1-1-2011

Essays on applications of behavioral decision making in public management and policy

Navid Ghaffarzadegan

University at Albany, State University of New York, navidg@gmail.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, Public Administration Commons, and the Public Policy Commons

Recommended Citation

https://scholarsarchive.library.albany.edu/legacy-etd/345

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.
Essays on Applications of Behavioral Decision Making in
Public Management and Policy

by

Navid Ghaffarzadegan

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

Department of Public Administration and Policy
Rockefeller College of Public Affairs and Policy

2011
ABSTRACT

This is a three-essay dissertation on examination of behavioral decision making phenomena in the contexts of public management and policy analysis. I study how policymakers, government employees, and citizens behave, make decisions and learn from decision outcomes. I move toward prescriptive analysis and offer solutions that can improve government performance.

In essay 1, I study behavioral decision making among airport security staff in a screening task. I develop a simulation model of screening and compare the results to data from a laboratory study of screening. The model explains how decision making complexity, uncertainty, and feedback characteristics lead to systematic bias and hamper performance. The lessons from the model are applicable to other public service contexts including law enforcement, social welfare, and public health.

In essay 2, I study how citizens react to warnings issued by the government. The purpose of this paper is to examine the tradeoffs inherent in issuing security warnings and to produce a set of policies that identify the optimal timing for warning issuance. I develop a simulation model of security based warning, and conduct several simulation analyses of the model. The results show that an underestimation of behavioral characteristics of the society can result in sub-optimal solutions and performance shortfall.

In essay 3, I study physicians’ decision making. I develop a theory that offers a new explanation for sub-optimal decisions in medicine, and proposes a new approach for health policy design. With the help of a simulation model, I show that practice variation and bias, two indicators of sub-optimal medical decisions, do not have to be caused by physicians’ personality traits and financial incentives, but can endogenously emerge through daily practices and outcome learning even for physicians with similar trainings.
working in the same region. Specifically, a physician’s exposure to outcome feedback, a physician’s ability to evaluate different forms of practice and a physician’s accumulated experience with a given approach all contribute to practice variation and bias. A preliminary validation of the results is achieved by comparing projected results with the actual data from cesarean section surgery in the states of New York and Florida.
ACKNOWLEDGMENT

My deepest thanks are to my dissertation adviser, Professor David Andersen. He always supported me to explore new ideas, to question common assumptions, and to operationally think about social systems. His expertise in understanding and analyzing policy problems and his visionary approach to guiding advisees were invaluable assets to me. I will never forget our weekly meetings at the school, the Muddy Cup, Barnes and Noble, Starbucks, and CTG, or our regular Skype meetings when he was in Pueblo or Montreal. In-depth discussions with David sometimes went beyond policy research, reminding me *why we live, what we want, how we can help the public, and why family should be our first priority*. I hope one day I will become as good an adviser as David.

I am also thankful to my other committee members, Professors Thomas Stewart and Erika Martin. Tom’s high quality class on judgment and decision making was a cornerstone in my research career, and thereafter, I have always considered the behavioral element of policy analysis crucial. I became interested in understanding human decision-making through talking with Tom and reading his high quality papers. Complementing, Erika played a unique role in my dissertation. As an expert in policy research and analysis, her insightful comments at different stages of my dissertation, and especially on the third essay, were thought-provoking and helped me to improve my research approach to the policy problems.

I was lucky that my studies at the University at Albany were during times when many worldly known scholars were researching and teaching there. I am sincerely grateful to Professor George Richardson. His excellent system dynamics classes and continuous modeling guidance were invaluable to me. I am also appreciative to Professors Stephen
Weinberg, Karl Rethemeyer, Mitch Abolafia, John Rohrbaugh, and many others for what they taught me and for always being available to help.

During my PhD, I had the privilege of working at the System Dynamics Society Office. Roberta Spencer, Robin Langer, Erin Sheehan, Joan Yanni, LouAnne Lundgren, and Jen Rowe were wonderful colleagues and friends at the office. Especially, I am thankful to Roberta, the Executive Director of the Society, for always being a support, caring about my education, and helping to make my experience in Albany a success. I learned a lot from working with Roberta. Her patience and understanding helped me to overcome my daily challenges as a PhD student, and to enjoy my life. Thanks Roberta.

Throughout the experience, I was nurtured with friendship and love from many friends. Different pieces of the journey remind me of different friends; some of whom are Amir Karimi, Amir Sadoughi, Andrew Whitmore, Ardavan Zandiatashbar, Bita Behforooz, Burak Türkgülü, Deborah Andersen, Farzad Houshmand, Gail Richardson, Hamed Parvaneh, Hassan Dibadj, Hasti Amiri, Hyunjung Kim, Jisung Kim, John Lyneis, Kamiar and Mahnaz Alaei, Lucy Dadayan, Michael Landon-Murray, Mohammad Mojtabahzadeh, Mohammad-Ali Poursina, Newshaw Bahreyni, Nici Zimmermann, Rod MacDonald, Russell Hassan, and Shahram Pourmand.

I would like to especially thank my parents, Sima Samadi and Issa Ghafarzadegan, who always encouraged me to keep a high work ethic, stay with the values they taught me, and devote myself to helping people. I am thankful to Mahshid and Ali, my sister and brother-in-law, for being extremely supportive during the time I was far away from my parents.
Most importantly, I would like to thank my wife, Niyousha Hosseini. Her support and love have been at the heart of this journey. She always encouraged me to continue, and she always buoyed me when I missed my home, my family members, or when things simply did not go as expected. She, also, tolerated my lack of cooking skills!

Finally, this dissertation has been done during a difficult time for many people in the Middle East, several of them bravely fighting against dictatorships and crying for freedom, human rights, and human respect. I dedicate this dissertation to them, and the families who lost their loved ones.
# TABLE OF CONTENT

<table>
<thead>
<tr>
<th>Essay Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction to three essays</td>
<td>1</td>
</tr>
<tr>
<td>Essay 1</td>
<td>8</td>
</tr>
<tr>
<td>“Stop, Search, and (Do not) Learn: Barriers to Learning in Security Contexts”</td>
<td></td>
</tr>
<tr>
<td>Essay 2</td>
<td>64</td>
</tr>
<tr>
<td>“An Examination of Complexities of Warning Issuance for Domestic Security in a Long Term Dynamic Context”</td>
<td></td>
</tr>
<tr>
<td>Essay 3</td>
<td>100</td>
</tr>
<tr>
<td>“Beyond Personality Traits and Financial Incentives: Bias and Variation in Medical Practices as a Systematic Result of Experiential Learning”</td>
<td></td>
</tr>
</tbody>
</table>
INTRODUCTION TO THREE ESSAYS
There are a lot of examples of public services were public employees, policymakers, or citizens must cope with uncertainty, limited information, time pressure, and environmental complexities and should make critical decisions in these situations. For example police officers in airports deal with complex and uncertain decision making conditions. They should make their decisions in a very limited time and the decisions can have significant consequences. On the other hand citizens should make proper decisions when coping with emergency situations. Their perception of risk and their reactions to the threats are central in decreasing the vulnerability of the public.

Limited information and limited cognitive capabilities to process available information lead to bounded rationality. Herbert Simon as one of the main initiators of behavioral decision making studies has stressed the importance of considering human decision making characteristics in understanding and improving public administration (Simon, 1947; 1965; 1973). In such conditions taking an approach of understanding the human element in decision making, i.e., a behavioral decision making approach, helps to improve human performance and ultimately the organizational performance (Simon, 1947; Edwards, 1954; 1961; Tversky and Kahneman, 1974; Hammond, 1996).

Despite the criticality, few studies address decision making performance of public employees and citizens and their effects on policy performance. On the other hand, the behavioral decision making studies have rarely been applied to the public management contexts, and have offered limited policy implications for governments. In this dissertation, I follow the classic view promoted by Herbert Simon to improve government performance through understanding the human element of decision making (Simon, 1947).
This is a three-essay dissertation directed toward examination of behavioral decision making phenomena in the contexts of public management and policy analysis. Specifically, I am interested to understand the decision making factors that can lead to mistakes, inconsistencies and inaccuracies in management and policy relevant contexts, and propose the ways that can improve government performance.

In the first essay, I model and analyze the results of a laboratory experiment on airport security screening. The paper examines the role of behavioral decision making and learning in complex, high security environments and uses insights from the study to generate a set of strategies to help improve the decision making of airport security staff. I continue a similar theme in my second essay where I develop a simulation model of security in public places and investigate effects of different decision making assumptions on optimal policies. The study produces a set of guidelines that identify the optimal timing for the dissemination of warnings. The third essay goes beyond the security domain and examines behavioral decision making and learning in another critical domain, healthcare management. In this essay, I examine the problem of healthcare disparities by studying practice variation among medical experts; basically, addressing why different doctors have different practice styles. Through these essays, I develop insights into the effects of the behavioral reactions of citizens and government organizations on government performance.

