cognitive feedback that is most important is that which illustrates the form of the functional relationship between cue and criterion, and the weight, or importance of a cue in determining the criterion.

In a similar work, Palmeri (1999) evaluated three category learning models by
comparing them to experimental data. The experiment is unique in that it queries the level of specificity subjects are able to learn, and at what rate. This study is partially in response to the challenge of how to account for the base level effect, whereby fewer errors were made at the more specific level than at the general level. Palmeri (1999) found that the newer generation of models, specifically the ALCOVE model (which is an earlier rendition of the RASHNL model described above) is able to far out-predict earlier modeling attempts. The author attributed this to ALCOVE’s ability to use selective attention to identify association strengths of cues with appropriate categories.

White and Koehler (2004) considered multiple cue probability learning when only partial information is provided. Often, in medical decision making settings and elsewhere, full information is unavailable: decision makers must make judgments based on what they think the value of the missing cue might be. The authors suggested that there are several ways the missing cues can be handled by decision makers, and integrate these strategies into a model. They used the Evidential Support Accumulation Model (ESAM) as a predictive tool. Three experiments are run, all requiring subjects to make judgments based on cue sets with missing information. ESAM is based on support theory (Tversky and Koehler, 1994), which suggests that when asked to choose among several hypotheses, the decision maker typically has one focal hypothesis and several alternative hypotheses. The decision maker is positively biased towards the hypothesis in question, and tends to discount the so-called residual hypotheses. Koehler, Tversky and Brehmer (1997) extended this theory and suggested that as the initial hypothesis gains support, the residual hypotheses are discounted even more heavily.

From a wider perspective, comparison between this work and Kruschke and Johansen (1999) is fascinating. Kruschke and Johansen (1999) proposed a model that allows for rapid and large shifts of attention among cues, limited overall attention, and a learning rate that slows over time. ESAM's difference is only slight, suggesting that it is the vector, or complete hypothesis, that becomes the target, and only feedback early in the learning scheme (provoking a big shift) will cause a change in strategy.

Researchers suggest that if we provide cognitive aids – such as clearly identifying what the weights are – there is often very little learning necessary. And, in most real-life situations, there is some indication of what the weights should be, and decision makers will identify a successful cue judgment strategy early, and stick with it. Moreover, when there is missing information – can we really know if a suboptimal birthing position is a life-threatening situation, even with ultrasound pictures – physicians can shift to other reliable cues, such as fetal heart rate, in order to make sound judgments?

2.3.2. Modeling threshold learning
Researchers have constructed simulation models of threshold learning for much the same reasons as researchers have built models of judgmental accuracy – they
represent virtual laboratories for exploration.

Busemeyer and Myung (1992) represented the start of the newest generation of models that propose to replicate how human decision makers behave when faced with an uncertain environment; specifically, the authors ask: how are cutoff rules selected? The “rule competition” model they offer considers the two tasks faced by a decision maker during a signal detection task: what is the optimum value of specified criteria for each decision rule, and then which rule is most effective?

Their model included both an adaptive learning component and a hill-climbing component. The adaptive model tests each of four “rules”. Rules, in this context, represent threshold adjustment strategies, such as if the judgment is below x, then this observation is noise, or if the judgment is between x and y, then this observation is a signal. The model systematically chooses each rule, and yields an estimate to how close it gets to the optimal value, with the results remaining in short term memory. Results suggest that the rule adjustment model is a simple yet accurate way of illustrating how decision rules evolve in tasks with outcome feedback. Interestingly, the rule adjustment portion of this model will be, for the most part, ignored by later modelers, and deemed inefficient when compared to their own models. It is unclear why this has been done, and most subsequent modeling articles unfairly dismiss the nature of hill climbing as a useful psychological construct.

Erev (1998; 2001) considered the same type of research question as Busemeyer and Myung (1992), but places it specifically in the context of signal detection theory. He proposed the Cutoff Reinforcement Learning (CRL) model as a superior model, suggesting that it will allow for a better fit to the data and more explanatory insight into why decision makers fail to find the optimal cutoff. Reinforcement learning models are based, in short, on the Law of Effect, which essentially states that the more an outcome is rewarded, the greater the probability it will be chosen next time. Erev (1998) was also specifically interested in providing a modeling environment that was superior to earlier generation “error correction” models that failed in one important way: they tended to focus on what happens after errors, but did not allow the learner to adjust after a successful strategy. (This is a point not forgotten for our purposes, as judgment scholars are equally interested in the feedback effect of negative feedback.)

Erev’s (1998) CRL model included several assumptions: there are a finite number of uniformly and symmetrically distributed cutoffs; the DM has an initial propensity to choose each of the cutoffs; and other assumptions about feedback in the form of reinforcement, learning and forgetting assumptions. Simply put, that strategy which is positively reinforced will have a higher probability of being chosen next time. Erev (1998) concluded that three principles have become clearer - that people will adjust their strategies, adjust them in response to incentive structures, and that cutoff strategies are often used.
Barkan and Erev (1998) applied results from Heinrich's (1931) study of industrial accidents to current experimental tasks, considering his assertion that most accidents can be characterized as a probabilistic result of human error. They consider his work an example of signal detection tasks: the environment can be safe or unsafe and the decision maker has the choice to act in a way that is safe or risky. In this paper the authors respond to a plea for more quantification and test four models - traditional signal detection with a static cutoff, a quantification of prospect theory, the CRL model, and the Hill-climbing model.

Results suggest that dynamic models may provide better fit and greater insight that models that utilize a static cutoff. “Dynamic”, in this context, refers to models that allow for feedback over time, creating a non-linear simulation environment. As is probably obvious, the complexity of non-linear simulation allows one to better replicate a learning process that is shaped by feedback, uncertainty, and the type of payoff matrix, among other factors. This type of environment has also allowed for more testing among models, in order to better understand which principles best explain learning behavior. For example, Barkan and Erev (1998) found that the hill-climbing model violates the two basic principles of the CRL model – hill climbing erroneously assumes that a reinforced cutoff may be discarded; and it assumes, by predetermining the learning pace, that stabilization is not a function of accumulated experience. Neither Busemeyer and Myung(1992) nor Erev (1998) varied the base rate or the degree of uncertainty during their modeling simulations.

An additional study by Barkan (2002), did, however adjust both the value structure and outcome feedback. Specifically, Barkan manipulated two variables: in one experiment she changed the payoff matrix to account for unnoticed risky errors, as an analogy to managerial expectations; in another experiment, she manipulated the amount of outcome feedback the subject received, to analogize the effect of supervisory feedback. Both experiments used only perceptual tasks. Barkan (2002) concluded that risky behavior is only partially eliminated when the positive reinforcement for "near misses" is removed: it decreased risk taking and improved the learning process, but only when there were enough accidents; and when simulating supervisory feedback, both penalty and error were able to decrease risk taking behavior.

So, to a great degree, Erev’s (1998) claims have been upheld: people do adjust their strategies (or cutoffs) based on feedback about the incentive structure, and cutoff strategies are often used. There is little to suggest that this cohort of models is optimal, however. Despite the CRL model’s ability to approximately match empirical data, the models themselves fail to represent transparent, psychological constructs, whereas hill climbing does seem to replicate actual decision maker strategy.

Simultaneously, game theorists Camerer and Ho (1997, 2000; Ho, Camerer and Chong 2002 ) proposed an “experience weighted attraction” (EWA) model, a hybrid approach utilizing both principles of belief based models and reinforcement learning
models. Belief based models represent a more cognitive approach to learning, suggesting that experience and beliefs about the knowledge that is being accumulated through feedback shapes learning. Learning is not, in other words, purely a reinforcement mechanism. The EWA has three key features that allow it to merge the critical components of both styles. First, reinforcement is crucial, but unlike traditional reinforcement learning, the EWA also reinforces the actions that are not taken. (Incidentally, the authors argued that this is a much better, real-world interpretation of reinforcement learning.) Second, the EWA allows the reinforcement growth rates to vary, and they are bounded by the limits of the payoff matrix. In other words, the attraction weights are necessarily products of the attraction decay rates, amounts of experience and past attractions. Third, the authors set an initial attraction parameter and identify an experience weight. Perhaps most notably, Camerer and Ho (1997) aim to develop each model parameter as a necessary and realistic interpretation of human learning.

Camerer and Ho (1997) concluded that reinforcement learning and belief based models are not fundamentally different. They illustrate how their model - the EWA approach - provides enhanced fit to data, surpassing reinforcement in all cases, and surpassing belief based in most.

Rustichini (1999) evaluated what he considers a “special class” of learning models: reinforcement learning models in the tradition of Thorndike and the Law of Effect, and more recently, as posited by Erev (1998). In this paper, the author considers the possibility that economists and psychologists may have different motivations in their study of learning models. Specifically, the author is motivated to explore the difference between expecting a subject to be “rational in her actions versus being rational in her procedures”. Similarly, the author considered whether the familiar models of reinforcement learning only treat payoffs as a source of motivation, even though subjects may actually be treating it as a source of information as well.

Chen and Khoroshilov (2003) asked how subjects learn under limited information in the laboratory. This paper evaluated three learning models and examines them against new data from two experimental games. They evaluated how well three learning models can replicate data - a simple reinforcement learning model, a modified experience weighted attraction model, and their payoff assessment learning model. The authors found that the payoff-assessment model, characterized by a strategy of hill-climbing, which allows the player to settles on a strategy that is optimal, or in the proximity of optimal, fairly quickly. The EWA model, followed by the reinforcement learning model, were next best. The authors suggest that this is because these models keep the possibility of other alternatives alive, and thus take a lot longer to settle on a strategy. They concluded, "To capture how real humans learn in such an unstable environment, we speculate that the direction and magnitude of a learner’s experimentation (or the noise term) should be a function of the payoffs received from each strategy. However, this aspect is not captured by any of the learning models we are aware of." This claim, that other models don’t settle as quickly, is reminiscent of the challenges raised by the cue learning theorists –
Chen and Khoroshilov (2003) suggested that the payoffs provide quick enough feedback to encourage real learners to settle on a threshold fairly quickly.

Maddox and Bohil, (2002; see also Markman, Baldwin and Maddox, 2005) focused on optimal categorization, which assumes many different information states but only two potential decisions, requiring the placement of the decision threshold along some relevant dimension. As discussed at length above, the base rate and the payoff matrix affect the location of the optimal criteria, or threshold. The authors suggested that there have been few other studies that manipulate both the base rate and the payoff matrix, despite its applicability to real-world scenarios. Their model is based on two principles: the flat maxima hypothesis, which is a response to the observation that subjects have a tendency to choose a more conservative cutoff than is optimal, and the COBRA (Competition between Reward and Accuracy) hypothesis, which is in response to the observation that subjects choose more optimally in unequal base rate than in unequal payoff situations.

The authors found significant support for the flat maxima hypothesis, which they tested alone, in addition to combining it with the COBRA hypothesis to produce several hybrid models. They found that it is clearly more optimal than the classic optimal classifier. The authors preformed a simple test of the Busemeyer and Myung (1992) Hill Climbing model and the Erev et.al (1998, etc.) CRL model and found that they predicted well, though not as well as their model. They also did not adjust parameters, as they assumed was done in the original modeling work. They suggested that far more research is needed and a more elaborate comparison between these models is necessary.

**2.3.3. Summary and modeling insights**

Cue modeling and threshold modeling research have remarkably little overlap – their separate literatures draw almost exclusively from different sources.

Judgment modelers have found:
- Judgment models that allow for shifting attention among cues, and include a mechanism that allows the model to search for optimal cue weights provides a good match to empirical data (Kruschke and Johansen, 1999; Palmeri, 1999).
- Others (White and Koehler, 2004) proposed a slight variation, suggesting that evidence is accumulated over time that supports a particular strategy, with heavy discounting happening to discarded strategies.

Threshold learning modelers have found:
- Hill-climbing, sometimes maligned, still remains a transparent psychological structure, and linked with Busemeyer and Myung’s (1992) rule adjustment mechanism, was one of the first to offer a model of threshold learning.
- Erev (1998) and Barken (2002) propose a type of reinforcement learning model, building on earlier generation error correction models, finding that
people do adjust their thresholds in response to feedback. Camerer and Ho (1997, 2001) suggest that experience weighted attraction, a hybrid approach embracing both belief based and reinforcement models, is superior.

- Chen and Khoroshilov (2003) suggest that each of these earlier attempts discount the decisiveness of early strategies: decision makers, they assert, often settle on a strategy fairly quickly. Maddox and Bohil (2002) suggests that hill climbing, as well as competition between reward and accuracy, shape learning strategies.

Despite the (generalizable?) insights gained from each of these modeling threads, there has been little attempt to draw them together. And after a thorough review, one wonders about the usefulness of continuing to model judgment and threshold learning separately. Is it possible to create one modeling strategy that absorbs both? I will attempt to draw together the work from different domains, extract the most relevant kernels and propose the key principles needed to construct a multiple cue threshold learning model.

3. Key elements to include in a model

Elements of a multiple cue threshold learning model are derived from an understanding of the tools used to explore learning in a probabilistic environment and from the insights accumulated from years of empirical testing and model building.

A thorough review of the literature has uncovered important modeling components in three overlapping categories – how feedback relates to accuracy, how feedback influences confidence (and its effect on the confidence – accuracy relationship), and to a lesser, but still important degree, how feedback affects implicit learning. A multiple cue threshold learning model should embody the following key principles, as extracted from the empirical literature:

- The model will have multiple cues – and simulated learners should “learn” to use those cues. (Summers, 1962; Hammond and Summers, 1965). Outcome feedback will usually mean quicker, more accurate threshold learning (Kubovy and Healy, 1978). Cognitive feedback will allow for better disentanglement of the cues (Hammond and Summer, 1972; Balzer, 1992). Taken together, these principles represent an essential modeling concept: we can assume that judges are continually hypothesizing about the difference between the current strategy and the strategy suggested by the feedback, and constantly adjusting to improve. (This is essentially what Busemeyer and Myung (1992) have proposed in their rule-testing hill-climbing approach.)

- The model will utilize the lens model as the backbone of the model (Hammond and Stewart, 2001; Lindell, 1976; Hammond, 1996)

- The model will accurately simulate the behavior of judges, who react differently to confirmatory versus disconfirmatory feedback. (Einhorn and
Hogarth (1978) This element is related to confidence: feedback will act to boost confidence, but does little to reduce it.