Besides addressing crucial policy problems, these essays have many methodological and contextual commonalities. Specifically, 1) they all focus on behavioral decision making theories at the individual level, 2) they consider dynamic changes in decision models, 3) they are concerned with policy implications, and 4) they take a systems view to the problem, and all use some forms of mathematical modeling and simulation of governmental problems as well empirical investigation of the findings.
The essays take a multi-method approach, and combine system dynamics simulation (Forrester, 1961; Richardson & Pugh, 1981; Richardson, 1991, 1999/2011; Sterman, 2000) with empirical data, such as data from a laboratory experiments (essay 1), and quantitative data from the field (essay 3). System dynamics is a computer-aided systems approach to policy analysis. The method applies to dynamic problems especially when there are feedback loops in the system, what is known to be at the heart of the system dynamics modeling (Richardson, 1991). Diagrams of loops of information feedback and circular causality, which are the building blocks of systems perspectives, are common in system dynamics. In addition, mathematical formulation helps to simulate the model and conduct counter-factual analysis with the model, also known as what-if analysis (Zagonel, Rohrbaugh, Richardson, & Andersen, 2004; Stewart & Mumpower, 2004; Schuman & Rohrbaugh, 2004). In a simulation based counter-factual analysis, a simulation model is developed that replicates the current stage of the system, and then used to test different policies through changing different parameters in the model. The approach helps conducting experiments in a faster rate, and of course avoids social costs of field experiments (Sterman, 2000; Ghaffarzadegan, Lyneis, & Richardson 2011). Such a simulation approach is present in all of the three essays, and in many places it is combined with quantitative data from a laboratory experiment or from the field.

The dissertation contributes to the literature and our understanding of theories of governance in several ways. As it was briefly stated, all of the essays are dealing with very important policy problems. Currently, a huge attention has been paid to the government actions for security in public places. Studying screening and warning as two of the most common actions that a government takes can help a proper policy
development. Desired policies are expected to result in unbiased decisions and reactions from both government employees such as police officers (in the case of screening) and citizens (in the case of public warning). Therefore, studies that lead to understanding the behavioral components of these decisions and give proper tools to improve judgment and decision making are necessary.

Further, the third essay which applies behavioral decision making phenomena to the context of health contributes to the literature of health policy, healthcare management, public administration and medical decision making. The issue of bias toward overutilization of medical resources is directly connected to the problems of efficiency and effectiveness, major concerns in the literatures of healthcare management and public administration. Further health disparities are signs of inequity and injustice in the society and needs proper governmental actions to be solved. Therefore understanding and solving the problems of practice variation and bias in the healthcare context is very crucial. As I will discuss in more detail, the literature of medical decision making lacks common understanding of the roots of practice variation and bias toward overutilization of medical resources.

In addition, the project contributes to the literature of systems science, system dynamics and simulation by developing simulation models of behavioral decision making. Many of the current system dynamics models are highly aggregated (at the levels of societies or communities) which may be less relevant to the concerns of this project. For example, the issues of heterogeneity in practice style (the third essay) need a model that addresses the cognitive dynamics behind decision making of physicians. These
models are rarely developed and never applied to the studies of public management and policy analysis.

Overall, through these essays, I develop insights into the effects of the behavioral reactions of citizens, government employees and government organizations on government performance in three important contexts. Furthermore, the lessons on how human elements of decision making are important, and how to use simulation models to study human decision making in public administration and policy can be applied to several other similar problems that link with theories of management, policymaking and governance.

References


ESSAY 1:

Stop, Search, and (Do not) Learn: Barriers to Learning in Security Contexts
Abstract

Public services include abundant examples of detection decisions. For example, a police officer continuously gathers information in order to detect criminal activities, and must decide whether to act on that information. I analyzed decision making performance in an important detection task: airport security screening. I developed a simulation model of decision making and learning and explain decision making errors. The model was compared to the results of a laboratory study. The results explain how decision making complexity, uncertainty, and feedback characteristics lead to systematic bias and hamper performance. The analysis also shows how overconfidence develops over time, limiting the ability of decision makers to learn from experience. I discuss how such overconfidence results in employees’ performance shortfalls. The results suggest ways to improve the decision making of security staff. The lessons are applicable to other public service contexts including law enforcement, social welfare, and public health.

Keywords: public employee performance, learning, detection tasks, public safety
1.1. Introduction

Many public service decisions involve elements of detection. In their simplest form, a
detection decision is based on an initial judgment and involves choosing one of two
actions. For example, in airport security, police security officer detects a suspicious
airline passenger and decides to conducts a search or allow the passenger to pass. The
decision to subject an airline passenger to an interview and search is based on the
officer’s judgment of degree of suspiciousness. In their daily stop and frisk operations or
preventing drunk driving, police officers observe suspicious cars and must decide
whether to stop them for additional investigations. In the context of social welfare, public
employees with the role of allocating financial resources to needy families, need to detect
need among families in a pool of applicants and allocate aid to the truly needy. Other
examples of detection tasks include investigation of public organizations as required by
accountability laws, detection of tax frauds among a large pool of tax payers, or decisions
about hiring/promotion/layoff in human resources departments. In all of these examples,
and many other situations, decision makers gather information from the environment and
make a judgment that leads to a decision.

Effective execution of detection tasks is a policy implementation concern and is
tightly connected with the public service performance. Improving learning and
performance on such tasks improves the quality of public services. There are many
studies of human performance on detection tasks, but these have not been applied to
public management contexts. The usual approach to improving decision performance
relies on technology to provide better information in a more timely manner (e.g.,
automatic number plate recognition systems for police that feed vehicle license plate
numbers directly into law enforcement data bases). These technological initiatives are important, but they are only part of the solution. Ultimately, a human decision maker must process the information and make a decision. This paper focuses on the human decision maker and takes a behavioral approach to understanding and improving detection decisions in a public service context. We focus on airport security as a critical detection problem where errors can have serious consequences. Later we discuss the applicability of the arguments to several other domains of public services.

**Security and Public Safety Detection Tasks**

Security decisions, such as airport screening, are about gathering information from the environment and detecting suspicious cases in order to take protective actions. The purpose of airport screening is protecting the safety and security of airplanes and passengers and prohibiting the transfer of banned items such as illegal drugs. Several recent terrorist attempts to attack mass transportation systems such as the attacks in New York, Madrid, London, Glasgow and Mumbai have raised the international concern about proper passenger screening procedures. Currently, under the Passenger Screening Program of the U.S. Department of Homeland Security (DHS), more than 700,000,000 airport passengers per year are screened using technologies such as advanced X-ray, bottled liquid scanners, and advanced imaging technologies (DHS, 2010). In 2009, the United States allocated $266 million to the Transportation Security Administration of DHS to accelerate investment and deployment of screening technologies (DHS, 2010, p.6).
In addition to advanced technologies, passenger screening based on observations of passenger behaviors such as the apparent level of stress and fear is common. A good example is DHS’s Screening of Passengers by Observation Techniques (SPOT) program in airports. Under this program, Behavior Detection Officers focus on passengers’ behaviors and appearance indicative of stress, fear, or deception to identify ones who may pose potential security risks. Individuals whose behavior passes some thresholds may be referred for additional screening and intervention (DHS, 2008). While the SPOT started in 2003 with a considerable federal budget ($212 million in 2010), no consensus exists about the effectiveness of such observational techniques for improving airport security (Government Accountability Office, 2010).

SPOT is a security based screening plan that depends heavily on human perception and judgment. Such dependence on perception and judgment is typical of most security related problems (Bazerman & Watkins, 2005; Sagan, 2004; Pinker, 2007; Ghaffarzadegan, 2008; Ghaffarzadegan & Andersen, 2009). Even if highly technological screening procedures are used, in the end, it is a person who interprets the information and decides if additional searching is necessary or not.

**Complexities of Security Detection Tasks**

Security officers must cope with uncertainty, limited information, time pressure, and complexity. In addition to these factors, the way that security officers receive feedback about the results of their decisions can limit their ability to make optimal decisions. In this paper, feedback is referred to information about the decision outcome. In many repetitive decision making tasks, a decision maker observes the results of her decision.
For example, if a police officer decides to stop a car to test for the blood alcohol content, after stopping, she will know if her decision to stop the car was a correct decision or not. The behavioral decision making literature has argued that learning from outcome feedback (here, the result of stopping a car) is not a straightforward cognitive task and it is often difficult for people (here, officers) to learn from feedback (Klayman 1988). In the case of airport screening to prevent illegal drug transfers and many other similar decisions, outcome feedback is not always available and is contingent on decisions. In this case, officers will know the result of their decisions only if they make a positive decision (here, to stop a car). When they make a negative decision (here, decide to let a suspect go without additional tests) they will not know if their decision was correct. This is referred as conditional feedback in the literature (Stewart et al. 2010) because outcome feedback is conditional on decisions and is usually limited to positive decisions.

Table 1.1 offers a few examples of security detection tasks and shows how feedback characteristics can be different across these decisions. In airport screening, searching to prevent illegal drug transfer is a conditional feedback situation where officers receive feedback if they decide to search a suspect. Most people, of course do not carry any drug (conditional feedback on low base rate). On the other hand, in a public place, such as a governmental organization, or a public event such as a concert, safety staff can refuse to admit people who seem trouble makers. In such a condition, feedback will be available about people who were allowed to participate. In such a situation a higher portion of cases provide feedback (conditional feedback on high base rate). It is similar for the case of visa issuance in the borders, as at least government will know what percentage of people who were issued a visa turned out to be trouble makers. Of course, for the people
who were not allowed to enter in the first place, no feedback about their true status is available. In contrast to these examples, in the case of airline bombing, security staff will receive feedback, no matter what their decision is: if they do not search a person who is carrying a bomb, they will later hear about the bombing news. This is referred as a full feedback condition.
<table>
<thead>
<tr>
<th>Examples of security detection tasks</th>
<th>Positive decisions</th>
<th>Negative decisions</th>
<th>Decision Characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Conditional feedback on low base rate: Airport screening to prevent illegal drug transfer</strong></td>
<td>Search</td>
<td>No Search</td>
<td>Feedback is only limited to the positive decisions. Lower proportion of decisions are positive than negative</td>
</tr>
<tr>
<td><strong>Conditional feedback on low base rate: Police officers’ stop and frisk actions</strong></td>
<td>Stop and question</td>
<td>No stop</td>
<td>Feedback is only limited to the positive decisions. Depends on locations and days. Usually, lower proportion of decisions are positive than negative</td>
</tr>
<tr>
<td><strong>Conditional feedback on high base rate: Public admittance to public events</strong></td>
<td>Admit</td>
<td>No admit</td>
<td>Feedback is available about people who are admitted. Usually, higher proportion of people is admitted</td>
</tr>
<tr>
<td><strong>Conditional feedback on high base rate: Visa issuance/Admittance to enter the borders</strong></td>
<td>Issue/accept</td>
<td>No admit</td>
<td>Feedback is available about people who are admitted to enter. Usually, higher proportion of people is admitted</td>
</tr>
<tr>
<td><strong>Full feedback: Airport screening to prevent airline bombing</strong></td>
<td>Search</td>
<td>No Search</td>
<td>Feedback is available for both positive and negative decisions. Lower proportion of decisions are positive than negative.</td>
</tr>
</tbody>
</table>
Current Study

As stated, even in highly technological screening procedures, in the end, it is a human who should make a decision. In such conditions taking an approach of understanding the human element in decision making, i.e., a behavioral decision making approach, helps to improve human performance and ultimately the organizational performance (Simon; 1947, Edwards 1954; 1961, Tversky and Kahneman, 1974; Hammond, 1996). This paper looks at how behavioral components of decision making can play a role in decision making pitfalls.

In this paper, we develop a mathematical model of airport screening to analyze the results of a laboratory experiment. The paper examines the role of learning in a high-stakes security environment. In the following, first we review the relevant theories. Then we describe our method and data. Next, we describe the simulation model of screening and examine the extent to which the model replicates the laboratory experiment results. Then, based on the model we conduct a set of simulation based analysis. Finally, we discuss how different management tools can help improve decision making in airport security.

1.2. Theory: Experiential Learning in Detection Problems

We build our analysis based on theories of detection decisions, and learning from outcome feedback.

Detection framework

Detection theory, also known as signal detection theory (Macmillan & Creelman, 1991; Green & Swets, 1966; Swets, 1991; Swets et al., 2000), divides the possible events
that are relevant to a particular decision into two categories, the ones we want to detect (positive events) and all others (negative events). For example, a police officer wants to detect guilty persons. In this theory, decisions are assumed to be based on a continuous score or judgment. For example, a decision to subject an airline passenger to an interview and search is based on the security officers’ judgment of degree of suspiciousness. In other words, officer makes a judgment based on the information that she receives from the environment about an individual. If the judgment exceeds some threshold (i.e., she is too suspicious), then a search is conducted. Otherwise the passenger is allowed to pass unchallenged.

In average, we can assume, police officers are more suspicious about guilty people that innocent people. In other words if we assign a number for how suspicious an officer is to each person of a sample of guilty people, and one number for how suspicious she is to each person of a sample of innocent people, the first sample will have a higher average of suspiciousness than the second sample. Figure 1.1 is a representation of detection theory for airport officers. The x-axis in Figure 1.1 represents suspiciousness. The distribution in the left depicts the distribution for innocent people, and the one in the right depicts the distribution for guilty people. It is shown the distribution of guilty people has a higher average suspiciousness than innocent people. However, due to environmental uncertainty it is not possible for an officer to perfectly distinguish all cases. Therefore, the distributions for guilty and innocent people overlap, and there will be innocent people for whom the officer is mistakenly more suspicious (see the overlap of two distributions in Figure 1.1). The detection framework assumes an officer has a threshold (the vertical line in Figure 1.1) for making a positive decision (e.g., search and interview a suspect)
and if the level of suspiciousness for a case exceeds that threshold, the officer makes a positive decision.

![Distribution of innocent and guilty persons in detection theory](image)

Figure 1.1: Distribution of innocent and guilty persons in detection theory

Figure 1.1 assumes normal distributions of suspiciousness for both innocent and guilty people (consistent with the original detection theory). The distance between the means of the two distributions is labeled $d'$.

The portion of the distribution of innocent people that falls above the threshold represents mistakes, i.e., innocent people who are stopped or arrested. Similarly, the portion of the distribution of guilty persons that falls below the threshold represents another kind of mistake—guilty people who are not stopped or arrested.

In any binary decision making situation, there are four possible decision outcomes. You can say "yes" and be right or wrong or you can say "no" and be right or wrong. These outcomes are often labeled true and false positives and true and false negatives. As
Table 1.2 shows, a police officer can decide to stop a person (a positive decision) and the person maybe guilty (true positive or “hit”) or innocent (false positive or “false alarm”). The police officer can also decide not to stop the person (a negative decision). And again the person can be guilty (false negative or “miss”) or innocent (true negative or “correct rejection”). Thus, there are two kinds of errors: false positives and false negatives.

<table>
<thead>
<tr>
<th>Decision</th>
<th>YES</th>
<th>NO</th>
</tr>
</thead>
<tbody>
<tr>
<td>YES (guilty)</td>
<td>False positive (false alarm)</td>
<td>True positive (hit)</td>
</tr>
<tr>
<td>NO (innocent)</td>
<td>True negative (correct rejection)</td>
<td>False positive (false alarm)</td>
</tr>
</tbody>
</table>

Table 1.2: Four possible outcomes

An important point is that different thresholds impose different error rates, and as the probability of one error decreases, the probability of the other error increases (see Figure 1.1). For example, if the threshold is increased (moved to the right) false positives will
be reduced, but false negatives will increase. Unless the distributions can be moved further apart (increasing d’), it is impossible to simultaneously decrease both errors by changing the threshold. It is expected for officers with a higher level of judgment expertise, d’ to be larger.

In this framework, the proportion of positive decisions is called selection rate (e.g. if 50 percent of people are selected for searching, selection rate is 0.5). On the other hand, the proportion of positive events (e.g., the proportion of suspects who are guilty) is called base rate (e.g. if 50 percent of people are guilty, base rate is 0.5).

For a given set of values/costs for each cell in Table 1.2, there is an optimal threshold location. An important research questions is whether individuals will learn this optimal threshold through experience.

**Learning in detection problems**

The learning process based on feedback about the results of decisions is called experiential learning (Cyert & March, 1963). Outcome feedback can help people learn to make more accurate judgment (increasing d’ in the detection framework) and it can help them learn better decision thresholds. In this paper we address the problem of threshold learning through experience (outcome feedback) in repetitive tasks.

Several mathematical models of threshold learning in binary choice problems have been proposed. One of the earliest models is the error correction model that assumes people learn from their errors and modify their thresholds in respond to feedback from false decisions (Kac, 1962; Kac, 1969). In these models threshold changes in pre-defined and constant steps and in the direction that the last information cue was observed.
Kubovy & Healy (1977) conducted more analysis on error correction models and find they are inadequate in explaining decision making especially after feedback about success.