- Similarly, a model’s simulated threshold is based, in part, on the decision maker’s level of confidence (Pease and Sniezik, 1991; Subbutin, 1996; Einhorn and Hogarth, 1978; Fischer, 2005).
- This model should embody the complexity of implicit, or intuitive learning, versus explicit (or analytical) learning, which is, in part, a function of the task structure (Dienes and Berry, 1997; Hammond, 1996).

The reason for reviewing the past modeling work devoted to understanding how judges treat cues, and how decision makers set thresholds, is straightforward: some of it has worked well, others not. The goal was to extract the principles that have withstood the test of time (and replication, and testing) and identify them by their merits. After a review of models, I conclude that a multiple cue threshold learning model should embody the following principles:

- Simulated learners will exhibit a shift of attention, rapidly, and sometimes rashly (Kruschke and Johansen, 1999; Palmeri, 1999) and will accumulate evidence to support a particular strategy (White and Koehler, 2004).
- Reinforcement learning (Erev, 1998) and hill-climbing (Busemeyer and Myung, 1992) are instantiated differently, but essentially produce the same result – judges will adjust their cutoffs in the same direction based on feedback about accuracy.
- The latest generation of models suggests that decision maker strategy will stabilize fairly quickly (Chen and Khoroshilov (2003), in part because of a reliance on feedback about reward and accuracy (Maddox and Bohil (2002).

A multiple cue threshold learning model is a substantial undertaking, especially if it can be constructed based on the key principles of learning identified in the empirical work. Yet, to a great degree, it is nothing more than the piecing together of multiple threads of the general learning and judgment and decision making literatures. As an understanding of the complex relationships between feedback and accuracy, feedback and confidence, and feedback and intuitive, or implicit learning, are disentangled, a simple, and accurate, representation of multiple cue threshold learning emerges.
4. Understanding the data

Prior to examining a multiple cue threshold learning model, however, we need some sense of how people might respond to the varying conditions we have alluded to in the literature. As discussed above, there is some clarity about the kinds of strategies we might suspect learners to employ when faced with a complex threshold learning task. Unfortunately, all previous experiments relied almost entirely on a limited number of variables. Specifically, researchers utilized a .5 base rate and with full feedback. They never adjusted the base rates, the rates of uncertainty, nor altered the value structure. And so how learners behave under other circumstances has been a research question that has remained, for the most part, unanswered. Consequently, all of the modeling attempts have mirrored that research, and only rarely adjusted task conditions.

Development of a research framework that considers multiple variables could have real world implications. In the example of physicians, doctors are often faced with the task of uncovering the base rate, by negotiating a landscape of multiple cues and limited feedback. And the value placed on their errors necessarily shapes their learning strategies and overall behavior. It is imperative, therefore, that we gain greater insight into their threshold learning behavior.

So this chapter has several objectives. Based on the new data, we will begin to answer some questions. We know that many real-life examples of complex learning tasks, such as the decision to choose a caesarean birth, have a very low base rate. But will having far fewer “signals” mean that physicians will have trouble learning where to place an optimal cutoff? And if they only receive feedback when they choose not to operate on a birthing woman, will this also hamper their ability to find the optimal threshold? Does the new data add anything to our understanding of how the value structure or changes in uncertainty affect learning?

We can also start to investigate whether theories about cue learning apply when there are several different cues, with different weights, within a more complex learning task. So far, theory regarding the effect of cues and their weights has considered only simple tasks. Lastly, based on an examination of the new data, can we draw any conclusions about the intuition/analysis continuum? Is there any support for the suggestion that there is implicit learning taking place – or any marked changes in behavior that could be attributed to movement along the intuition – analysis continuum?

4.1. The task

The experiment was conducted as a computer simulated learning environment.  

9 Participants were asked to imagine there were customs officers in a busy

9 Experimental data used with permission of Dr. Thomas Stewart, principal investigator.
checkpoint. The experiment consisted of 20 blocks, with 25 trials in each block. Subjects were given three cues – Clothing, Emotionality, and Bulging – and asked to move a slider bar to indicate their judgment. A sample screen is shown in Figure 6.

Figure 6: Sample of subject screen in experiment
Figure 7: Subjects had the option of viewing information about the cue weights

In order to provide additional task information, participants had access to the cue weights, provided as a bar graph (see Figure 7). After the judgment was made, the participant was asked if this particular passenger should be searched. Then, depending on the particular experimental condition, the participant would either be told whether or not she was correct, or that no information was available, and accrues “points”, based on the payoff structure for that particular set of trials. After each block, each participant received information about her judgment accuracy, total points, and bonus points. Similar feedback was provided at the end of the experiment.

Uncertainty was determined exogenously: cues were assigned to replicate a Taylor Russell-type framework, where the $R^2$ statistic could be manipulated. In the experiment that follows, .7 refers to a moderate level of uncertainty (which is treated like a base run), and .5 and .9 refer to more difficult, and easier, levels of uncertainty, respectively.

In the results that follow, graphs will indicate the amount of points accrued over time, with points added for correct responses, and subtracted for incorrect responses. “Average points” refers to the mean amount of points accrued over particular trials, across participants. The percentage of correct responses refers to the fraction of responses, across, participants, that were correct for each block of trials.
4.2. Results from base rate changes

To start with the simplest results, (see Figure 8) when points are averaged across all trials and all subjects, we find that at the commonly tested, and most difficult, .5 base rate, participants scored the least points. The .1 base rate, when only 10% of the trials yielded signal events, led to the highest scores, with the .8 base rate condition a close second.

What does the pattern of errors look like? As shown in Figures 9 through 11, when there are a lot of signal trials, participants will make a lot of false positive (or false alarm) errors. (They will indicate “yes” too often.) When there are very few signals, participants will make a lot of false negative errors (or misses). (They will choose “no” too often.) Moreover, recall that Tanner and Rauk (1970) concluded that base rate matters in that both hit rates and false alarms increase as the probability of signal increases. As shown in Figures 9 and 11, this experiment concludes the same.

This pattern of errors also persists over time, with little indication that participants learn to minimize errors as they gain experience. Figure 11 indicates that the correct responses (true positives and true negatives) have at least two notable characteristics – they fail to increase over time, at least after the first block or two, and there is significant variation among blocks. This variation might indicate an inability to balance the inherent uncertainty in the task with a participant’s confidence in her choice of the appropriate range of thresholds. Or, as discussed in detail earlier, it can be attributed to the finding (see Chen and Khoroshilov, 2003) that decision makers tend to settle quickly on a desired threshold, and learning tapers off quickly. This potential evidence of “sticky” thresholds, or compromised learning, may also be critical for real-world decision makers. Do physicians become comfortable with a certain threshold, perceiving it as the most accurate, and stop learning?

Another interesting result of empirical testing is the obvious difference between initial thresholds, illustrated in each of the figures below, which assumes that there is very fast initial learning happening, and subjects set their thresholds early. As shown in Figure 12, there is an initial block reaction to the environment – learners seem to quickly gain some sort of insight into the approximate base rate, and then head in that direction. The results indicate that there is an attempt to match the pattern of errors – although subjects continued to accumulate errors, they have a general sense of how often they should be identifying positive responses. Or, in other words, they were, to some degree, trying to identify a threshold that would match the base rate.
Figure 8: Overall performance (points) for three base rates, across blocks of 100

Figure 9: False alarm rates, across of blocks of 25, indicating differences among base rates
Figure 10: Misses over all base rates, across blocks of 25

Figure 11: Hit rate for all base rates, across blocks of 25
4.3. Results from changes in feedback

The following graphs address the results from changes in feedback. We might expect more feedback to necessarily induce more learning, for the simple reason that there are more opportunities to correct. Specifically, in the cases that follow, the amount if feedback is either “full”, or provided after each trial, or “conditional”. Recall that in cases of conditional feedback, information about the results of a decision is dependent on the decision. As shown in Figure 13, though, there is little to support that claim. Moreover, when tested at the .1 and .8 base rates (see Figures 14 and 15), participant performance is almost unaffected by the level of feedback. The only exception to this might be made at the .8 base rate, when conditional feedback scores are slightly less than full feedback, from initial trial to the end.

At the .8 base rate, a higher threshold during conditional feedback trials leads to far more misses, but fewer false alarms, leading to a slightly lower hit rate. So, as previously discussed, these are two interwoven issues. First, this is clearly a case of inevitable error – in attempting to minimize one sort of error by only providing conditional feedback, we aggravate another. Acknowledging this, we then need to decide the type of error that is more important, and that depends on the context of the problem. In the case of medical decision making, misses can much more dire consequences than false alarms. For example, excessive false alarms translate directly into an ever-increasing caesarean rate in the US, but still pale in comparison to high and rising rates of birth-related mortality in some developing nations.
These results provide some insight into our learning strategies. Contrary to widely held beliefs, due in large part to the unlikely but often employed .5 base rate, conditional feedback does not significantly impair learning. People are learning, even without explicit feedback, suggesting that some sort of implicit learning is happening under these conditions. It also suggests that early on in the experiment, participants gain an adequate amount of confidence in their decisions, and the inherent uncertainty in the experiment does little to dissuade them from their initial strategy.

![Figure 13: Examination at the .5 base rate](image-url)
Figure 14: Examination at the .8 base rate

Figure 15: Examination at the .1 base rate

4.4. **Results from changes in uncertainty**

The results obtained from the uncertainty conditions are somewhat ambiguous. Figure 16 seems to provide some evidence for the assumption that as the task becomes more uncertain, there will be less (or at least slower) learning. By the end
of the trials at the .5 base rate, there are marked differences among the three conditions, and just in the pattern one might predict. There is little to suggest that the subjects were able to counteract the challenges of uncertainty by employing better strategies.

At high or low base rates, however, the outcome is far less clear. As shown in Figures 17 and 18, there is less evidence of learning, especially after the first two blocks, and little difference between results when the task is made more uncertain. Less uncertainty (an $R^2 = .9$) does produce slightly better outcomes, but more uncertainty does not seem to produce reliably worse outcomes. In 6 of the 20 blocks, for example, at both the .8 and .1 base rate conditions, the higher uncertainty condition produced better higher scores than did the base, or moderate levels of uncertainty.

Moreover, there is little change at all in the patterns of learning between the higher uncertainty conditions at the .5 base rate when differences in feedback are added (shown in Figure 19), with the exception of the fact that the initial learning is delayed until the 3rd or 4th block. The trends then indicate little additional progress is made. This contrasts with results of conditional feedback trials at either the .1 or .8 base rates, shown in Figures 20 and 21, where accuracy can be characterized by an upward trend from the beginning.

\[ \text{Blocks} \]
\[ \text{Average points per case} \]

Figure 16: Different levels of uncertainty ($R^2 = .5$, $R^2 = .7$, and $R^2 = .9$) at the .5 base rate
Figure 17: Different levels of uncertainty ($R^2 = .5$, $R^2 = .7$, and $R^2 = .9$) at the .8 base rate

Figure 18: Different levels of uncertainty ($R^2 = .5$, $R^2 = .7$, and $R^2 = .9$) at the .1 base rate

Figure 19: Examining behavior at the .5 base rate with more uncertainty ($R^2 = .5$), for full and conditional feedback
A final series of conditions investigated the effect of changes in the values placed on errors: one condition placed a greater penalty on false positive errors, and another placed a greater penalty on false negative errors. Based on previous literatures, and some intuitive sense of what should happen, we might assume that larger penalties on particular errors would translate into an enhanced desire to minimize those errors. This is similar, in this way, to the earlier trials, where conditional learning necessarily produced error-minimizing behavior. Figure 22 indicates at least one obvious characteristic: at first glance, the results are nearly the same regardless of the values placed on penalties. The greater question becomes: are there any penalties, under any conditions, that can improve outcomes?

We can begin to answer that question. Figure 23 adds additional complexity – and when faced with conditional feedback, one particular outcome proves interesting.

Figure 20: Examining behavior at the .8 base rate with more uncertainty (R^2 = .5), for full and conditional feedback

Figure 21: Examining behavior at the .1 base rate with more uncertainty (R^2 = .5), for full and conditional feedback

4.5. Results from changes in values
Recall that at the .1 base rate, participants learn quickly that most trials are noise trials. In other words, during most trials they would choose “No” and opt out of further screening. They also, as shown earlier, make a lot more false negative decisions. But penalizing them for making those excessive false negative decisions, or misses, results in fewer correct responses throughout. Participants are not able to receive this feedback as corrective (as suggested by Rustichini 1999) and thus fail to adequately correct their threshold.

Furthermore, fewer correct responses do not correspond equally to fewer points, as discussed earlier by Maddox and Bohil (1998), leading to a classic example of decision makers forced to balance a need for accuracy with a need to score higher. As shown in Figure 24, when we penalize false negative errors at the .1 base rate, and compare full and conditional feedback, we find that participants minimize their errors more, and become more accurate, when receiving only conditional feedback. But this “correction” leads to fewer points, as shown in Figure 25.

Examining the error rates directly provides some insight. Recall that even in the most basic scenario (see Figures 11 and 12), there is a pattern of fewer false alarms over time, as the participants gain experience at the .1 base rate, but there is little evidence of an attempt to also minimize misses. (Note, too, that in a block of 25 there are also few opportunities to actually improve, as there are so few signal trials, and this would explain some of the erratic behavior shown in those graphs.) During these trials, full or conditional feedback, with a penalty on misses or a penalty on false alarms, will enhance a participant’s ability to reduce false alarms over time. And, as shown in the graph, subjects behave predictably: a penalty on false positives has the greatest effect when combined with full feedback, and conditional feedback, especially when combined with an error on false negatives, has the least effect on minimizing false positive errors. But none of the conditions shape behavior in a way that alters the rate of false negatives, or misses. Again, there are so few of them, especially when faced with conditional feedback, that it is understandably difficult for subjects to reliably reduce their rate of misses.