More complicated learning models are also offered. The hill climbing model (Busemeyer & Myung, 1992) and the reinforcement learning model (Erev, 1998) are two of the most discussed models. In the hill climbing model, the main idea is that the model tests possible solutions that are close to the current solution (searches the neighborhood of the current solution) in order to find a better answer, and continues until it is assured that the current answer is the best possible decisions. In the reinforcement learning model, individuals are assumed to assign a value for each member of a range of possible thresholds, modify those values after each trial, and then randomly choose one of the thresholds based on probabilities determined by the values. While most of these studies have compared their models with the past models, no universal agreement exists on which model works better, especially in a wider range of cases of decision making and in different conditions.

Although these studies make different assumptions about how people learn, they share an important assumption about the nature of feedback. They assume full feedback, that is, a decision maker always receives feedback about whether a decision was right or wrong. But full feedback may be the exception outside the laboratory.

*Conditional feedback in detection problems*

As stated, in many real world situations, decision makers receive feedback only for positive decisions. There are very few studies in the domain of behavioral decision
making on conditional feedback, most of which have not been applied to public management problems yet (Einhorn & Hogarth, 1978; Tindale, 1989; Elwin et al., 2007; Dalgleish & Smillie, 2006). Einhorn and Hogarth (1978) claim that, in general, in conditional feedback (they call it partial feedback) people receive less disconfirming information that they would receive in a full feedback condition, so they become trapped in the illusion of validity. They do not empirically test their arguments. Fischer and Budescu (2005) make a similar argument based on a set of experiments. They study learning and development of confidence under different types of feedback and in different base rates. They argue that under full feedback (they call it a discrimination task) people are able to learn quickly and there is a high correspondence between their performance and confidence, but in conditional feedback (they call it a screening task) it seems there is less correspondence between performance and confidence.

The type of feedback also affects performance in a learning task. Tindale’s (1989) study of 48 trials of decision making in a recruitment task with eight information cues and feedback resulted in better individual performance under conditional feedback, but other studies show the opposite. Elwin et al. (2007) and Henriksson et al. (2010) investigate empirically the effects of conditional feedback on decision making. While observing that people underestimate the base rate (the ratio of signals to total observations), they argue that people assume their negative decisions, for which they do not receive feedback, are true. In Appendix 1.1, I question the validity and the generalizability of this argument in high base rates, as it imposes an ever-declining threshold. Overall, while studies of conditional feedback are increasing, there is little

---

1 During finalizing the dissertation, A newer draft of Appendix 1.1 got published as a commentary in the Journal of Experimental Psychology: Learning, Memory and Cognition. In the published manuscript we
agreement about how decision makers learn under conditional feedback, and studies are not applied to the public management contexts.

1.3. Method

Data description

This study is a part of a larger project on threshold learning and uses data from an extensive laboratory experiments conducted at [University Name]. Participants were graduate and undergraduate students of [University Name] and members of the community. They were paid between $20 and $30 based on their performance. The experiment took up to 90 minutes.

The experiment used a simulated airport screening task. Participants were asked to assume the role of customs inspector in a major international airport. They were shown 500 hypothetical passengers on the computer screen. On each trial, they were shown three information items about the appearance of a hypothetical passenger: clothing, emotionality, and bulging. Information on clothing represented the appropriateness of passengers clothing for the season and the possibility of wearing special clothing in order to hide smuggled items. Emotionality represented how nervous the passenger appeared when questioned. Bulging referred to bulges on the passenger that may relate to hidden items. As Figure 1.2 shows, each information item was presented as a number between 0 and 7. The numbers varied randomly for each case.

also offer a few suggestions to improve constructivist coding. For complete discussion please see Ghaffarzadegan and Stewart (2011).
Figure 1.2: Example case: A passenger is represented by three information items.

The subjects’ task was to judge the suspiciousness of the passenger on a scale of 1 to 10 (10 most suspicious), and then to decide whether to allow the passenger to pass or subject the passenger to further searching. In order to limit the experiment to a threshold learning experiment and avoid the need to learn about the importance of each information cue, the subjects were provided “ideal weights” to use in their judgments. The weights provided were: Clothing (20), Emotionality (50), and Bulging (30) representing the proportional importance of each item. A bar chart of these weights was available to subjects as a reminder during the experiment. Knowing the proportional weights, the subjects’ main task was to learn the appropriate threshold for making a search decision.

Outcome feedback was provided in different forms to facilitate learning. Our focus on the part of the experiment that involves full feedback and conditional feedback.

Under the full feedback conditions, participants were shown the correct decision after each decision. Full feedback represented conditions where an experienced inspector was helping them to learn and giving them the correct answer after each decision. Under conditional feedback, subjects were notified about the correct decision if their decision
was to search the hypothetical passenger. Conditional feedback represented a more realistic experience whereby outcome feedback is provided only in the cases that officer decides to search a passenger; otherwise the passenger goes away without revealing his or her status. Subjects received points for their decisions: +100 for each correct decision, and –50 points for each incorrect decision. As a performance incentive, a bonus payment was provided based on the subjects’ score.

Correct outcomes for the 500 trials were determined based on the ideal weights (20, 50, 30) adding some random noise to the system (normally distributed) representing real world uncertainty due to imperfect information. The subjects were informed about imperfectness of information.

The study examined different base rates, i.e., the proportion of people who actually should be searched. Base rates were 0.1, 0.5, and 0.8. Different base rates can represent a wider range of conditions that police officers face every day (refer to Table 1.1). There were 143 subjects in the experiment, with at least 18 subjects in each of 6 (2 feedback conditions X 3 base rates) conditions.

Analysis of the laboratory experiment shows discrepancies between full feedback and conditional feedback. Figure 1.3 shows that the selection rate is lower for all base rates with conditional feedback than with full feedback. This difference is greater for 0.5 and 0.8 base rates than for 0.1 (base rate by feedback interaction F(2, 137) = 6.82, p < .001). Conditional feedback also resulted in a significant decrease in performance as measured by the percent of correct decisions (F(1,137) = 11.605, p < .001). This is an important finding in the experiment as conditional feedback represents a situation close to the real world screening experiment whereby feedback is limited to the cases that officers
conduct searching. In other words, the experiment reveals possibility of systematic bias in airport screening.

![Figure 1.3: Selection rate for last 200 trials for 6 different conditions: three base rates (0.1, 0.5 and 0.8) x two feedback types (full and conditional feedback)](image)

The observed bias in conditional feedback raises the importance of why such a bias emerges and how it can be mitigated.

**Mathematical modeling**

A mathematical model that can replicate the dynamics of decision making under conditional feedback can help understanding why the observed biases emerge in
conditional feedback and in all base rates. Further, a public management approach requires a normative investigation of the question and finding ways to mitigate the problem. A simulation based counter-factual analysis approach can provide the needed insights. Counterfactual analysis (also known as what-if analysis) has been used as an important tool for policy analysis avoiding costs of experimenting in real world and to perform tests in a faster and less expensive rates (Zagonel et al., 2004; Stewart & Mumpower, 2004; Schuman & Rohrbaugh, 2004; Ghaffarzadegan et al., 2011). We conduct a set of simulation based analyses and, based on the results of those analyses, propose management tools to improve security screening tasks.

Our modeling approach process as follows: 1) modeling, 2) replication of data and model calibration, and 3) simulation experiments (counter-factual analysis). We first develop a discrete, stochastic simulation model of threshold learning. The goal is to capture the dynamics of decision making and threshold adjustment in full feedback and conditional feedback. Then, we examine the extent to which the model replicates the data. Later, having a mathematical model that represents how individuals make decisions under screening tasks, we can move toward recommendations through simulation experiments and counterfactual analysis.

1.4. A Model of Airport Screening

We develop a model of airport screening that is general enough to represent a wider range of security detection tasks (as was illustrated in Table 1.1) by changing a few parameters. The model overview is shown in Figure 1.4, and the detailed formulation is offered in Appendix 1.2.
The basic logic of the simulation model is as follows: The model assumes each officer has a decision threshold where if their suspiciousness (judgment) about a passenger exceeds the threshold they search the passenger. In each time period, the model creates a one-dimensional judgment, based on the provided information cues from the environment, representing the officer’s judgment about a passenger. Then the model makes a decision by comparing the decision threshold with the judgment. If the judgment exceeds the decision threshold, the model makes a positive decision (search), and if not it makes a negative decision. The decision is followed by decision result; it is either correct or wrong.
The decision outcome is formulated through using the two-by-two table shown in Table 1.2. Then, as Figure 1.4 shows, the model has a perception of decision outcome either through feedback from decision outcome or from self confidence. Under the full feedback condition, it receives the result of the decision; under the conditional feedback situation, the model receives the results only if the decision was positive, but guesses the results based on confidence if the decision was negative. Finally, based on the perception of decision outcome, the model corrects its decision threshold if it is necessary.