Clearly, how one interprets these results depends on what, exactly, is being accomplished. If we apply the caesarean section example, the question becomes how do we reduce infant and mother morbidity and mortality, while protecting as many mothers and infants as possible from unnecessary surgical deliveries? This data suggests that even with penalties, reducing the false negative (or miss) rate, for problems with a very low base rate, is difficult. Perhaps as a reaction to that knowledge, physicians have continued to lower their thresholds, as misses are perceived to be a far greater threat to lives than false alarms, or unnecessary caesareans, are. So even with the pressure of a heavier penalty on false positives, it is, and will continue to be, difficult to reduce the false positive rate at the .1 base rate, especially considering physicians are only receiving conditional feedback. From a policy perspective, stemming the continued rise of caesarean births may prove to be an intractable problem, as moving the threshold back up will be interpreted as risking more missed caesareans.
Figure 22: Examining behavior at the .1 base rate, with full feedback, and differences in penalties

Figure 23: Examining behavior at the .1 base rate, with conditional feedback, and differences in penalties

Figure 24: Examining behavior at the .1 base rate, with full and conditional feedback, and penalties on false negatives, by rate of false negative errors
Based on this data, we can make the following observations.

- The data indicates that base rates matter. Base rates that are higher and lower than .5 produce behavior that suggests that there is either less learning happening, or that decision makers choose a selection rate that approximates the location of the optimal threshold fairly quickly. A model should be able to replicate that finding. The model should also be able to replicate the patterns of errors. When there are a lot of signal trials, participants will make a lot of false positive errors and when there are very few signals, participants will make a lot of false negative errors (or misses). The greatest learning seems to happen in the first block or two, and there can be significant variation among blocks.

- The data also indicates that feedback effects are more complicated than previously assumed. The tests above indicate that there tends to only be slight effects in overall score differences, mainly because of the issue of inevitable error. There are shifts in the numbers of false alarms and misses, which the model should replicate. Additionally, it provides some insight into our overall learning pattern: participants were clearly learning, even without full feedback, suggesting that there is implicit learning happening. The data also suggest that people learn quickly, settle on an approximate selection rate, or threshold, and then learning ceases.

- Uncertainty changes are far less clear. At the .5 base rate, there are marked differences on learning behavior across the uncertainty levels, which matches prior assumptions about the influence of uncertainty on learning. Differences in uncertainty at the .8 and .1 base rates produce ambiguous
behavior; it is far from clear that a higher level of uncertainty will create a more challenging task. Again, the model should be able to replicate these findings.

- Asymmetrical values, or the placement of higher penalties on errors, allows us to investigate how we can shape learning behavior by encouraging the participant to minimize error. The question here - of which penalties, under what conditions, can improve outcomes – must also be replicated in the model environment. Just penalizing misses as a way of improving accuracy does not significantly alter behavior, as there are fewer correct responses. (Though it is equally important to note that fewer correct responses do not necessarily mean a lower score, in this data.) Even more importantly, when applying examples, it becomes clear that at the .1 base rate and conditional feedback, reducing misses is “easier” (at least in this laboratory setting) than reducing the false alarm rate, which is much more intractable. It is critical that the model be able to reproduce such selection rate changes.

If the earlier literature review were a foundation, a skeletal framework, then the data serves to add more flesh. The model serves as an additional tool – a way of interacting with the data in a way that provides more explanatory power and allows us to tell a fuller, richer story.
5. Model Construction

So far, I have explored a wide swath of judgment and decision making literature, examining both empirical studies and models. I have also dissected a recent data set, leading to some general conclusions about threshold learning under complex conditions. A model, at this point, will allow for rigorous theory testing, as well as enabling further experimentation beyond that which was conducted in the earlier laboratory experiment. This chapter will provide details about model construction, identification of variables used, a thorough explanation of the model's “learning” process and a summary of test variables used to judge model behavior. It will first identify the specific model sectors, exploring the rationale for particular structures. Then, I provide a brief explanation of the model process, retracing a sample model trial.

Generally speaking, the structure adopts an anchor and adjustment strategy, and is partially a product of Busemeyer and Myung's (1992) hill climbing model as well as Erev's (1998) cutoff reinforcement learning model. Like these earlier models, the cutoff is a function of performance-related outcome feedback. Unlike these earlier models, the model will be able to be manipulated across 4 different experimental conditions – base rate, feedback, payoff structure, and uncertainty. The assumed goal of the decision maker is to find the optimal threshold, while also conserving an awareness of how the uncertainty in the model, the base rate, the value structure and the type of feedback provided shape learning behavior. To briefly revisit the goal, Figure 26 illustrates the standard signal detection construct, with a line indicating the optimal threshold. The actual decision depends on whether the judgment is above or below the threshold. Two kinds of errors can be made: rejecting a signal event (a false negative error) and accepting a noise event (a false positive error). Placement of the threshold determines the relative proportions of these two errors (as well as the proportions of correct decisions). Following earlier model builders, I hypothesize that the model responds in a way that is similar to an actual decision maker, in that it attempts to move closer to the optimal threshold by adjusting based on feedback.
To be useful, a model ought to be more than the sum of its parts. Indeed one goal of system dynamics modeling is to identify not only a structure that can approximate the patterns of actual performance, but also to shed some light on the causes behind behavior. Although later chapters examine behavior in great depth, it is also important to have an understanding of the model learning process.

The model, in its simplest form, is shown in Figure 27. A cutoff is selected, performance evaluated, and then a new cutoff is selected in an attempt to improve performance. However, as noted above, there are many elements that shape the learning behavior implicit in this simple loop. Most notably, we miss the complexity of the process that happens between the cutoff and the performance.

Figure 28 adds judgment and the cue variables, leaving the model with a more complete environment structure, but still fails to identify how the model gets from a judgment and consequent decision to “performance”.

Figure 29 adds feedback, separating payoff and accuracy from simple “performance”, both of which provide outcome feedback and exert pressure against the cutoff. As indicated, the amount of feedback will also affect the threshold adjustment.
Figure 27: Basic structure of threshold learning model

Figure 28: Basic structure, expanded to include cues, cue weights, and judgment
Figure 29: Basic structure, with influence of feedback on threshold adjustment
5.2. The environment sector

Actual simulation requires more detail than causal loop diagrams, however. As shown in Figure 30, this multiple cue threshold learning model needs a sector (or piece of the model) that provides the exogenous information that will later be translated into a threshold adjustment strategy. Or, in other words, it needs the first part of the lens model – the cues.

In the illustration above, the cues are derived exogenously, from a linked spreadsheet. (An identical model, using randomly drawn cues, also acts as a laboratory for experimentation. Additional information about use of both models is available in the appendix.) They are the identical cues used in the experiment in the previous chapter, allowing for even greater comparison between the model and the experimental data. Each cue has a weight, as some cues are more important, and should be valued differently. The model then constructs a weighted average of the cues.
The payoff sector

Figure 31: Summarized view of the payoff sector

“Payoff” in this case (see Figure 31 above) refers to the process by which responses are mapped to the appropriate payoff values. The simulated decision maker (SDM), after constructing a weighted average for the cues, characterizes the event as either signal or noise, depending on whether the weighted average of the cues falls above or below the threshold. “DM distribution choice” means that if the weighted cue average is higher than the threshold, the SDM chooses signal; if below, then the SDM chooses noise.

The payoff is based on the parameters within the value structure. As shown in Figure 32, this model, in its base run and in many additional scenario tests, utilizes the equal condition payoff structure in Stewart (2007). Later tests examine the effects of greater penalties for false negatives and false positives on SDM learning behavior, especially when combined with such added complexity as conditional feedback and changes in uncertainty.

<table>
<thead>
<tr>
<th>Payoff condition</th>
<th>False Negative</th>
<th>True Positive</th>
<th>True Negative</th>
<th>False positive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Equal (VE)</td>
<td>-50</td>
<td>100</td>
<td>100</td>
<td>-50</td>
</tr>
<tr>
<td>Greater penalty for false negatives (VN)</td>
<td>-100</td>
<td>100</td>
<td>100</td>
<td>-50</td>
</tr>
<tr>
<td>Greater penalty for false positives (VP)</td>
<td>-50</td>
<td>100</td>
<td>100</td>
<td>-100</td>
</tr>
</tbody>
</table>

Figure 32: Table of penalties

As the model “knows” whether a signal was truly presented, it determines how to reward the simulated decision maker for its choice of signal versus noise. So, for example, if the SDM identified the trial as representing a signal event, and the
environmental distribution choice was also signal, the payoff would be the value of a true positive; if the SDM identified the trial as a noise event, and was correct, the payoff would be the value of a true negative. Both false negative and false positive errors are calculated in the same way.

5.3. The threshold adjustment sector

The final sector, the threshold adjustment sector (shown in Figure 33 above), directs the “action” in the model – determining how the simulated decision maker will adjust and correct the threshold to maximize performance. From a technical perspective, the change in the cutoff is based on changes between two sets of variables: following Busemeyer and Myung’s (1992) formulation, I reconstructed a hill climbing function that indicates the direction of the next shift by calculating the product of the difference between last trial's cutoff and this trial's cutoff, and between last trial's payoff and this trial's payoff. This model borrows Busemeyer's “h” function:

\[ \text{chng in payoff} \times \text{chng in cutoff (Busemeyer/Myung h function)} = \text{change in payoff} \times \text{change in cutoff} \]

Adaptation of Busemeyer & Myung's h function only partially provides a foundation for behavior, however: the model presented in this dissertation also relies on accuracy to provide some indication of the size of the next shift. As the frequency of correct responses increases, the desire to alter the threshold declines.

In other words, the model is charged with “remembering” two pieces of
information. One piece is the hill climbing calculation, which determines whether recent changes in threshold and payoff have produced an outcome that encourages movement in the same direction or the opposite direction. The threshold adjustment is a function of size and direction:

\[
\text{IF THEN ELSE}( \text{chng in payoff} \times \text{chng in cutoff} = 1, 0, \text{IF THEN ELSE}( \text{chng in payoff} \times \text{chng in cutoff} > 1, \text{Size up, Size down})
\]

Payoff, in the model, is a function of whether or not the decision maker identifies correctly:

\[
\text{Payoff} = \text{IF THEN ELSE}(\text{Environmental distribution choice} = \text{DM distribution choice}, \text{Payoff for correct identification}, \text{Payoff for incorrect identification})
\]

where

\[
\text{Payoff for correct identification} = \text{IF THEN ELSE}(\text{Environmental distribution choice} = 1, "\text{Value on signal, correctly identified (true positive)}", "\text{Value on noise, correctly identified (true negative)}")
\]

And \[\text{IF THEN ELSE}(\text{Environmental distribution choice} = 1, "\text{Value on signal, incorrectly identified (false negative)}", "\text{Value on noise, incorrectly identified (false positive)}")\]

The second piece is the error correction component, which accounts for the model’s ability to both adjust the size of shifts as well as dampening the effect of hill-climbing when the chosen threshold produces a history of correct responses.

Feedback is a crucial element of this model of detection learning, and to that end, it is constructed so that I can easily test the differences in behavior between full feedback, no feedback and conditional feedback conditions. As shown in Figure 33, “feedback condition” is a parameter allowing the user to identify the particular form of feedback to be generated. Based on that parameter, the threshold adjusts, either responding to no feedback, feedback provided on every trial, or conditional feedback, either only feedback after positive decisions.

Lags, or delays, are equally important. In this particular situation, the model has been initially constructed so that one time step is equivalent to one trial. In other words, every time the model takes a step forward in time, it provides a new trial. There is a small delay on the cutoff adjustment, allowing some smoothing of the cutoff. The delays used to derive the recent cutoff and the recent payoff are both set to 1, in keeping with Busemeyer’s original formulation. Practically speaking, a 1 time step delay can be interpreted to mean the delay between “last trial's endpoint” and the “last trial's payoff”. Lags and delays provide useful insight, and will be described in further detail in pages to come.
6. Model evaluation

The purpose of this chapter is to provide a thorough explanation of the model behavior, to provide enough tests and examples of model runs to ensure confidence in the reader that this model runs effectively. In other words, the model should pass certain reality tests, illustrating what it will do, under many different circumstances, and (most importantly) why it behaves in the way it does. It is useful to consider this chapter an exposé on the “model as subject”, and so another way to read this it is as a way of exhibiting the model over a course of conditions, just as we might a real subject, and explore the differences in “learning” behavior.

In other words, you might imagine putting our “model as subject” through a round of testing, just as you would a real physician. These experimental tests – first identifying how well a physician treating laboring women might respond when faced with low risk pregnancies, or women with a history of easy, quick births – provides some insight into how well doctors learn, under which conditions. I also examine how well the same imaginary physician might respond if we offered more complicated cues – such as women who have been diagnosed with gestational diabetes (a vague and imprecise cue in itself), or whose babies are in sub-optimal birthing positions at term. And I can reduce the feedback given to our hypothetical physician, making the cues more complex, or altering the value structure. How does this “physician” respond? Does this additional complexity alter behavior significantly; does it make learning more difficult? And what insight can I derive from this model about how real-life physicians might respond to such complex learning tasks?

6.1. A simple, full feedback single cue model

This first test of the model examined how well it runs under the very simplest of conditions. Recall that the actual experimental task required participants to form a judgment based on a number of cues. Cue weights were provided. In this first test, the three cues have been set to values of either 1 or 7, depending on whether there is a noise or signal trial provided, which effectively turns the multi-cue model into a single cue model, and effectively removes any uncertainty. (To clarify, for each trial, the cue values are the same. So, for example, during trial 1, cues 1, 2 and 3 would all be 1, eliminating all variation among cues.) Recall that the parameters are identified as parameters, and in the runs depicted below, there is full feedback, equal penalties and a moderate level of uncertainty. Model parameters will be adjusted as more complicated conditions are explored, such as when the model is run under conditional feedback conditions.

Figures 34 and 35 illustrate the model running, in the simplest of conditions. They are utterly uninteresting graphs, as they should be: the cutoff starts off and remains within an optimal range; the proportion of positive responses (or the number times
the simulated decision maker chooses “yes” or “signal”) moves to around .5, matching the exogenous base rate condition. Indeed, the cutoff illustrations will look like this regardless of the base rate, as it is unnecessary to change the cutoff; the model, under these conditions, just needs to place the weighted cue judgment into the correct category: signal or noise.

Why does the model behave like this? As described in the previous chapter, the model is charged with “remembering” two pieces of information. One piece is the hill climbing calculation, which determines whether recent changes in threshold and payoff have produced an outcome that encourages movement in the same direction or the opposite direction. The second piece is the error correction component, which accounts for the model’s ability to both adjust the size of shifts as well as dampening the effect of hill-climbing when the chosen threshold produces a history of correct responses.