Detailed formulation of the model is as follows. Vensim coding is offered in the Appendix 1.2.

**Decision Making Process**

From detection theory, an officer has a threshold and makes her decision by comparing the observation with the threshold. If the observation is greater than the threshold to search, she makes a positive decision; otherwise she makes a negative decision. We assume the existence of a single threshold which can be formulated by an if-then-else decision rule:

\[
\begin{align*}
    d &= 0 & \text{if } x < C \\
    d &= 1 & \text{if } x \geq C
\end{align*}
\quad (\text{Equation 1.1})
\]

where \(d\) represents a decision, and is 1 for positive decisions and zero for negative decisions. \(x\) is the observation, and \(C\) is the officer’s threshold to search. We show the true state of the world by \(Q\) which will be either 1, for positive events (guilty persons), or zero, for negative events (innocents).

Based on Table 1.2, we have four possible decision outcomes: true negative, false negative, true positive, and false positive. Decision payoffs (i.e., the value/price that one
receives/pays for each decision outcome), are set to be consistent with the experimental condition (payoff for true decisions =100 and payoff for wrong decisions = -50). The model is not restricted to these values, and other value structures will be explored in the future if the model proves useful.

Feedback

Under the full feedback condition, the formulation is straightforward: whatever the decision is, the model gets to know if the true status of the airport passenger (Q), and therefore gets informed if the decision was correct or not. However, the conditional feedback condition is different.

The important issue in modeling conditional feedback is about how people judge (code) the results of negative decisions for which they do not receive feedback. In the absence of feedback, would an officer judge that all of her negative decisions allowing passengers to pass were correct? Constructivist coding is proposed as a hypothesis about how people code their no-feedback decisions. It hypothesizes that decision makers assume all of their negative decisions were correct (Elwin et al., 2007; Henriksson et al., 2010). Ghaffarzadegan and Stewart (2011) argued that constructivist coding (at least in its original form) is not a good representation for high base rate conditions.

Here, we develop a more generic form of constructivist coding. We define $p$, to represent subjects’ confidence on their negative decision, the decisions for which they do not receive feedback. In other words, $p$ represents the probability that a subject assumes her negative decision was correct. So, if $p$ is 1, the model assumes there is no false
negative decision and when is equal to 0 the model assumes all of its negative decisions were wrong. Mathematically, equation 1.2 represents this formulation:

$$\bar{Q} = Q$$ \hspace{1cm} \text{if } d = 1 \hspace{1cm} \text{(Equation 1.2.a)}$$

$$p(\bar{Q} = 0) = 1 - p(\bar{Q} = 1) = p.$$ \hspace{1cm} \text{if } d = 0 \hspace{1cm} \text{(Equation 1.2.b)}$$

where $\bar{Q}$ is an officer’s perception of decision outcome. Equations 1.2.a and 1.2.b read as: if a police officer makes a positive decision (Equation 1.2.a), she gets to know the real outcome ($\bar{Q} = Q$), and if she makes a negative decision (Equation 1.2.b), with the probability of $p$ she guesses the decision was correct and the passenger was not carrying an illegal drug ($p(\bar{Q} = 0) = p$).

**Learning and adjusting threshold**

We define payoff shortfall as the difference between the maximum possible payoff and the actual payoff obtained for the decision made. We can formulate the process as follows:

$$\Delta V = V_{\max,Q} - V_{Q,d}$$ \hspace{1cm} \text{(Equation 1.3)}$$

where $\Delta V$ is payoff shortfall, and $V_{\max,Q}$ is the maximum possible payoff, the payoff that one would have received if made a correct decision (here, +100). Knowing that an officer has made a wrong decision ($\Delta V > 0$), the model assumes that the decision threshold will be amended toward the observation. The size of the change depends on several factors. Considering such a process, we can say:

$$\Delta C = k(x - C) \hspace{1cm} \text{if } \Delta V > 0$$ \hspace{1cm} \text{(Equation 1.4)}$$
where \( k \) is the correction ratio (the speed of adjusting threshold to an observed information cue or judgment) and we have \( 0 < k \leq 1. \) The correction ratio can depend on many factors, such as the personal characteristics of the decision maker, her confidence, and the amount of experience. Larger \( k \) represents people who change their threshold more drastically in response to outcome feedback. Our simple model assumes fixed \( k \). In the real world condition, \( k \) might change as one become more confident about her decision style.

In order to represent passengers passing an officer, we produce a set of random negative and positive events representing innocent and guilty passengers respectively. Consistent with detection theory negative events are normally distributed with the mean of zero and the standard deviation of 1, and the positive events are normally distributed with the mean of \( d' \) and the standard deviation of 1. We choose randomly from these distributions with a ratio that creates the desired base rate.

The stated four equations are enough to re-build the model. We provide more details about coding a program that facilitates reproducing the model with all stochastic components in the appendix.

Next we examine the model’s capabilities in replicating the data.

1.5. Data Replication

We examine the model’s ability to replicate the laboratory experiment results for each individual subject. For each subject, the model predicts each decision after receiving the 

\[ \text{This assumption can be considered close to the formal assumption of individual learning in error correction models (Kac 1962, Kac 1969) however in this model threshold adjustment is assumed proportional to the magnitude of the gap.} \]
weighted average of the information cues that the subject sees (weighted as proposed in
the experiment instruction). For full feedback condition the model estimates \( k \) for each
individual and the initial value of her threshold. For conditional feedback, the model
estimates \( p \) is addition to those parameters.\(^3\) We conduct this process for 60 subjects
(randomly chosen from our sample) in the described six conditions. As stated each
subject had 500 decisions (data points per subject).\(^4\)

We are interested to see the percentage of trials that the model’s output is exactly the
same as a subject’s decisions. Those are conditions in which the model and the subject
both make a yes decision (search) or both make a no decision (no search). Table 3
displays the results for all of the six conditions. For example, the small table at top left
reports the results for full feedback condition under the base rate of 0.1. As we see, in 427
times both the model and the subject made a positive decision (to search), and in 127
trials the model’s outputs were positive (to search) while subjects made negative
decisions (no search).

\(^3\) In simulation practices, estimation of parameters is referred as model calibration and it is conducted to get
the best fit between the model and the data. When the dependent variable is continuous, parameters are
estimated in a way that results in the least mean square error, in many aspects similar to a regression
analysis and coefficient estimation. The difference is that an endogenous simulation model creates
dependent variables in each time interval, and then uses the created value as its independent variable in the
next trial to predict the next value for the dependent variable. When the dependent variable is a binary
variable, as it is in our case, parameters are estimated to get the highest percentage of correct decisions.

\(^4\) More subjects could be tested, but we were satisfied after conducting the test for 60 subjects and observed
low marginal return with doing more calibration.
<table>
<thead>
<tr>
<th>BR = 0.1</th>
<th>Model</th>
<th>positive</th>
<th>negative</th>
<th>Model</th>
<th>positive</th>
<th>negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full Feedback</td>
<td>subjects</td>
<td>positive</td>
<td>negative</td>
<td>427</td>
<td>127</td>
<td>206</td>
</tr>
<tr>
<td>Conditional Feedback</td>
<td>subjects</td>
<td>positive</td>
<td>negative</td>
<td>449</td>
<td>63</td>
<td>161</td>
</tr>
<tr>
<td>BR = 0.5</td>
<td>Model</td>
<td>positive</td>
<td>negative</td>
<td>2275</td>
<td>290</td>
<td>371</td>
</tr>
<tr>
<td></td>
<td>subjects</td>
<td>positive</td>
<td>negative</td>
<td>2288</td>
<td>274</td>
<td>296</td>
</tr>
<tr>
<td>BR = 0.8</td>
<td>Model</td>
<td>positive</td>
<td>negative</td>
<td>3886</td>
<td>177</td>
<td>220</td>
</tr>
<tr>
<td></td>
<td>subjects</td>
<td>positive</td>
<td>negative</td>
<td>3040</td>
<td>256</td>
<td>216</td>
</tr>
</tbody>
</table>

Table 1.3: Results of replication of data for different base rates (BR) and different feedback types.