![Figure 34](image_url)  
*Figure 34: Indicates the simulated threshold at the .5 base rate with simple cue values for 25 trials*

![Figure 35](image_url)  
*Figure 35: Indicates the proportion of positive responses at the .5 base rate with simple cues for 25 trials*
This type of behavior is replicable at the lower and higher base rates as well. As shown in the composite graphs in Figure 36, at the lower base rate, the model is selecting only about 10% of events as signals, and at the higher base rate, it is correctly selecting between 80 and 100% of trials as signal events.\textsuperscript{10}

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{figure36.png}
\caption{Results of simulation over 25 easy trials at lower (.1 base rate on the left) and higher (.8 on the right) base rates. The graphs indicate the selection rate.}
\end{figure}

\textbf{6.2. Three cues, at .5 base rate, and with full feedback}

Decision tasks experienced in real life often involve more than one cue, however. Our hypothetical physician would (hopefully) consider an array of cues before suggesting a caesarean delivery. The following figures show how well the model does when additional, and more complicated cues are added.\textsuperscript{11} As shown in the preceding chapter, the model constructs a weighted cue average as a judgment, and then selects signal or noise, depending on whether the judgment is above or below the current threshold.

A common misconception is to assume that once the optimal, or most accurate, threshold is found, the decision maker will not only adopt it, but remain there

\textsuperscript{10} Additional details about step by step changes in the model are available in Appendix C.
\textsuperscript{11} These are the same cues that are used in the experiment.
forever. Unfortunately there are rarely signposts, other than continuous feedback, to indicate when the optimal threshold is reached. So our simulated decision maker will grope around for the best threshold, often finding a range of thresholds that are adequate. As is equally clear, even though the optimal threshold may be reached, and even though the model seems to have “learned” the optimal threshold, there is no guarantee it will remain there. Instead it continues to examine new evidence about the accuracy of threshold placement, and to change threshold placement as indicated. Again, in our model and in real life, there is no clear signal that one has, in fact, found the “optimal threshold”. Constantly changing a threshold, even by a small degree, indicates that there may, in fact, be new information that requires a strategy change.

In this case, the thresholds chosen allow the model to perform well, meaning that it performs within an acceptably optimal range, and garners a fairly high level of accuracy. It doesn’t find the optimal threshold, though it comes close, for several reasons. This model assumes that as the simulated decision maker gets close to the optimal threshold, the size of adjustments (determined by the amount of errors made) as well as the amount of hill climbing required will taper off.

It is even more useful, when specifically considering the model performance, to evaluate how well the model is doing from other perspectives. The simulated decision maker, like experiment participants and like real world decision makers, is constantly adjusting the decision strategy based on perceived reward and accuracy. The percent of correct responses maps out how threshold changes result in more or less correct responses, and how the model rebounds, trying to remain within a range of thresholds that are appropriate, garnering the most correct responses as well as the highest payoff.

### 6.3. Changing the base rate

Decision makers, such as our physicians in the caesarean example, are unlikely to face a problem with a .5 base rate. For this reason, it is important to understand decision maker behavior at lower and higher base rates. Although there is far less literature to rely on here, we can make several assumptions.

The model, like an actual decision maker, will adopt a strategy that places the current threshold within a range of acceptable thresholds, as indicated by the perceived count of correct responses. Moreover, it is important to understand that even during modest levels of uncertainty, there are limits on accuracy, even when decision makers (either simulated or real) are using the optimal threshold. Behavior at the .1 and .8 base rates will reflect the fact that there are fewer opportunities to make threshold adjustments, as the type of event is either usually noise or usually signal. Still, the model should search for, we can expect that the model should grope around and find a threshold that allows it to minimize errors and maximize payoff.
As shown in Figures 37 and 38, both model runs indicate that the model is performing as expected – the threshold chosen heads in the direction of the optimal threshold. In Figure 38, the .8 base rate simulation slowly moves in the direction of the optimal. At the .1 base rate, it declines below optimal and then moves back up. Additionally, it is traditionally assumed that the .5 base rate is the most difficult base rate, and the lower and higher base rates will be easier for human decision makers to grasp, as they will quickly adjust to an environment where they should either always choose signal (i.e. .8 base rate) or noise (.1 base rate). Although this is clearly a question that requires experimental investigation, early simulation results appear to dispute it.

Figure 38 also shows the conservative cutoff placement result found in empirical studies. This happens because the nearer the model gets to optimal, by achieving fractional increases in points as well as accuracy, the less the pressure to search for a better threshold. Moreover, as indicated in Figures 38 and 39, the percentage of correct trials indicates that although the thresholds are moving fractionally in one direction, the model accuracy is also driving the search, ad preventing it from straying too far from the optimal threshold.
Figure 37: Simulated threshold outcomes at the .5 base rate with full feedback. Graph on left compares the simulated threshold to optimal. Graph on the right indicates the percentage of correct responses, averaged by block.

Figure 38: Simulated threshold outcomes compared to the optimal threshold at the .1 (left) base rate and at the .8 (right) base rates.
Figure 39: Comparison of percentage of correct trials between the low and high base rates
6.4. **Conditional feedback**

As conditions become more difficult, we can assume that the model performance may degrade. In the following runs, the amount of feedback will be limited, and we can expect that the model will be less likely to find the optimal threshold range, or take much more time to find it, as it is not getting outcome feedback after every trial. Specifically, during the conditions explored below, the simulated decision maker will only receive feedback on trials where the decision is positive. From a practical perspective, physicians will only receive feedback about their performance after they determine that a caesarean must be performed. (Note that there is another way of considering this problem as well; later, in more complex scenario runs, we will consider what happens if physicians only get feedback when they say no, which provides immediate information about misses.)

As indicated in Figures 40 and 41, the threshold shifts are still in the correct direction, but they tend to not get as close to the optimal threshold, and they tend to have a more difficult time remaining there, as predicted. Recall Figure 33 above: payoff and threshold changes together determine the direction of the threshold. And accuracy, defined here as whether the simulated decision maker was correct or not, determines the size of the threshold change. If information about feedback is missing, then the simulate decision maker is uses a random function to assume that when there is no feedback, it is correct some fraction of the time (the default is 65%).

And as shown in Figure 42, there are fewer correct responses, but the performance is still adequate: the model successfully identifies an optimal proportion of positive responses, which, in turn, yields fairly high percentage of correct responses. The model also accurately replicates the conservative cutoff, so, as shown in Figure 42, the threshold for conditional feedback at the .8 base rate is higher than the threshold for full feedback, meaning that the simulated decision maker is making fewer positive decisions.

The model behavior is solely due to the simple learning strategies employed – as stated earlier, it is relying on two mechanisms. The hill climbing function calculates whether a previous move has resulted in an increased payoff, and the error correction component moves to dampen any lingering adjustment if there is a history of correct responses, while enhancing the size of adjustments if there is a history of errors. Of particular interest is how conditional feedback affects the count of correct responses. Recall that Elwin (2005) suggests that there is some discrepancy in the way people will label, or encode, no-feedback responses – some may assume 100% accuracy when they fail to receive feedback, and others may assume that they are correct some proportion of the time, perhaps based on some intuition about how accurate they have been in the past. Also recall the literature regarding confidence, which bears some similarity to Elwin (2006): a decision
maker’s confidence will grow, even in the absence of feedback; or, reinterpreted, they will, over time, move towards an encoding strategy that will assume substantial accuracy, even in the absence of feedback. Dalgleish, too, suggests that there is a “fill in the blank” phenomenon that occurs in the absence of feedback, suggesting that as confidence increases, so does one’s belief in the inherent accuracy of a strategy. Initial analyses suggest an assumed overall accuracy of 65%; sensitivity analyses were performed to assess the implications of other assumptions.\(^{12}\)

Assumptions about accuracy will differ in their effect on particular conditions. A run where there is only feedback offered after a positive response will have a greater effect on higher base rate runs than lower base rates runs, which accounts for some of the degraded performance at the .8 base rate. Note that if feedback was only available when a decision maker identified an event as a signal trial (i.e. this laboring mother should be delivered by caesarean), and the decision maker assumed that she was correct whenever there was no feedback, there would be virtually no alteration in the selection rate at the .1 base rate. This could be one reason why obstetricians are not “learning” to adjust their thresholds – they are faced with a low base rate condition (women who need to deliver via caesarean to protect the life or health of mother or child) with only conditional feedback (they only receive feedback about false negatives, those women who should have delivered surgically; for those that are subjected to caesarean deliveries incorrectly, it is rarely perceived as an error). The tendency would be to avoid false negatives, and allow more false positives, which would necessarily lead to a higher fraction of mothers receiving (unnecessary) caesareans.

Assumptions about accuracy will differ in their effect on particular conditions. A run where there is only feedback offered after a positive response will have a greater effect on higher base rate runs than lower base rates runs, which accounts for some of the degraded performance at the .8 base rate. Note that if feedback was only available when a decision maker identified an event as a signal trial (i.e. this laboring mother should be delivered by caesarean), and the decision maker assumed that she was correct whenever there was no feedback, there would be virtually no alteration in the selection rate at the .1 base rate. This could be one reason why obstetricians are not “learning” to adjust their thresholds – they are faced with a low base rate condition (women who need to deliver via caesarean to protect the life or health of mother or child) with only conditional feedback (they only receive feedback about false negatives, those women who should have delivered surgically; for those that are subjected to caesarean deliveries incorrectly, it is rarely perceived as an error). The tendency would be to avoid false negatives, and allow more false positives, which would necessarily lead to a higher fraction of mothers receiving (unnecessary) caesareans.

![Comparison of low base rate thresholds, with full and conditional feedback](image)

**Figure 40: Comparison of low base rate thresholds, with full and conditional feedback**

\(^{12}\) I hypothesize that there are no perfectly positivist or constructivist encoders, that we will oscillate depending on how the circumstances affect the decision maker’s level of confidence. If a decision maker has been doing well, she will tend towards constructivist encoding, and risk becoming overconfident. If she has been doing poorly, then she will tend towards a more conservative positivist posture. Moreover, the distance between the threshold and the judgment may factor into how one encodes. This piece will be explored more fully in future work.
Figure 41: Comparison of high base rate thresholds, with full and conditional feedback
6.5. Changes in uncertainty

Uncertainty, in this case, can be a tricky variable to measure. Much like the decision “illusions” discussed by behavioral economists, the assumption that we are making the task easier by altering uncertainty is not as uncomplicated as it sounds. The logical assumption that we are making when we “increase the uncertainty” is that the task is harder for the experiment participant, or our hypothetical physician to understand. We are, in other words, making it more difficult for someone to find that hidden threshold.

Remember that the goal here is to find ways that that the model is both coherent – it should be running according to the principles that have been outlined in the literature chapter – as well as being able to correspond to real life data, but also to identify ways that it is not coherent, if any. In the case of uncertainty, then, it is important to define exactly what I mean. Recall that the first test made the task easier by manipulating the cues to make it easier to discern whether the event was a signal or noise event. In this case, the cue values do not change, but the differences between the perceived distributions do. As the two distributions move apart, and overlap less, it should be easier for the discerning participant to identify the
difference between the signal and noise events. Yet there are several problems with this assumption.

The optimal threshold is still hidden – and with ambiguous cues, the slight improvement or impairment caused by the shift of distributions will probably not be enough to affect behavior. (Recall that we are measuring uncertainty through the use of \( d' \), or the difference between the means of the two distributions.) Moreover, there is an additional problem as the distributions move further apart – the range of appropriate thresholds becomes much bigger, even as optimal range of thresholds may produce nearly the same performance as the optimal one. So a greater the \( d' \) usually leads to better overall performance, but does not necessarily mean the decision maker will have any easier time finding the optimal threshold. In other words, there is just a larger area where there is lower motivation for learning because the feedback suggests the learner is within a correct range.

Moreover, making the task more complicated by moving the distributions closer together would probably make the task slightly more difficult. But it is also useful to remember that the differences between a \( d' \) at 1 and \( d' \) at .5 are not significant enough to make a tremendous amount of difference in the learning behavior.

Figure 43 illustrates predicted behavior. At the more difficult uncertainty level, it performs the worst, although it does resume a fairly high degree of correct responses. By the final 4 or 5 blocks, the model is essentially running as well as it did during the less challenging uncertainty levels, with behavior for all three uncertainty levels clustered together.

---

13 Graphs of the differences between .5, 1 and 1.5 \( d' \) are appended.
As shown in Figure 44, the higher base rate condition, with conditional feedback, across uncertainty levels, is much more challenging for the model, and it performs far less well during these conditions. From a structural perspective, this behavior is explainable. The model begins with a cutoff that is higher than optimal, and since it is only receiving feedback on signal responses, a higher than optimal cutoff essentially means that most of the responses will be noise responses. Moreover, the model is assuming a decision is correct much of the time (65% of the time) it doesn’t receive feedback. Regardless of the level of uncertainty, the model’s decision making structure is essentially getting no feedback at all, and random searching is proving ineffective.
6.6. Changes in payoff values

So far the model has been tested and explored over different base rates, with conditional feedback, and over different levels of uncertainty. The combination of hill-climbing and error correction does a fairly good job of “learning” an adequate threshold, or identifying a range of adequate thresholds, and I can generalize that physicians may use this combination of learning strategies to determine their own cutoffs.

A final exploration examines whether or now altering the value structure, or penalties on incorrect answers, might also alter overall learning behavior. Can we effectively change a threshold by altering the value structure? Theoretically, for example, we should be able to reduce the amount of false alarms by enacting a penalty. Figure 45 indicates that it is not that easy to alter learning behavior. At the .5 base rate, with only conditional feedback, placing a penalty on false alarms does result in a higher threshold, but with it, the accompanying higher error.

![Figure 45: Percentage of correct responses and proportion of positive responses for results at the .5 base rate, with conditional feedback, comparing the effect of a penalty on false alarms](image)

Why does the model behave like this? For at least two reasons – at this base rate, it is starting at a threshold that happens to be approximately correct, so conditional feedback, though an impediment to optimal performance, does not degrade it significantly. But when the penalty placed on the false positive error is larger, the
model has an even more difficult time finding the optimal threshold. On one hand, the model is hill-climbing to find the best place for a threshold based on the values, while it is also shaping a strategy based on error correction. To make matters more complicated, as the threshold rises in response to the heavier false positive penalty, there is progressively less feedback available.

How does the model respond to the other base rate conditions? The same complicating factors are evident, as well as the fact that at the very high or very low base rate conditions, there is even less opportunity for feedback. Yet penalties on either error fail to have much of a result – there is little deviation from the original behavior shown in the conditional feedback run.

6.7. Discussion

The goal of this chapter has been to examine various aspects of model validity by comparing its behavior to common sense and the literature – to illustrate how well our simulated decision maker “learns” before it is compared to any actual decision makers. The model has been run over a series of tasks, from very simple to more complicated, to show how the model performance degrades as it faces more difficult challenges.

Specifically, single cue model runs perform well – the simple learning mechanisms employed here facilitate quick learning. As more, and less certain, cues were added, and these cues became more uncertain, the model was run over 500 trials, or 20 blocks, to more fully illustrate behavior. The model continues to perform well, as it successfully moves in the direction of optimal, often very quickly. The model was also run over different base rates, and although there is less literature to rely on, it seems to run according to predictions.

When there is limited feedback, conditional on positive responses, the model performance degrades. It tends to produce fewer correct answers, but still show some indication of learning over time. And in general, model performance matches what we suspect human performance would achieve, in the more difficult conditions, although this will be explored more fully in the next chapter.
7. Model comparisons to data

This final chapter will explore model performance when compared to subjects’ performance in the multiple cue threshold learning task, with the goal of synthesizing a number of different concepts, derived from the literature, and from the empirical and modeling investigations, into a cogent whole. In particular, I am interested in identifying when the model behavior closely matches the experimental data, as well as when (and why) it fails to match. If the model was a perfect replication of human behavior, it would match behavior exactly. This model does not offer a perfect replication, but provides insight into where models can be useful, where additional empirical investigation is needed, and where there are gaps in theory. It is one way of exploring whether the combination of two well-known models (hill climbing and error correction) are sufficient tools with which to explain the patterns of behavior exhibited by real-life decision makers. Overall, in the majority of cases, model behavior matches the pattern of learning as well as the overall level of accuracy, allowing for exploration the structural basis for such behavior. Perhaps most importantly, such an exercise provides a foundation for understanding some of the behavior of real world decision makers, offering some insight unto which behavioral changes might improve the performance of physicians, and quell the rapid rise in caesarean deliveries.

In this set of scenarios, the model is being run with the same set of cues as the subjects, effectively making it another subject. For the most part, the following scenarios will mimic the experimental tests used on the human subjects, including the same base rates, levels of uncertainty, feedback and payoff structure. The simulated decision maker will be compared to the average of all the human subjects.

7.1. Comparison: base rates

As discussed, not only are we interested in exploring how well the model performs at the .5 base rate, but also at the higher and lower base rates. Figure 46 explores the similarities and differences between the experiment and simulation data at the .5 base rate. Although there is a similar pattern of learning, there is some discrepancy between the selection rates. The simulation tends to make many more sweeping changes before settling on an acceptable threshold. This result is not altogether unexpected: experimental data is an average of subjects, where the simulation data is based on a single run. Figure 47 illustrates this, indicating the wide degree of variation among subjects at the .5 base rate condition. In general, as shown in Figure 48, the simulation tends to settle on a threshold that balances both rewards and accuracy (Kruschke, 1999), resulting in a strategy that is more successful than that of the subjects.

The figures can also be deceptive. The left frame of Figure 48, for example, suggests that the simulation’s threshold is not a good match to the subjects’ threshold; it chooses a selection rate that is quite different from the subjects’ average

---

14 See appendix for a table of the number of subjects in each condition.
selection rate. Yet, as shown in the right frame of Figure 48, by the last half of the experiment, the simulation’s percent correct score is on par with what the subjects’ are scoring. Essentially, although the subjects are choosing fewer events as signals, they are not getting any (or many) more correct. This is a critical insight, suggesting that even though the combination of error correction and hill-climbing represents a strategy that correctly seeks the optimal threshold, and therefore tends to match decision maker behavior, this simulation suggests that it is not necessarily the strategy (or at least the complete strategy) that real life decision makers are using. This “fit” issue is due to the model’s structural rules – hill-climb while also balancing accuracy. The inherent uncertainty in the environment dampens any speedy movement towards optimal.

The propensity of experimenters\(^{15}\) to use the .5 base rate is also somewhat deceptive in that decision makers usually choose an initial threshold that is close to the middle, making the task a bit easier. As the very same initial threshold is used throughout all model simulations, the “easier” initial threshold for the .5 base rate becomes a challenge for the model at the lower and higher base rates as an initial threshold of 3.5 is too high for the .8 base rate (which has an optimal of 2.7) and too low for the .1 base rate condition (which has an optimal threshold of 6.1). In other words, the model has more to “learn” at the higher and lower base rates. One gap that needs to be filled by more empirical investigation is to uncover what transpires in the initial block (or blocks) of the experiment. It seems that subjects learn very quickly to reduce or raise their threshold to within an acceptable level during that initial block of trials. This is potentially insightful with regard to real world decision making as well: physicians begin their careers primed with the expectations of the current environment. If the environment includes an assumption that there will be a high number of caesarean deliveries, then that initial “sticky” threshold might shape all future experiences.

Still, as shown in Figures 48 and 49, the model and subjects seem to match fairly closely. From a structural perspective, the model is maximizing rewards and minimizing errors, which, at least in this particular condition, results in behavior that appears to match the subject’s learning behavior. Specifically, the model is altering the threshold based on the rewards until it corrects enough to dampen the hill climbing effect. In other words, the model is receiving feedback about how accurate its judgment is, and moving, per the hill-climbing equation, in the direction that will maximize points in the next trial. When the model receives feedback that it is correct, it will dampen the hill-climbing effect, or limit searching for a more optimal threshold. The inherent uncertainty in the environment, though, necessarily means that even when the model is close to an optimal threshold, it can receive feedback that it should resume searching.

So, to clarify the components of the model that produce this behavior, consider Figure 33 again. The model is given an initial threshold, and considers the weighted average of the cues. Throughout most of the early blocks, it is assuming a threshold

---

\(^{15}\) See, for example, Busemeyer and Myung (1992), Ever (1992) or Barkan and Erev (2001)
that is too high, and the trend is to respond to payoff feedback that indicates that the rewards increase as the threshold declines, by reducing the threshold. Moreover, as the more the threshold declines, resulting in a much higher fraction of positive responses, the percentage of correct responses increases, acting the dampen the hill-climbing effect by making the step changes smaller.
Figure 46: Comparison of simulation and experiment results at the .5 base rate, with full feedback. Proportion of positive responses is shown on the left, and percentage of correct responses is shown on the right. (Moderate uncertainty and symmetrical values)

Figure 47: Comparison of 6 different subjects at the .5 base rate, full feedback condition. (Moderate uncertainty and symmetrical values)

Figure 48: Comparison of simulation and experiment results at the .8 base rate, with full feedback. Proportion of positive responses is shown on the left, and percentage of correct responses is shown on the right. (Moderate uncertainty and symmetrical values)
7.2. **Comparison: feedback**

How the model fares when compared to actual decision makers on tasks that involve conditional feedback is of even more practical interest. As previously discussed, physicians are often faced with problems that involve only limited feedback, conditional on their decisions. Figure 50 indicates some interesting variation between the simulation and experimental results. Again, there are some fairly sizable changes in the simulation that do not occur among the subjects, but this could be due to the effect of averaging. Figures 51 and 52 seem to indicate that conditional feedback may limit some of the excessive searching that happens during the full feedback conditions.

This could mean two things, and more empirical work is needed to untangle the relationships. On one hand, with less feedback, there are fewer opportunities to learn, but there might also be more incentive to make each opportunity mean more. It may induce more analysis and less random searching, or less reliance on intuitive snap decisions that lead to swinging changes between thresholds (Hammond, 2001). On the other hand, limited information, especially information that is offered through the use of multiple ambiguous cues, is traditionally thought to induce more intuitive behavior.

In general, the investigation thus far suggests that a bit of both might be true. Initially, the ambiguity of the cues, coupled with a visually stimulating environment and very little in the way of an organizing principle probably induces more intuitive behavior, which is marked by the greater variation in the first half of the task. As the decision maker develops a more complete strategy, there is less variation, and the decision maker and the model tend to settle around a small range of adequate thresholds. Although the model cannot embrace these decidedly human cognitive traits, it does mirror the process. It, too, tends to explore widely, especially during the .5 base rate, during the first half of the task, then, as the error rate declines, search is dampened.
Figure 50: Comparison of simulation and experiment results for the .5 base rate, with conditional feedback

Figure 51: Comparison of simulation and experimental results for the .8 base rate, with conditional feedback

Figure 52: Comparison of simulation and experimental results for the .1 base rate, with conditional feedback
7.3. **Comparison: payoff**

Experimental data for the effect of changes in penalties is available for the .1 base rate, and as shown in Figures 53 and 54, it represents a challenge for the model. Recall that during the experiment, when a penalty was enacted on false negatives, it meant the subject should be more careful to avoid misses, so she should lower her threshold and say yes more often. The scenario is inversely repeated with false positives – a subject should avoid false alarms, so she should raise her threshold and say yes less often. This is directly applicable to doctors as well – we can assume that if a penalty is enacted on errors where a woman should have received a caesarean, but did not, a doctor, or a practice or a hospital might enact policies that would lower the threshold and encourage more caesarean deliveries. If doctors (or the general public) were to decide that far too many caesareans were taking place, they might enact a penalty on false positives, or women who were having surgical births erroneously, and push the threshold up, discouraging caesarean births.

In the experimental data presented here, the description offered above seemed to happen for misses, but not for false alarms. This makes sense, though, because there are simply not very many opportunities to improve on the current rate of false alarms, because the subject is making very few of those errors. You'd also not expect her to lower her threshold and say yes more when penalized for false alarms, and somewhat interestingly, the data suggests people did do that.

The model, on the other hand, is trying to balance both accuracy and rewards through error correction and hill-climbing. At the .1 base rate, especially when only receiving conditional feedback, there are just not a lot of options - for example, eventually there are no false positive errors, so that penalty cannot effect behavior at all. And there are very few false negative errors, so even when the elevated penalty suggests there should be a slightly lower threshold to choose more signal events, the error correction functions disputes it, and dampens (in this case, eliminates) any chance of doing that. So even though it seems unlikely - that during this condition both penalties result in the same behavior - it is a reasonable thing for the model to be doing.

So why are subjects, and probably real life decision makers, such as doctors to laboring women, not choosing this reasonable strategy? Because they are not in a position to maximize rewards, they need to strictly minimize one particular error – that of false negatives. So even though the model’s reasonable strategy provides the most accurate overall strategy, and severely limits the amount of false positive errors (an error for which doctors are rarely penalized for), it would necessarily lead to some false negative errors.
7.4. **Summary**

A number of interesting insights emerge, as the simulated subject is compared to the experimental results.

- Again, the base rate matters: at the .5 base rate, simulated subjects make large and sweeping oscillations prior to settling on an acceptable threshold. As the thresholds move further from the .5 base rate, the variation between experiment and simulation diminish. In general, the simulated decision maker also tends to identify a slightly more optimal threshold than the experiment participants, suggesting that even through the combination of error correction and hill-climbing represents a fairly accurate strategy, it isn’t necessarily the complete strategy used by experiment participants, or real-life decision makers.

- The effects of limiting feedback are of even more practical interest. Despite early assumptions that limiting feedback should significantly degrade performance, the comparisons shown here contradict that proposition. It is possible that the ambiguity of the cues, complicated with a (somewhat) visually stimulating
environment and difficulty identifying an organizing principle induces more initial intuitive behavior, leading to larger threshold shifts. Later, as experiment participants settle on a compromise position that meets the need for both accuracy and rewards, fewer larger shifts are evident. The model results tend to mirror this process.

- Lastly, and perhaps most illustrative with regard to the dilemma of caesarean deliveries, are the results of changes in payoff. Even though the model chooses a strategy that is very reasonable, and, for the most part, quite accurate, the experiment participants choose a strategy that is less optimal. This less than optimal strategy mimics the real world physician’s strategy, however, which is not to maximize rewards, but to focus on minimizing one particular error: that of false negatives, or women who should have received a caesarean but did not. Complicating matters is the fact that penalties placed on false positive errors, in the case of physicians accompanying laboring women, are very rare, and the penalty placed on false negative errors is far larger than anything simulated here.
8. Model Improvements

While the development of a model that can replicate data is important, it cannot claim to be generalizable unless it is capable of passing a number of other critical tests. As the original model, with the basic hill-climbing formulation, we found to have several important flaws, most notably that it resulted in different behavior when run with different initial noise seeds, it became necessary to investigate further. In the following chapter, the model presented is identical to the earlier version, with the exception of the construction of the environment sector. Instead of extracting the same cue values that were used in the aforementioned empirical work, the model draws random cue values from three cue distributions. Such distributions are formulated to best recreate the environment created by the experiment, with similar ranges, means and standard deviations, and are shaped as normal distributions.

This model represents, to a great degree, a working laboratory, where the base rate, amount of feedback, degree of uncertainty, and amount of feedback can be easily altered and manipulated. Recall that all generations of earlier models have neglected to allow for repeated testing of different conditions, so their results have primarily provided insight about behavior at the .5 base rate. In recent years, models have begun to examine the effects of conditional feedback, but none so far have attempted to do this while also manipulating levels of uncertainty and payoff, and at different base rates. Moreover, as this is a multiple cue threshold learning model, the weights placed on the cues can also be altered. As physicians often face conditions where they are receiving only limited feedback, with high levels of uncertainty, a value structure that may penalize certain types of errors, and with a range of cues, an exploration that only examines behavior at the .5 base rate, with none of the other complexities promises to yield less satisfactory results.

Each cue, in this working laboratory version of a multiple cue threshold learning model, has a signal and a noise seed. To facilitate this more rigorous evaluation of the model, only sensitivity tests are shown, to illustrate the model’s ability to find appropriate threshold’s despite variability in the noise seeds for the each cue’s signal and noise distribution. Simply put, this means that the model is allowed to choose different starting places for each of those 6 variables. The resulting graphs will illustrate the convergence, or lack of, around a particular threshold. In the sensitivity graphs shown, each will indicate the middle 50% of 200 runs, the middle 75%, middle 95% and 100%. The single line running through the center of the runs represents the mean of the 200 runs.