Note: Each table compares the model’s outputs and subjects’ decisions. For example, the table at top left shows under full feedback and base rate of 0.1, at 427 trials the model and the subjects made a positive decision, while in 127 trials the model made a positive decision for the trials that subjects made a negative decision.
Table 1.4 reports the percentage of correct replication of the data in each condition. Table 1.4 also gives a baseline to compare our model’s performance. It can be done by comparing our result (the third column) with the results from a static model of threshold placement (the fourth column). In the static model, threshold is fixed at the center. The model does a far better job in replication of the data in comparison to the static model in five conditions (conditions 1, 2, 4, 5, and 6). It has almost an equal performance in the full feedback condition for base rate of 0.5 where the optimal threshold is close to the middle (0.5).

<table>
<thead>
<tr>
<th>Conditions</th>
<th>Number of subjects</th>
<th>Total Number of trials</th>
<th>The model performance (the ratio of correct replications)</th>
<th>Comparison baseline (the ratio of correct replications by a static model)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Full feedback, BR=0.1</td>
<td>10</td>
<td>5000</td>
<td>93%</td>
<td>53%</td>
</tr>
<tr>
<td>2. Conditional feedback, BR=0.1</td>
<td>10</td>
<td>5000</td>
<td>95%</td>
<td>52%</td>
</tr>
<tr>
<td>3. Full feedback, BR=0.5</td>
<td>10</td>
<td>5000</td>
<td>87%</td>
<td>85%</td>
</tr>
<tr>
<td>4. Conditional feedback, BR=0.5</td>
<td>10</td>
<td>5000</td>
<td>89%</td>
<td>84%</td>
</tr>
<tr>
<td>5. Full feedback, BR=0.8</td>
<td>10</td>
<td>5000</td>
<td>92%</td>
<td>75%</td>
</tr>
<tr>
<td>6. Conditional feedback, BR=0.8</td>
<td>10</td>
<td>5000</td>
<td>91%</td>
<td>84%</td>
</tr>
</tbody>
</table>

Table 1.4 – Summary of replication of data for different feedback types and base rates (BR) for 10 randomly chosen subjects in each condition (6 conditions)

Note: The third column (the ratio of correct replications) shows the percentage of trials in which the model exactly replicated individual decisions. The final column
(comparison baseline) shows the percentage of trials in which a “static” model exactly replicated individual decisions, used as a base line to compare our model’s performance. In the static model of threshold placement threshold is fixed at the center.

Figure 1.5 shows the average results from simulation. Comparison of Figures 1.3 and 1.5 shows that the model captures average individual’s selection rate. As we see the model replicates the results of the experiments and predicts lower selection rates in conditional feedback.

Figure 1.5: Selection rate in the simulation model for last 200 trials for 6 different conditions: three base rates (0.1, 0.5 and 0.8) x two feedback types (full and conditional feedback).
As discussed, the model is able to replicate how people make screening decisions in a laboratory context. Using the calibrated model, we can conduct counterfactual analyses to investigate management tools that improve decision making for security officers.

1.6. Simulation Experiments

After model validation and gaining confidence on model’s ability to replicate human behavior for screening task, we can conduct simulation experiments with the model. This type of analysis is commonly known as counterfactual or what-if analysis. We report results from two sets of simulation experiments with the model. The main purpose of this section is to thoroughly investigate effects of confidence (p) and the level of expertise (d’) on learning and decision outcome. A summary of these simulation experiments are reported in Table 1.5.

<table>
<thead>
<tr>
<th>Experiments Goal</th>
<th>Simulation Experiment 1</th>
<th>Simulation Experiment 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Independent variables</td>
<td>p = 0, 0.1, .., 1.</td>
<td>p = 0, 0.05, ..., 1.</td>
</tr>
<tr>
<td>d’ = 0.5, 1, 1.5, and 2</td>
<td>base rate=0.5</td>
<td>base rate=0.5</td>
</tr>
<tr>
<td>feedback type=conditional</td>
<td>feedback type=conditional</td>
<td></td>
</tr>
<tr>
<td>Controlled variables</td>
<td>Threshold</td>
<td>percent of correct decisions</td>
</tr>
<tr>
<td>Sample size</td>
<td>1,100 (100 simulations per condition)</td>
<td>8,400 (100 simulations per condition)</td>
</tr>
<tr>
<td>Result</td>
<td>Threshold dynamics is influenced by the value of p, i.e, the level of confidence.</td>
<td>In different values of d’, there is an optimal value for p. As d’ increases total of correct decisions increase.</td>
</tr>
</tbody>
</table>

Table 1.5: A summary of two sets of simulation experiments
First, we investigate the effects of confidence on threshold learning dynamics. We run the model for different values of p under conditional feedback (1,100 simulation runs for the base rate of 0.5, $d'$ of 1, and for $p \in \{0, 0.1, 0.2, \ldots, 1\}$, that is 100 runs for each $p$).

Figure 1.6 illustrates the dynamics of threshold for different values of $p$. The x-axis is the trials. Each graph depicts the results for one specific $p$, representing an average subject’s performance when her confidence on her negative decisions is $p$. For example, the model predicts that for a subject who assumes 0.8 of her negative decisions are correct ($p=0.8$), if she starts from the initial threshold of 0.5, her threshold will oscillates around 0.5 for the whole experiment.

![Figure 1.6: Possible threshold dynamics for different $P$s (base rate = 0.5, $k=0.2$)](image)

As we see in Figure 1.6, the results are sensitive to $p$, which means the way that people interpret their past negative decisions can substantially influence their next
decisions. The figure also shows the optimal threshold, which for the base rate of 0.5 and $d' = 1$ is 0.5. At two extremes ($p = 1$ and $p = 0$), people who believe their negative decisions were always right or wrong end up with a considerable systematic bias (see graphs that are labeled as $p = 0$ and $p = 1$).

There are two main reasons for why there is a possibility for officers to code absent feedback differently and have different values of $p$. First, different people have different personality traits; some are more risk averse, presumably, coding more false for their negative decisions. Second, people’s confidence can change and influence their coding, and it can result in different averages of $p$ for different people in different conditions.

Following the previous simulations, we investigate the interactive effects of confidence and level of expertise ($d'$) on decision outcomes. We conduct a set of simulations for different values of $p$ and $d'$ under conditional feedback and compare the total percentage of correct decisions in each condition (total of 8400 simulation runs for 21 values of $p$ and 4 values of $d'$, that is 100 simulations for each condition). This helps to more precisely examine the effect of $p$ (confidence on negative decisions) on total performance under conditional feedback.

Figure 1.7 shows percentage of correct decisions for different values of $p$ and $d'$. Each graph in Figure 1.7 shows how, in a constant value of $d'$ (level of expertise), a change in $p$ (x-axis) affects the percentage of correct decisions in the last 200 trials (y-axis). For example, for $d' = 1.5$, if a decision maker is totally confident about her negative decisions ($p = 1$) she will end up with less than 70% correct decisions. For the same person, the graph shows less confidence, around $p = 0.85$, would have maximized her performance.
Figure 1.7: Percentage of correct decisions in last 200 trials for different values of $p$ (confidence on negative decisions) and $d'$ (level of expertise).

Note: $p^*$ is the optimal $p$, that is the value for $p$ that maximizes the percentage of correct decisions. For example, $p^*$ for a person with the level of expertise of $d' = 1.5$ is $p^* = 0.85$.

There are two major points in this Figure. First, if we look at each of the four graphs in Figure 1.7, there is a point where the percentage of correct decisions is maximum. In other words, for any value of $d'$, there is an optimal value of confidence, $p^*$, that results in the best performance. For example, $p^*$ for a person with the level of expertise of $d' = 1.5$ is $p^* = 0.85$. Larger or smaller values of confidence affect decision outcomes. Second if we compare the graphs, we see as the level of expertise ($d'$) increases and
people weight information cues more correctly the percentage of correct decisions in conditional feedback improves.

Overall, these simulation experiments stress the importance of helping decision makers to set their confidence to the optimal value avoiding over/under confidence. We will discuss the problem of overconfidence and the management tools to mitigate it in more details in the discussions section.

1.7. Discussion

This study was a behavioral decision-making approach to one of the most common decision problems in public service: detection tasks. The study focused on detection decisions in security as a common and important policy area. We argued that the nature of detection tasks consists of many behavioral decision-making components, and therefore policy design and implementation can benefit from taking those behavioral components into account. Then, we proposed a mathematical model of airport screening. The model was used to analyze the results of a laboratory experiment of airport security screening. To a considerable degree, the model replicated the human behavior in the laboratory environment. In following we discuss the role of overconfidence in screening performance, management tools that can help to solve this problem, and future avenues of research.