It is important to also be clear that there is randomness in the starting point, as well as randomness within the distributions themselves. So while the following graphs will illustrate behavioral differences due to the randomness in the starting point, this model is also based on random signal and noise distributions for three different cues. This is the very same logic used to determine behavior in the earlier model; while the earlier rendition of the model draws from exogenous cue values chosen to represent a random
environment, this model chooses from endogenous random distributions.

The table below (see Table 1) illustrates an initial testing procedure, where the very basic model is examined first, with additional pieces tested and manipulated over time. The purpose of such a follow-the-breadcrumbs approach is to provide a replicable model, grounded in theory and easily explainable to the layman. The first step begins with the original formulation constructed by Busemeyer and Myung (1992), which includes their random weighting structure placed on the indicated threshold, as well as length of lags which mimic the “what I decided last time” convention. I will also start with the original testing parameters and begin at the .5 base rate traditionally used in prior simulation work.

Table 1: Procedure for model testing

<table>
<thead>
<tr>
<th>Scenario (.1 base rate)</th>
<th>Normal weight on ITC</th>
<th>Lag length of recent cutoff and payoff</th>
<th>Time over which to consider prior cutoffs</th>
<th>Time over which to consider next direction</th>
<th>Adding correct responses smoothing time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Examination of original Busemeyer formulation</td>
<td>.95</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>n/a</td>
</tr>
<tr>
<td>Busemeyer with longer lags and smoothing times</td>
<td>.95</td>
<td>10</td>
<td>3</td>
<td>10</td>
<td>30</td>
</tr>
<tr>
<td>Busemeyer with longer lags and smoothing times, and without a weight on the indicated threshold change</td>
<td>1</td>
<td>10</td>
<td>3</td>
<td>10</td>
<td>30</td>
</tr>
</tbody>
</table>

8.1. Taking apart the original formulation

The purpose here is to examine how each of the pieces of the original formulation contributes to the behavior of the model. Recall that the real brain of the model, the structure that contributes to the sense that the model is “learning” over time how to set a correct threshold, is located on the threshold setting screen. The purpose here is not so much to tell a story, but to examine each component of the model that shapes the threshold choice and understand it’s contribution. As shown in Figure 55, the key elements are represented by several smoothed variables and lags. The smoothed variables identified here – the recent cutoff and the recent payoff – are shaped by their time constants. Those time constants, the lag for recent cutoff and the lag for recent payoff were, in the original Busemeyer and Myung equation, a way to formulate the concept of “what I decided before”. In other words, we (and they) are interested in modeling the idea that last time’s cutoff was important. This model goes one step further, allowing the model to embed a more complex form of “memory”. Because this model is constructed to calculate the change between, for example, the current cutoff choice and last time’s cutoff choice, the number chosen for the lag refers to how many prior choices we are using to construct a running average. In simplest possible terms, this running average represents a memory of recent choices. If the value of those lags were set to 1,
for example, we would expect the model to have no memory at all, and there is no expectation that it could find a reasonable threshold.

From a more technical perspective, each of these memories or “smooths”, captured as a weighted average of a new data point and the previous value of the average, is actually an exponentially weighted average of all previous data, with the exponential weight declining for the previous data points. So, the "memory" is strong for recent data and weaker and weaker as data recedes into the past. How quickly that memory fades for past data is dependent on the time constant, which is what will be adjusted in the following runs. This type of formulation represents, to a great degree, how an actual decision maker might treat new and old information, and therefore, how she might adjust her threshold. Recall that earlier models (Erev 1998, for example) have considered this question of updating and chosen modeling structure that, while effective for some base rates, does not seem to replicate the strategies used by human decision makers.

Several other smooth variables are key to producing learning behavior with this model. The smoothed cutoff, shaped by its time constant, time over which to consider prior payoffs acts as a similar type of model memory. This memory variable acts to dampen new information, based on the memory of prior cutoff information; a longer memory effectively means the new information will have less weight. Conversely, a small smoothing time means that new information will be treated in the same way as old.

\[\text{Recent} \quad \frac{\text{lag for recent cutoff}}{\text{lag for recent payoff}} \quad \text{change in cutoff} \quad \text{indicated direction of cutoff} \quad \text{effect of ratio of current to desired correct responses on size} \quad \text{size up} \quad \text{size down} \quad \text{norm size up} \quad \text{norm size down} \quad \text{remembered frequency of correct responses}\]

**Figure 55: Threshold setting screen**

\[\text{Smoothed cutoff} \quad \text{time over which to consider prior cutoffs} \quad \text{indicated direction of next cutoff} \]

\[\text{Smoothed payoff} \quad \text{lag for smoother payoff} \quad \text{change in payoff} \quad \text{indicated direction of cutoff} \quad \text{effect of ratio of current to desired correct responses on size} \]

\[\text{Recent cutoff} \quad \text{lag for recent cutoff} \quad \text{time over which to consider prior cutoffs} \quad \text{indicated direction of cutoff} \quad \text{effect of ratio of current to desired correct responses on size} \]

\[\text{Base Rate} \quad \text{indicated cutoff} \quad \text{feedback condition} \quad \text{time over which to consider next direction} \quad \text{time over which to consider prior cutoffs} \quad \text{size up} \quad \text{size down} \quad \text{norm size up} \quad \text{norm size down} \quad \text{remembered frequency of correct responses}\]

16 It would be prohibitively time and space consuming to examine every single result for every potential lag time. The ones chosen here attempt to convey to the reader a concise and careful experimental procedure – moving from 1 to 10 – without showing the potentially exhaustive list of runs. The model is attached to this document, so do examine more, as well as replicating the ones shown here.
information, or memory. If the value of this smoothing time were set to 1, then there would be no memory of prior actions, and we expect the model to be unable to find any optimal cutoffs. All weight would be placed on new information, in other words, with no weight on past experience.

A last smooth, also representing memory, is also critical to learning behavior: the **smoothed indicated direction of the next threshold** is shaped by the time constant, **time over which to consider the next direction**. Again, these variables allow the model to have an embedded sense of time and memory. In this case, the memory of which direction was tried over the last n trials allows the model to calculate the direction that is most likely to lead to higher payoffs and greater accuracy.

As shown in Figure 56, the original Busemeyer/Myung parameters are used first. Although the mean value of the 200 sensitivity runs suggests that the model finds thresholds that near an optimal range, what’s more interesting is the lack of convergence around any particular threshold. Recall that the white area in the graph represents the middle 50% of all runs, with other sections representing the middle 75%, middle 95% and 100% of all runs. In the graph at the left in Figure 56, 50% of all runs fall between 3 and 6, an unlikely amount of variation. This large amount of variation in the smoothed cutoff is significant, and suggests that the original formulation can be improved upon. It also suggests that if random seed changes can create this amount of variation, sensitivity analysis is a more appropriate tool for examination. The graph on the right of Figure 56 also suggests that there is little convergence, even towards the end of these runs, with only a slight tendency for the 50% band to become more visible. Such a lack of convergence draws into question whether the original Busemeyer and Myung (1992) hill climbing formulation is appropriate here, or whether some improvement is necessary.

![Figure 56: Base runs for the medium base rate. The graph on the left indicates the smoothed cutoff over time, for 200 runs; the graph on the left is the Indicated directed of next cutoff.](image)

Before improvement is possible, however, it is critical to understand why the model is

---

17 In the model runs examined below, the .1 base rate is used as the default base rate; later in this chapter, the model will be run over multiple base rates.
behaving as it does. As far as system dynamics models go, after all, this one isn’t exceptionally complicated – to a great degree, it is simply responding to the payoff and levels of accuracy and ratcheting a threshold value up or down. But understanding what compels it to do that is critical. Figure 57 provides some additional clarity: as a threshold moves into a more optimal range, we should see some indication that the model is achieving more accurate results: there should be more true positives and negatives, over time, for example, and a decline in the amount of errors. In the series of graphs shown in Figure 57, there is little indication that the model is moving the threshold at all: as indicated by the black line in each of the graphs below, the threshold produces results that indicate little improvement over time.

Figure 57: A series of graphs illustrating the number of true positives, true negatives, false positives and false negatives in the base run.

Figure 58: Medium base rate, with longer lag values on the recent cutoff and the recent payoff.
Allowing the model to have more “memory”, a sort of proxy for experience, should enable it to find a more optimal threshold, even at the .5 base rate previously used by decision and judgment simulation modelers. Figure 58 illustrates the effects of learning behavior when the lag values for the recent payoff and cutoff are moved to 10, while all others remain at 1. As shown in the graph on the left, there is a tremendous amount of variation, and constant changing, as shown in the graph on the right. Smoothing the payoff does provide a better picture of the history of the payoff, however, as shown in Figure 59.

In Figure 60, if we return all the smooth values to 1, and examine the influence of only the time over which to consider prior cutoffs we find that lengthening only the influence of this memory variable leads to a mean response that’s moving in the correct direction, but also a very large range of potential runs. This result is critical, as it suggests that a potential problem with this formulation: convergence should result, as regardless of the starting points of any of the noise seeds, the model should be able to find an optimal range of thresholds most of the time. If we only focused on this smooth, we’d find that the range of 95% of potential cutoffs chosen by the model would vary between a threshold of 3 and 7, a difference of 4 thresholds!

**Figure 59: Recent payoff - at medium base rate with longer recents**

**Figure 60: Smoothed cutoff and indicated threshold change at the medium base rate with**
A longer value on time to consider prior cutoffs

A last examination reveals that, with all other lags remaining at 1, altering the memory of the next direction alone does little to improve behavior. As shown on the right in Figure 58, changes to only the **time over which to consider the next direction** minimizes the amount of variation significantly, forcing the model to only make small negative changed, driving the threshold down.

![Graph showing smoothed cutoff and indicated threshold change](image)

*Figure 61: Smoothed cutoff (left) and indicated threshold change (right) at the medium base rate with longer value on the time over which to consider the next direction*

If the lags on the recent cutoff and payoff, and the time over which to consider the next direction are both increased, however, there is both less variation and the sense that the model is moving in the correction direction. And if we allow the third memory variable, the time over which to consider next direction to dampen the searching, by giving slightly more weight to old information, we find the results (Figure 62) indicate that the model is also settling on a threshold that nears optimal. Such a threshold is therefore able to garner the expected amount of rewards – as shown by the frequency graphs in Figure 63 – and minimize errors. Note, too, that because the threshold is still slightly below optimal, there are more false positives than desired.
Figure 62: Smoothed cutoff at the medium base rate with all smoothing times lengthened

Figure 63: Series indicating the frequency of true positives, true negatives, false positives and false negatives.

For purposes here, how the model learns at a base rate that is more likely to be experienced in the real world, is even more important. Figure 64 indicates that behavior at the .1 base rate, adjusting for longer lags, is in line with our expectations: there is some convergence, and the optimal threshold of approximately 6 is achieved.
But the question remains: can the model be improved further? There is still more variation than is warranted, even after altering the times constants of the smoothed variables. Moreover, can the model more effectively (and more quickly) find optimal thresholds?

8.2. Reformulating for a better model

It is crucial, in devising a model that is capable of replicating human behavior, as well as producing reasonable results where there is an absence of empirical work, that the structure be as simple and concise as possible. One component of Busemeyer’s original formulation: a small weight placed on the indicated threshold change, resulted in interesting, and very disconcerting results in our laboratory model. When placed at 1, for example, this weight was eliminated, and the model ceased all searching. According to the original formulation, that small weight was critical to producing enough “noise” in the system to provoke continuous searching.

The weight on the indicated threshold change has been removed in the following graphs. One simple change – the original subtraction functions that calculate the change in cutoff and change in payoff have been changed to ratios instead, and the need for the indicated change in threshold weight was eliminated entirely. As shown in Figure 65 below, there is improved behavior, and with a cleaner, more concise model. Figure 66 indicates that although the model chooses a threshold that is too high, thereby missing some true positives, it also effectively identifies most true negatives, while minimizing errors.

Changes in these equations, in general, makes for a more robust model – we can remove

\[ h = (c_{\text{cutoff}} - c_{\text{cutoff, last}}) \times (p_{\text{payoff}} - p_{\text{payoff, last}}) \]

So, to clarify, the original h function is the product of subtracting the last two values of the cutoff and the last two values of the payoff. The amended h function is the product of ratios, instead: the current cutoff divided by last trial’s cutoff and the current payoff dived by last trial’s payoff.
structure that was providing nothing useful to the behavior of the model, and can construct a simpler, more elegant model that is easily followed by the layman. It also makes any insights derived from multiple cue threshold learning easier to replicate.

![Diagram showing smoothed cutoff after Busemeyer/Myung weighting function is removed](image1)

**Figure 65:** Smoothed cutoff after Busemeyer/Myung weighting function is removed

![Diagram showing frequency of true positives, true negatives, false positives, and false negatives](image2)

**Figure 66:** Series indicating the frequency of true positives, true negatives, false positives and false negatives

### 8.3. Further exploration: an improved model over different conditions

#### 8.3.1. Changes in base rate

With better structure in place, the model can now be run over the different experimental
conditions. Recall that although experiments, as well as models, have historically focused on the .5 base rate, we are interested in the more extreme base rates that are more applicable to modern decision dilemmas. During high base rate conditions, the optimal threshold is approximately 2.3. As shown in Figure 67, model behavior does not diverge much from a small set of thresholds, the majority of which fall on or near the optimal threshold. Note, too, that this version of the model allows for a better potential fit with empirical data; prior renditions struggled to find an appropriate threshold during the higher base rate runs. Figure 68 indicates that the improved structure results in better outcomes at the higher base rate as well: the chosen threshold minimizes errors while maximizing rewards.
A final interesting discovery worth noting: the time over which to consider prior cutoffs is dependent on the base rate, to some degree. The more extreme the base rates – or, in other words, the further the base rates move from the .5 base rate, in either direction – the longer the lag values. This is completely intuitive, for two different reasons: the more extreme base rates require a longer memory of past events, in order to find a more optimal threshold. In the case of the .1 base rate, for example, there are only signal events approximately once every 10 events. Similarly, in the .5 base rate, a shorter memory allows for more effective threshold changes, as there are an equal number of signal and noise events.