Overconfidence as an important source of the problem in conditional feedback

There are two major findings in this paper. First, a screener’s confidence on her negative decisions plays a critical role in her performance. As in conditional feedback,
people do not receive direct information about their performance when they make negative decisions, their own judgment of their performance becomes critical. If they overestimate their performance, they will make fewer positive decisions than optimal, and if they underestimate their performance, they will make more positive decisions than optimal (Figure 1.6). Overall, the analysis showed that one’s performance depends on her confidence, and there is an optimal level of confidence which results in best decision making performance (Figure 1.7).

Second, our empirical investigation with the model shows that decision makers are on average overconfident on their negative decisions. What makes the situation more complicated is that an increase in confidence, results in fewer positive decisions, which in turn results in less feedback, as feedback is conditional on positive decisions. Thus, it is critical to help security officers to adjust their confidence with their performance in order to improve their decision making performance. The behavioral decision making literature can help in this respect.

In fact, overconfidence has been one of the research areas in behavioral decision making since sixties and seventies when different studies found that people systematically think that they know more than what they actually know (Adams & Adams 1961; Oskamp, 1965; Lichtenstein & Fischhoff, 1977). Since then, a vast literature has been developed on understanding overconfidence. Effects of different variables on confidence have been examined including decision makers’ characteristics (e.g., personality, skill, and experience), task characteristics (e.g., task difficulty, number of information cues, task domain), and cultural characteristics (Soll, 1996; Klayman et
al., 1999; McKenzie et al., 2008; Fischer & Budescu, 2005, Tsai et al., 2008; Yates et al., 1997).

Most of the behavioral decision making studies are concerned with the question of how people judge their performance without knowing about it. Our study moves forward, and claims that people’s misjudgment of their performance hampers their performance under conditional feedback situations. In conditional feedback, asymmetries in feedback (having feedback only on positive decisions) result in misplacement of decision thresholds and sub-optimal decisions. Therefore, our argument is not solely about whether officers know about their performance or not, but we show that inaccurate beliefs about one’s performance can result in dynamically worsening performance. As a result, management tools that address overconfidence can actually improve the performance of security officers.

**Management tools to reduce overconfidence**

Various strategies have been suggested to help people to overcome the problem of overconfidence. We review four of these strategies. First, asking for justification is one of the common techniques to mitigate overconfidence (Lichtenstein & Fischhoff, 1982). When people are asked to report their reasons for their decision their confidence declines. In fact in many settings, decision makers are required to justify their decisions. Further, it is shown that when individuals are told that they would be required to defend their decisions in front of a group, their confidence on their decisions has declined (Arkes et al., 1987).
Second, it is also observed that issuing cautions about the problem of overconfidence can help mitigate the problem. If decision makers know that people are prone to become overconfident, they may more easily accept the possibility of their own mistakes.

Third, decision makers can be trained to calibrate themselves and have a more accurate estimation of their performance (Lichtenstein & Fischhoff, 1980). This is usually done in training sessions by giving a report on decision makers’ deviation from their actual performance in different decisions. As decision makers usually are more confident in difficult tasks, these reports usually consist of a graph that shows how decision makers’ performance and confidence deviate in different levels of task difficulties.

Finally, feedback has been argued to help adjusting confidence to actual performance (Lichtenstein et al., 1982; Fischer, 1982). It is expected that once decision makers see what percentage of their past decisions were correct, or if they receive feedback after each decision that they make, they estimate their ability in making correct decisions more precisely. However, the empirical evidence on effects of feedback on confidence is mixed (Subbotin, 1996). Winman and Jusline (1993), Gonzalez-Vallejo and Bonham (2007) and Rose and Windschitl (2008) find support for corrective effect of feedback on overconfidence. They all use general knowledge questions in their studies. Fischer and Budescu’s (2005) study is the only one that compares overconfidence in full feedback and conditional feedback. The study shows that overconfidence in conditional feedback is more difficult to mitigate than in full feedback.
Following this thread of research we can expect that training sessions with a full feedback condition can help screeners to adjust their level of confidence. As biases emerge in the process of daily duties, periodic training with full feedback can help de-bias screeners. In addition, this training should include calibration information, and cautious about the problem of overconfidence. In these training sessions officers can be asked to justify their decisions based on a checklist that is developed for such a purpose. Training sessions can use sophisticated simulation software to represent different types of passengers. Further, in the field, grouping officers, suggesting random checks, and conducting periodic discussion sessions about officers’ performance can help to mitigate the problem of overconfidence. Finally relevant public organizations such as TSA should estimate the optimal threshold for searching and optimal selection rate (the percentage of people that should be screened). Deviation from these measures should be considered as a sign of threshold misplacement.

**Future Avenues**

There are several possible ways to extend this study and more contribute to the public policy literatures. First, personality traits are shown to influence learning in binary decision making and the level of confidence (Klayman et al., 1999). So, we can expect personality traits to affect how people interpret their negative decisions and learn in conditional feedback situations. In airport screening, therefore, some officers should make better screeners than others, and knowing the personality characteristics’ effects on screening performance can help in the hiring of more proper officers for screening. In further studies, data on subjects’ characteristics can be gathered. Different parameters can
then be compared. Testing a hypothesized relationship between some of the Big Five personality characteristics (like openness) and the way that people code negative decisions \((p)\) is another possible and interesting way to extend this study.

Second, the policy recommendations should be empirically tested. We suggest the policies to be tested in a laboratory or a (close to) field context. For example, effect of periodic training sessions that provide calibration or complete feedback to screeners, can be examined.

### 1.8. Conclusion

This study has theoretical and practical contributions at two different levels. First, we contribute to the understanding of decision making through extending the theoretical analysis of how people learn and how they shift their decision thresholds. While many scholars have emphasized the negative effects of misperception of delays (Rahmandad, 2008; Rahmandad et al., 2009), noise in feedback (Bereby-Meyer & Roth, 2006), and feedback asymmetry (Denrell & March, 2001) on learning, our work builds on few studies of conditional feedback and its effect on learning. This study does not reject other theories, but sheds more light from a new perspective on the problem of barriers to learning. The simulation outcomes and the replication of data show that conditional feedback can result in bias and underestimation of the base rate. Basically, assuming people learn from their incorrect decisions, in conditional feedback, most of the negative decisions are treated as correct ones. Therefore, there is always a force from false positive results to correct the threshold, while there is no accurate information about false negative results. Thus, for an overconfident person, who overestimates the results of her
negative decisions, threshold increases too much. This implies that conditionality of feedback for screeners such as security staff can result in misperception of performance and overconfidence.

Second, this study contributes to the theory and the practice of policy implementation. We extend theories of security to the micro level of behavior—i.e., how officers’ behavior results in policy shortfalls, the area that has been underestimated in many policy studies. We also suggest several practical applications and management implications in order to increase screening performance. The significance of this contribution resonates with the current international concern about investment in security.

Overall, our main claim in this paper is that conditional feedback is a barrier for improving security detection tasks, and that it results in a systematic bias and overconfidence. In addition, helping officers to correct their confidence in their negative decision will result in less bias in threshold placement and more accurate decisions. We propose management tools that can help to mitigate overconfidence problems in order to improve screening accuracy.

Acknowledgments

This material is based in part upon work supported by the National Science Foundation under Grant Number SES 0451850. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

I am thankful to Professors Thomas Stewart, David F. Andersen and Erika G. Martin for their helpful comments. The first drafts of the paper were presented in the Academy
of Management Conference in 2009, the Judgment and Decision Making Conference in 2009, and the System Dynamics Conference (SDC) in 2008. The first draft of the paper received the Barry Richmond Award for its contribution to systems thinking. We are thankful for all of the comments that we received through these conferences.
Reference


Appendix 1.1:

A Problem with the Constructivist Coding Hypothesis as a Learning Model for Selective Feedback When the Base Rate is High

In many natural settings, people receive outcome feedback only when they make positive decisions. For example, a police officer who stops suspicious drivers to test for the blood alcohol content only receives feedback when she decides to stop and conduct the test (positive decisions). For the cases that she does not stop a car, she will never know if her decision to let the car go (negative decision) was correct or not. In this example and many other real world conditions, outcome feedback is selective and is usually limited to positive decisions. Several studies suggest that under the selective feedback condition, people make fewer positive decisions than would be optimal (Elwin et al., 2007, Henriksson et al., 2010, Stewart et al., 2010).

Elwin et al. (2007) offer the constructivist coding hypothesis to describe how people code the outcomes for decisions when there is no feedback. Under this hypothesis, in the absence of feedback people assume their decisions are correct. The result is a bias in selection rate and too few positive decisions. Elwin et al. (2007) found empirical evidence for this hypothesis and Henriksson et al. (2010) provide additional support.