8.3.2. Changes in feedback

Another critical test, now that the model structure is satisfied, examines model behavior at the conditional base rates. As comparing two sets of sensitivity runs on one graph is nearly impossible to decipher, the following scenario depicts the sensitivity analysis for the .8 base rate run with full feedback, with another single run with only conditional feedback on positive responses. The purpose here is to show that the model is able to replicate the results of the empirical work shown earlier: conditional feedback at the .8 base rate, as shown in Figure 69, does result in a slightly higher mean threshold, and fewer positive decisions. Also shown, in Figure 70, the model identifies a threshold that minimizes errors and maximizes rewards over time.

Figure 68: Series indicating frequencies of true positive, true negative, false positive and false negative
Although not shown here, the results of conditional feedback over the other base rates indicate similar results to the empirical work: through much of the simulation, there is similar behavior to the base run. Note, too, that because this model is now based on a random environment, there is substantially more power to shape experimental conditions. While we were limited to three base rates and two feedback conditions, this model allows
us to alter the base rate anywhere between .1 and .9, and alter the feedback between full feedback and feedback based on whether there was a positive or negative decision. One can also choose no feedback, as an extreme test of the model; during this condition the model, having received no feedback, searches briefly randomly and then chooses a suboptimal threshold and ceases all searching.

8.3.3. Changes in uncertainty

Altering the degree of uncertainty in the model is also substantially easier. In the runs shown below, the level of $d'$ has been altered on each cue. The default value of $d'$ is 1; in Figure 71, a higher (.5 for each value of $d'$) and lower (2, for each value of $d'$) level of uncertainty is examined. The graph on the left indicates that there is very little change from the default. As the base run achieved substantial convergence, as well as finding an optimal range of thresholds, an increase in $d'$ did little to change behavior.

![Figure 71: Changes in uncertainty as represented by $d'$ values of the cues](image)

The graph on the right, however, indicates that increasing the level of uncertainty does have some effect on model behavior. Thoroughly in line with earlier suggestions, increasing $d'$, contrary to what is often expected, does not produce “better” behavior. In fact, if you recall an earlier explanation of $d'$, we find that a larger range of thresholds is exactly what one might expect, as there is a much wider region of potentially acceptable thresholds. From another perspective, however, this is yet another reason to not rely on individual runs when examining this particular model.

8.3.4. Changes in value structure

Again, earlier models have been greatly limited in their ability to test behavior over time and over different values of penalties. As shown in Figure 72, both types of penalties act to reduce the threshold. This is completely unsurprising; although we’d expect that the threshold would drop when false negatives are penalized, penalties at the .1 base rate when there is only conditional feedback cannot be tremendously effective as there are very few false positive errors available to penalize.
Figure 72: Low base rate, with conditional feedback, with penalties on false negative errors (above, with smooth threshold on the left and the indicated change on the left) and false positive errors (below, with smooth threshold on the left and the indicated change on the left).

As shown in Figure 72, changes to the value structure can lead to spectacular differences from earlier runs. The graphs shown above indicate a low base rate, conditional feedback scenario run, with changes to the penalties placed on false negative (default is -5; changed here to -50). To place this complicated run in perspective, consider it a way of examining what might happen if a hospital were to penalize doctors for missing too many cases where women should have received cesarean section deliveries.

Interestingly, shown in Figures 73 and 74, providing greater rewards for true positives and true negatives yields substantially more optimal behavior than larger penalties for misses and false alarms:
Figure 73: Low base rate, with conditional feedback and greater rewards for true positives and true negatives. Smoothed cutoff is shown on the left and the indicated direction of the cutoff is shown on the right.

Figure 74: Results, clockwise from top left: True positive, true negative, False negative, False positive.

Although there was no such run in the experimental data presented above, the simulation results suggest that additional testing along this vein might prove interesting. From a policy perspective, this simulation suggests that doctors who are rewarded for greater accuracy instead of penalized for errors find a more optimal threshold, which, in this case, would be a higher threshold and a lower rate of cesareans. The reason for this is purely practical: the type of errors we are trying to avoid are difficult to learn from, when there is such a low base rate, because there are so few of them. Again, this is another reason why an entire generation of models that focused on a .5 base rate were of little
value to policy makers: not only does the value structure need to match the goals of the audience and seek to achieve the best outcomes, but they must also consider the structure of the task and the environment.

8.3.5. Changes in cue weights

Having tested numerous other conditions, we can now also test the effect of changes in cue weights, completing our goal of examining a true multiple cue threshold learning model. In all previous runs, the cue weights have mimicked the empirical work – Cues 1, 2 and 3 have been set to .2, .5 and .3, respectively. They have standard deviations of 3 and a value of 1 for each d’.

In Figure 75, I have altered the structure of the heaviest cue, substituting a small standard deviation. One can interpret this in a number of ways, but perhaps the simplest explanation is that we have gotten a bit better at measuring this particular cue and it is providing a smaller range of accurate data. Better data ought to make for more accurate decision making, and in this case, less divergence from an optimal threshold.

As shown in Figure 75, the resulting behavior indeed suggests that alteration of the cue forms have a substantial impact on behavior, as the model is able to find the optimal range most of the time.

8.3.6. Comparing with current events

A final simulation attempts to convey how physicians and families ended up in a position where there is a far higher selection rate than optimal, with limited feedback. Prior to now, we have reasonably assumed that physicians will usually get feedback when they decide to do a cesarean. There are some concerns with this assumption, not the least of which is that when they do judge a laboring woman as a candidate for a c-section, they are usually rewarded, even though many of them are probably done in error, or
unnecessarily. In the following simulation, that assumption is questioned, and the simulation is set up so that feedback is provided on a “no” response. In other words, a physician receives outcome feedback only when cesareans are not performed. In this way, there is a heavy penalty for false negatives, and little penalty (or a small reward!) for false positives.

As shown in Figure 76, the result is a much lower than optimal threshold, and a higher than optimal selection rate. Indeed, the selection rate is close to our current selection rate.

![Figure 76: Turning conditional feedback on its head: examining the selection rate when feedback is only on "no"](image)

### 8.4. Discussion

This chapter seeks to accomplish a few fairly important goals. First, a thorough assessment of structure is critical if one is seeking generalizable results. Even though earlier chapters examined a model that fits data nicely, and provides useful behavioral insights, it was able to be substantially improved. Indeed, the earlier rendition of this model should garner almost no confidence in the policy maker, as the original Busemeyer and Myung formulation was very likely to produce widely divergent behavior depending on the starting points of any of the noise seeds. Recalibrating the model, as well as correcting structural flaws enables much greater generalization of results. In general, one should still be wary of any single run from this model, and instead examine sets of runs, such as the sensitivity analyses of 200 runs shown throughout this chapter. In itself, this is a useful insight: should any learning model of this variety be examined through single runs? In the environment is based on random inputs, even random within a specified structure, and especially when there are multiple cues represented, sensitivity analysis would seem to be the only way to gain confidence in average behavior.

Moreover, this improved structure allows us to experiment much more widely: where we were trapped within a set of fairly strict parameters, limited to the values set in the empirical work, this improved structure creates a laboratory for creative discovery. In practice, we are able to adjust base rate, feedback, uncertainty, value structure and cue
weights, individually or in any number of combinations. This allows us to better replicate the kinds of conditions that policy makers are likely to face in the real world, as well as replicating more extensive experimental conditions.

Furthermore, we are able to back up a lot of the prior conclusions we made about behavior, based on the empirical investigations and the earlier rendition of the model, with greater confidence in the results. Runs indicate that model behavior finds an optimal threshold range, with only slight deviations when the model becomes more complicated by varying levels of uncertainty, payoff and feedback. Unlike in earlier tests, this model version has cue weights that are manipulatable, so additional testing across cues and weighting structures are possible.

Even more interesting, though are the insights derived about how people consider their threshold choices. Smoothing, as a behavioral construct, allows us to manipulate the simulation’s “memory” in a way that replicates human memory within a task. Moreover, we are able to consider different smoothing variables during different decisions within the model: how does our memory of prior cutoff values differ from our memory of whether we moved our threshold up or down last time?

This improved model also lends additional insight into more practical questions: What might people do when they are facing difficult challenges, especially ones that compel them to make tradeoffs, alter their decision threshold, and choose from murky options? Further research in this vein might examine how their own “smoothing times” or memory functions allows them to make more accurate decisions.
9. Discussion and conclusions

The purpose of this investigation has been multifold. In the previous pages, I have examined the literature, and identified principles and modeling rules based on the large volume of prior research. I have identified the key elements of decision maker behavior, based on an analysis of recent experimental data, applying, wherever possible, those elements to a current problem in public policy and model decision making: the significant rise in caesarean deliveries. Finally, I presented a simple model, based on earlier attempts, that combines error correction and hill-climbing principles. I conclude here by offering some general conclusions about what we know about this sizable intellectual landscape, in light of the distantly related, voluminous literature. And I will offer some prescriptions, based on empirical and model based exploration, about where we are at, and what we still need to uncover, in order to make better decisions.

Throughout this work, I have also referred to a particular practical example of our struggle to make good decisions in an uncertain environment: the dramatic rise in the rate of caesarean births. There are no demons here, only rational actors, trying to minimize particular errors and maximize rewards, which, in this case, are safely delivered babies. There are many reasons why this fascinating problem is a reasonable result of our policy decisions, and our value system. For the purposes of this paper, I have attempted to set aside the details of what has become a very emotional debate for the physicians and birth attendants, and the women and their families. I have assumed that people are doing their best, within a complex and complicated environment, to accurately judge cues, identify thresholds and make decisions that are best for all involved.

9.1. Overview of results

I began with the following questions, and revisit them here. First,

Under what conditions does feedback improve judgmental accuracy and under what conditions does feedback improve accuracy of threshold placement?

Research has indicated that we can create a comprehensive cue judgment strategy, using various types of feedback and induce greater accuracy (Summers, 1962; Hammond and Summers, 1965; Summers, 1972; and Lindell, 1976). Additional literature suggests that various strategies exploit the feedback available, in order to garner greater accuracy: Harvey (1995) suggests a simple anchor and adjustment strategy not unlike the dynamic decision rule proposed in Kubovy and Healy (1977). Einhorn and Hogarth (1978) remains one of the most critical pieces of scholarship, as they propose, among other useful hypotheses, that our ability to respond to feedback is heavily dependent on receiving positive feedback because judges are notoriously unable to adequately utilize negative feedback.

The modeling insights presented here (Figure 41), and confirmed by the empirical work (Figures 13-15), suggests that limiting feedback has less of an effect on accuracy than
previously assumed. Model results suggest that the higher threshold result during conditional feedback is an inevitable result of the system structure.

Moreover, when there are additional complicating factors, such as changes in the levels of uncertainty, especially at a higher or lower base rate, limiting feedback becomes almost irrelevant. And, for the most part, the reasons for this are sound: at the high and low base rates conditions, there are often far fewer errors from which to learn from. This is backed up by the lab model as well, which replicates those same findings, even when there is a random environment.

It has been tempting to assume that particular problems would have better outcomes, if only the decision makers had access to more feedback. Model results suggest that such an assumption is not necessarily true, especially at low base rates. This is good news for a medical decision making community often faced with limited feedback problems, since they can now rule out one complication as a source of judgment errors. Furthermore, the type of feedback is critical: providing cognitive feedback in a form that allows decision makers to actively consider choices and consequences is superior to simply providing knowledge of results. This is also good news, as this represents an important lever with which to improve outcomes.

How does feedback affect confidence in judgment and threshold placement?

This is a fascinating question – and obviously one that deserves substantially more investigation than is offered here, as there are tremendous gaps in the literature. Sharp, Cutler and Penrod (1988) challenges the assumption that simply providing feedback will reduce confidence and improve accuracy, as intuitive, but generally not supported. Importantly, even in conditions of limited feedback, confidence and consistency will increase (see Pease and Sniezek, 1992; and Fischer, 2005). Subbutin (1996) suggests there may be additional complexity: underconfident decision makers may be aided by additional feedback, but overconfident decision makers will not.

The model suggests that there is often quick, early learning, even without full feedback, then a decision to settle on a threshold. See, for example, the comparisons in Figures 46, 48 and 49, where there are large early threshold shifts towards the optimal threshold, and then fewer and less dramatic shifts during the second half of the simulation. This is supported by the empirical work as well, as indicated in the same figures. The model’s simple learning rules - to find a best threshold that balances the rewards gained from correct responses and different errors, with the greatest accuracy – tend to fit the overall learning patterns indicated in the empirical work. As shown in the lab model, smooths as a depiction of human memory have proven to be a useful construct. Here we can replicate a moving average, indicating a decision maker’s ability to remember the history of changes, without the weightiness of assuming that someone can remember the details of the last 10 trials. Here we assume, like modelers before us (See Camerer and Ho, 1998, for example), that we are incrementally updating our experience.

One future modeling challenge would be to unexpectedly alter the base rate after the
simulation has settled on a particular strategy, and examine how quickly it responds, and how fast it settles on a new strategy. These results could shape a new round of similar empirical testing.

**What role does implicit learning play in improving judgmental accuracy and threshold placement?**

This is a complicated relationship that a number of authors have addressed. Dienes and Berry (1997) profess to have uncovered a simple strategy that suggests why implicit learning is happening: even in situations where judges express an inability to understand how learning is happening, they seem to rely on certain cue-outcome combinations, and then generalize to a broader strategy.

Dalgleish (2006) suggests that there might be an interesting interplay between confidence and our ability to “fill in the blanks”, when there is limited feedback and only a tenuous grasp on an organizing principle: as we gain confidence, we will necessarily assume greater knowledge, and limited feedback matters less and less when a decision maker is sure of his strategy.

And, again, the empirical work presented here does not indicate that limiting feedback has a substantial effect on accuracy, as smart decision maker simply settle on a particular strategy, discounting negative feedback, and “fill in the blanks” when feedback is missing. This is a useful strategy, especially in high stress environments that do not lend themselves to repetitive analysis; unfortunately, in the case where thresholds have drifted from optimal, it also means that it will be difficult to convince decision makers to readjust their strategy. This is vividly clear from the lab model results: although we are able to achieve convergence, there is often about 5% of runs that depict a run where the model has gotten so far off track that it cannot find it’s way back to optimal.

**How do we conceive of a multiple cue threshold learning model capable of reproducing human behavior when faced with multiple cues and a threshold placement requirement?**

There has been a number of models constructed over the past several decades, most of which are interested in some piece of the problem presented here. In general, the simple model offered is able to mimic the strategies employed by some human decision makers. It integrates multiple cues and a mechanism for changes in the threshold based on both the rewards for correct decisions as well as accuracy. And, as shown, it reproduces human behavior fairly well, indicated that decision makers are probably using a hybrid strategy that takes advantage of both hill climbing and error correction.

There is some evidence that there are pieces missing – that even hill climbing and error correction together are not giving us the full picture. We don’t really understand how and why people learn so quickly in the first few blocks of empirical testing, for example. It is also clear that both the current model, and experiment participants, tend to settle over time. There is some literature that suggests why this is: confidence grows and a strategy
becomes firmly grounded. Further research is needed to explore how quickly decision makers would adjust if there was evidence that the current strategy was producing inferior outcomes.

This investigation has made some strides in this regard, though: as shown in Figure 67, we are able to determine that alteration of the cue structure has obvious and immediate effects on the threshold strategy. Similar to and earlier suggestion, another useful experiment could test how changing the cue structure after a simulation has already begun might affect behavior. How quickly could it readjust? From a policy perspective, we are interested in how quickly physicians and their patients might adjust, if given more accurate tools or education.

9.2. Conclusions

With these practical insights in mind, there are several conclusions that can be drawn with regard to the dilemma of caesarean deliveries. Holding aside, for the purposes of this conversation, those physicians who may be interested in making more money by scheduling caesareans, or those physicians who are oversensitive to potential liabilities, one is still struck by the difficulty of the environment, from the perspective of an obstetrician.

As pointed out above, base rates do matter – and since fewer than 15% of women are candidates for caesareans – the low base rate should aid in making good decisions. One assumption doctors could make, after all, is that most laboring women do not represent signal events. Unfortunately, since the threshold seems to be set at around 30% of all laboring women, policy changes would need to help adjust the rate down to a more reasonable level, a formidable task in itself.

Other aspects of this problem make judgment and decision making far more complicated. Feedback, in particular, represents an interesting dilemma – clearly, in the case of physicians, there is a lack of feedback: they will only get feedback after the worst possible error – when a delivery should have had surgical assistance and did not, resulting in injury or death to the mother or child. Both empirical investigation and modeling suggests that feedback, and the lack of it, has significantly less of an effect on accuracy than previously assumed. This, and the fact that there is a tendency to learn early, and then settle to a compromise position, is somewhat alarming for those of us interested in improving medical outcomes.

Recall, too, that according to prior research, and supported by the empirical and modeling work offered here, one of the reasons why learning seems to happen early and then pause, or settle, on a particular strategy or threshold is that enough experience and confidence as accumulated. Errors, at this point, are having very little effect, and what effect they may have will push the threshold in the wrong direction. Specifically, false alarms are rarely even registered as errors, and have little effect at reducing the > 30% caesarean rate; a
false negative error resulting in the death or injury of a woman or child does register, but will only push the threshold down and lead to even more caesareans.

Additionally, there are no repercussions for false alarms errors; indeed, caesarean sections are often billed at a higher rate, leading to a potentially perverse incentive to perform more. Moreover, the values placed on misses matters two-fold – not only does it force a physicians hand in individual cases, but politically, it makes it very hard to convince the establishment that rule changes are necessary. Regardless of how such a sentiment is phrased, encouraging physicians to alter their thresholds to allow for fewer false alarms also raises the specter of some potential misses. As popular opinion rejects any increase in the degree of morbidity or mortality caused by raising the threshold, especially considering the already dismal rates of US infant and mother morbidity and mortality, and appears to not be as concerned with the morbidity and mortality incurred by false alarm c-sections, any changes will be difficult and unlikely.

The situation is far from hopeless, however. Improving judgment in medical decision making has had tremendous benefits (see Swets, Dawes and Monahan, 2000, for example), and would undoubtedly improve caesarean rates. Greater awareness of the particular cues, and the types of cognitive feedback that would make those cues more accurate, would undoubtedly affect decisions, even in high stress hospital births. Even early modeling results shown here indicate that during low base rate conditions, making more of an effort to reward physicians who make correct decisions and find a more optimal threshold is more effective that penalizing the errors.

Unfortunately, awareness of the high rate of caesareans as a problem has to come first, as well as a willingness to accept different decision making strategies. Luckily, there is some guidance, as other medical decision making dilemmas have received a great deal of attention, with some conclusive results. Quanstrum and Hayward (2008), in a recent article in the New England Journal of Medicine, respond to the controversy regarding altered mammography screenings. They offer strong support to a view that uses evidence-based practices and expert panels as the logical guide to policy formation. They propose a methodology that assumes a continuum of care, from clear treatment to no treatment. Critically, they insist that there is a wide swath of discretion between those two poles; the sharp dichotomy of mammography screening, beginning at age 40, is inappropriate and ignorant of current research.

Clark (2008) addresses similar concerns, with regard to the obstetric specialty:

In short, our specialty is faced with unsatisfactory perinatal outcomes and an ongoing malpractice crisis despite increased costs and more cesarean deliveries. (Clark 2008)

The proposals to address this challenge are similar. Among several recommendations, Clark (2008) suggests that there should be uniform processes and procedures in place, as variation tends to lead to poor practice and lesser quality care. Clark (2008) also suggests that peer review is an intrinsic and very important feedback component to quality care. Yet, it is important to note that there is still little indication that physicians have an understanding of the environment – how the selection rate for caesareans is directly
related to the environment, to the reality of inevitable error, and the intractable situation that develops as soon as the threshold creeps lower. Should this be made clearer, it may have a direct effect on how we shape our policies.

To conclude, this is a huge intellectual landscape, with various fields contributing in many ways. The goal here was to illustrate, through an assessment of the literature, and through model building, supported by empirical work, that key principles can not only be derived from such disparate forces, but molded into a cogent whole. Such a practice not only provides the research pieces with more force, but it allows us to bridge the gap between research and practice.
Appendix A

Figure A: High uncertainty (dprime at .5)

Figure B: Medium uncertainty (dprime at 1)

Figure C: Low uncertainty (dprime at 1.5)
Appendix B: Guidebook to model scenarios

Summary: There are two different models that can be run to explore the structure of this model. This dissertation is based on one that extracts the data from exogenous cues, the same cues that are used in the Stewart (2010) experiment. There is also an additional more general version of the model, which relies on endogenous cues, drawn from random distributions within the model. The latter will be used during the defense to explore the model structure and behavior.

Both versions have the identical structure. Both require the user to specify several parameters, though sometimes in slightly different ways:

**Base rate**
The exogenous cue model draws the cue values from an external spreadsheet. In Vensim, clicking on the variable “Environmental distribution choice” opens the equation.

GET XLS DATA( 'Data for cues.xls', 'Data', 'B', 'N2')

The last alpha numeric value – highlighted in the equation above, specifies the particular base rate. When running the full feedback condition, use M2 for .5 base rate, K2 for .1 base rate and N2 for .8 base rate.

The endogenous cue model is more simple – just click on base rate and choose a base rate between 0 and 1.

**Feedback**
Feedback can altered in several ways: the user can specify full feedback, one of two types of conditional feedback, or no feedback. For either model, simply choose the base rate, and then click on “Feedback condition”. If feedback is set to 0, then there is no feedback. If set to 1, then the model offers only conditional feedback, specifically feedback when the decision is positive. If set to 2, the model offers only when the decision is negative. If set to 3, then the model offers full feedback.

**Uncertainty**
In the endogenous cue model, specifying uncertainty is also a function of the “environmental decision choice”. Recalling the equation above:

GET XLS DATA( 'Data for cues.xls', 'Data', 'B', 'N2')

Replace N2 with the base rate and level of uncertainty of your choice. For a higher R and less certainty, insert either Q2 for .5 base rate, O2 for .1 base rate or R2 for .8 base rate. For more uncertainty, and a lower R, select I2 for the .5 base rate, G2 for the .1 base rate,
and J2 for .8 base rate.

For the endogenous cue model, simply adjust the level of d`. The default is a moderate level of uncertainty of ~1. Lower amounts, such as .5, increase uncertainty, while higher amounts should decrease the amount of uncertainty implicit in the task.

**Value structure**

The value structure is the same regardless of the model. Simply click on "Value on signal, correctly identified (true positive)", "Value on noise, correctly identified (true negative)", "Value on signal, incorrectly identified (false negative)", or "Value on noise, incorrectly identified (false positive)". The default values award 10 points for each correct identification and subtract 5 points for each error.
Appendix C: Stepping through the model

As it is crucial that the reader understand what the model is doing, from initial cue judgment through threshold setting, the following section attempts to recreate, using the lab model, the steps involved in moving from judgment to decision to judgment.

1. Begin with an initial judgment. Recall that cues are determined by distributions installed within the model. A shown in the figure above, the components of the cues can be altered, so the shape and size of the cue distributions can be changed to match particular cue circumstances. In all cases, the three cues – Cues 1, 2 and 3 – and their respective weights are combined to form a weighted cue average.
2. In the Payoff screen, the weighed cue average is compared to the smoothed threshold. (During the first trial, the weighted cue average is compared to the initial threshold.) If the weighted cue average is higher than the threshold, we consider it a signal trial; otherwise, it is a noise trial. As the environment determines whether each trial is actually a signal or noise trial, the model determines whether the simulated decision maker was correct or not, during this trial, and as shown in the Payoff screen above, determines reward or penalty.
3. The changes in rewards and threshold (there would be only slight changes during the first trials, as we set an initial threshold) is determined using Busemeyer’s h function, or the change in payoff multiplied by the change in threshold, which determines the next direction of the threshold. The next direction is accentuated by the step size, which grows and recedes in response to feedback about accuracy.

To make this even more clear, the chart below is indicative of 20 trials, about half way through the run of 500 trials, at the .1 base rate, for a full feedback condition:
<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>217</td>
<td>4</td>
<td>7</td>
<td>Noise</td>
<td>Noise</td>
</tr>
<tr>
<td>218</td>
<td>5</td>
<td>7</td>
<td>Noise</td>
<td>Noise</td>
</tr>
<tr>
<td>219</td>
<td>4</td>
<td>7</td>
<td>Noise</td>
<td>Noise</td>
</tr>
<tr>
<td>220</td>
<td>5</td>
<td>7</td>
<td>Noise</td>
<td>Noise</td>
</tr>
</tbody>
</table>

As stated earlier, the model calculates a weighted cue average, determines whether it is correct or not by comparing the simulated decision to the environment, and awards a payoff. An indicated next direction is determined, and the smoothed next direction slowly creeps up or down, in an attempt to get as close as possible to an optimal cutoff.
Glossary
Definitions, jargon, and technical terms, as used in the above draft.

**Base rate**: frequency of signal events in the environment.

**Conditional feedback**: Refers to the conditions when knowledge of results is not available after every trial. Feedback, for example, may be provided after only signal, or “yes” responses, and it might also be offered after only “no” or noise events.

**Correct rejections (true negative rate)**: frequency of correct negative responses. In other words, events that indicated a surgical delivery was not necessary and it was not performed.

**Cutoff**: see threshold

\( d^* \) : represents a measure of uncertainty. As shown in Appendix A, \( d^* \) is equal to the mean of the signal plus the noise distribution minus the mean of the noise distribution.

**False alarm rate (or false positive rate)**: frequency of incorrect positive responses. In other words, the frequency of events that did not require a surgical delivery, but it was done anyway.

**Full feedback**: Refers to the condition when knowledge of the results of decisions is offered after every event, or trial.

**Hit rate (true positive rate)**: frequency of correct positive responses. In other words, events that indicated a surgical delivery was necessary, and a surgical delivery was performed.

**Miss rate (or false negative rate)**: frequency of incorrect negative responses. In other words, the frequency of events that did require a surgical delivery, but it failed to be done.

**Noise distribution**: representation of noise in the environment, as depicted as the negative distribution in the figures in Appendix A.

**Noise**: Noise is random variation that is always present. When the signal is not present, there is only noise. When the signal is present, there is signal plus noise.

**Optimal selection rate**: the selection rate that maximizes some criterion. For example, in the experiment described above, the subjects attempt to maximize rewards and minimize errors.

**Percent correct**: percentage of correct responses
**Points**: rewards accrued based on the payoff structure

**Threshold**: also referred to as the cutoff, the threshold determines the value of the judgment (see Appendix A) at which action is taken.

**Selection rate**: proportion of positive decisions.

**Signal distribution**: representation of occasions when the signal is present in the environment, as depicted by the positive distribution shown in the figures in Appendix A. The signal distribution (really the signal plus noise distribution) is assumed to be obtained by adding a constant (d’) to the noise distribution.

**Signal**: refers to the conditions where an event is present. In the example examined here, as signal event refers to circumstances when a laboring mother requires a caesarean in order to safely deliver her baby.

**Uncertainty**: refers to the degree of randomness or unpredictability in the environment, given the information available to the decision maker. In this investigation, uncertainty is measured using d’.


Chen, Yan and Yuri Khoroshilov. Learning under Limited Information. *Games and Economic Behavior* (44) 2003 1-25


Fischer, Ilan and D. Budea. When do those who know more also know more about how much they know? The development of confidence and performance in categorical decision tasks. *Organizational Behavior and Human Decision Processes* 98 (2005) 39–53


Kubovy M. and Healy, A F. The decision rule in probabilistic categorization: what it is and how it is learned. Journal of Experimental Psychology: General, 106, (1977)427 – 446


Lichtenstein and Fischhoff Do those who know more also know more about how much they know? Organizational Behavior and Human Performance. 1977. 20 (159-183)


Paul W. Pease and J. A. Sniezek Influences on the appropriateness of confidence in judgment: Practice, effort, information, and decision-making Organizational Behavior and Human Decision Processes, Volume 48, Issue 1, February 1991, Pages 100-130


Sharp, Glen L., Brian L. Cutler and Steven D. Penrod Performance feedback improves the resolution of confidence judgments Organizational Behavior and Human Decision Processes, Volume 42, Issue 3, December 1988, Pages 271-283


Tanner and Rauk. Signal recognition as influenced by information feedback. Journal of