Both studies used tasks with a base rate of .5 – a common practice in studies of category learning. Studies restricted to base rates of .5 or lower are justified when feedback is given on every trial because a base rate above .5 (say, .7) can always be

---

5 During finalizing the dissertation, a newer draft of Appendix 1.1 got published as a commentary in the Journal of Experimental Psychology: Learning, Memory and Cognition. In the published manuscript we also offer a few suggestions to improve constructivist coding. For complete discussion please see Ghaffarzadegan, N., & Stewart, T. R. (2011). An extension of the constructivist coding hypothesis as a learning model for selective feedback when the base rate is high. forthcoming in the Journal of Experimental Psychology: Learning, Memory & Cognition.
reformulated as a base rate below .5 (in this case, .3) by simply switching events. But when feedback is selective, the resulting asymmetry prevents this reformulation, and base rates above .5 must be considered. Such high base rate conditions are common outside the laboratory. For example, in a public place, such as a night club or a bar, screeners will attempt to bar admittance to trouble makers, but most patrons will not cause trouble. The screeners receive no feedback about patrons not admitted.

We will show that for high base rates (.7 or higher) the constructivist coding hypothesis overestimates base rate bias.

**Constructivist Coding in a Learning Process**

The foundation of the constructivist coding hypothesis is that people attempt to match their selection rate with the base rate that they infer from the combination of a) feedback and b) their assumption that they are always right when there is no feedback.

This inferred base rate will be:

\[ b = p \cdot s \]

(Elwin et al., 2007, equation 2)

where \( s \) is the selection rate in the learning phase, and \( p \) is the ratio of correct decisions to all positive decisions. For an experiment with a learning phase followed by a testing phase, \( b \) is the inferred base rate after the learning phase, and therefore also the selection rate for the test phase. To represent series of blocks of learning trials and the dynamics of learning after each block, we use block index for the variables (e.g., \( b_n \) to represents inferred base rate after block \( n \).)
Constructivist coding predicts that the selection rate will not increase, and will probably decrease after each block of trials. Therefore, subjects will almost certainly end up with lower selection rates than they start with. The inferred base rate at the end of block \(n\) \((b_n)\) is determined by the selection rate in the previous trials \((s_n)\) and the proportion of correct decisions in the last trials \((p_n)\). Since \(p_n\) cannot be greater than 1.0, a simple analysis shows that the selection rate should decrease with each block\(^6\). If (as is generally the case) people start from a selection rate around .5, they should end with a lower selection rate even if the base rate is very high.

There is a possibility that the learned selection rate could increase if there are several true positive outcomes by chance, and the selection rate is updated after every trial. In order to address this possibility, we examined a simulation model representing trial by trial updating. An algorithm for this simulation is provided in Henriksson et al.

\(^6\) If we assume people learn in large enough blocks of trials, we can assume stochastic components to cancel out. Based on this assumption, subjects exactly select \(b_n\) portion of the cases in the block \(n+1\). Therefore, for the blocks \(n\) and \(n+1\), we can write equations 1.A.1 and 1.A.2:

\[
b_n = s_n p_n \quad (1.A.1) \\
\frac{s_{n+1}}{n+1} = \frac{ns_n + b_n}{n+1} \quad (1.A.2)
\]

Equations 1.A.1 and 1.A.2 result in equations 1.A.3 to1. A.4:

\[
s_{n+1} = s_1 \prod_{i=1}^{n} \frac{i + p_i}{i + 1} \quad (1.A.3) \\
b_{n+1} = s_1 \left( \prod_{i=1}^{n} \frac{i + p_i}{i + 1} \right) p_{n+1} \quad (1.A.4)
\]

Note that \(\left( \prod_{i=1}^{n} \frac{i + p_i}{i + 1} \right) p_{n+1} \leq 1\). Equation 1.A.4 shows that the final inferred base rate \((b_{n+1})\) is sensitive to the initial selection ratio \((s_1)\) and \(b_{n+1} \leq s_1\). In other words, subjects who usually start an experiment with the selection rate of 0.5 cannot finish with an inferred base rate of more than 0.5. Stewart et al.'s (2010) empirical investigation shows that although base-rate bias exists in the base rate of 0.8, but the selection rate of subject in their last block of experiment is significantly higher than 0.5. The constructivist encoding hypothesis cannot account for this.
We use the algorithm and apply it for a single cue (stimulus) task in a simulation model\(^7\).

The simulation was run 1000 times for each of 4 different base rates (.5, .7, .8, and .9). Simulated subjects started learning with a selection rate of .5. Each simulation run included 240 learning trials and 100 no-feedback testing trials. Figure 1.A.1 shows the 99% confidence interval for the final inferred base rates. The results, while replicating Elwin et al.’s (2007) and Henriksson et al.’s (2010) results for the base rate of .5, show that for higher base rates constructivist coding predicts inferred base rates lower than .5, which implies a very large bias. Of the studies of selective feedback that have included base rates above .5 (Stewart et al., 2010; Griffiths & Newell, 2009; Fischer & Budescu, 2005), only Stewart et al. included selection rates in their results. Mean selection rate for their subjects when the base rate was .8 was .68, a result virtually impossible to obtain with constructivist encoding (p < 0.001).

---

\(^7\) We use the algorithm offered in Henriksson et al. (2010, p.15-16) and apply it for a single cue (stimulus) task. For each trial, the model compares the similarity of the stimulus with the previous experiences. Equation 1.A.5 represents the distance between the current trial’s stimulus \((y)\) and the stimulus \((x_k)\). The distance function is used in equation 1.A.6 to formulate the similarity between stimulus \((y)\) and the stimulus \((x_k)\), equivalent to equation A3 in Henriksson et al. (2010 p.15).

\[
d_{y,k} = |y - x_k| \quad (1.\text{A.5})
\]

\[
Sim(y, x_k) = e^{-d_{y,k}} \quad (1.\text{A.6})
\]

Based on theses equations, our model calculates similarity between the current stimulus and the past ones and uses the equation for probability of selection in Henriksson et al. 2010 (A1, p.15) to calculate the probability of making a positive decision in each trial. We also test the sensitivity of the results to the equation 1.A.6 by testing a few other functions, and find qualitatively consistent results.
Inferred base rate at the end of 240 trials

Figure 1.A.1: Results of a simulation model of constructivist coding for different base rates. Whiskers show 99% confidence intervals. The simulation model for the base rates of .5, .7, .8, and .9 predicts the inferred base rate to be 0.31, 0.38, 0.43, and 0.46 respectively (i.e., the bias is 0.19, 0.32, 0.37, and 0.44 respectively). Thus, for the high base rates, constructivist coding predicts a large base rate bias.

**Conclusion**

The constructivist coding hypothesis imposes an overall decreasing trend in selection rate in a learning process. Therefore, this cannot represent learning in high base rates such as .8, where people learn to increase their selection rate. Studies have shown that the hypothesis has merit in explaining behavior in tasks with .5 base rate, but it will not explain behavior when the base rate is high.
References for Appendix 1.1


Appendix 1.2: Detailed coding of the model to replicate the results

Note: The following coding is in Vensim language. Adapting the codes to other software packages is straightforward and needs only translating the functions of “IF THEN ELSE”, “RANDOM NORMAL”, “RANDOM UNIFORM” and “INTEG” (integral) into the new language.

Generator of Detection Distributions (Figure 1.1)

Xsignal= RANDOM NORMAL(-10, 10 , d prime, 1 , random seed )

Xnoise= RANDOM NORMAL(-10, 10 , 0, 1 , random seed )

true status of a passenger (q)= IF THEN ELSE( RANDOM UNIFORM(1, 100 , random seed)>(100*(1-base rate)), 1,0)

observation (x)= IF THEN ELSE(true status of a passenger (q)=1, Xsignal , Xnoise )

Decision Making Process (equivalent to equation 1.1)

positive decision (d)= IF THEN ELSE(observation (x)>screener's threshold (C), 1 , 0 )

Feedback (equivalent to equation 1.2)

RND= RANDOM UNIFORM(0, 100 , random seed)

feedback availability= (1-switch to conditional feedback)+switch to conditional feedback*positive decision (d)
screener's understanding of outcome = feedback availability * true status of a passenger

(q)+(1-feedback availability)*IF THEN ELSE (RND>100*confidence in negative decisions (p), 1,0 )

**Learning and adjusting threshold (equivalent to equations 1.3 and 1.4)**

screener's payoff estimation (V')= IF THEN ELSE (screener's understanding of outcome=0, IF THEN ELSE ( positive decision (d)=0, Vtn, Vfp), IF THEN ELSE ( positive decision (d)=0, Vfn, Vtp))

payoff shortfall (delta V) = maximum payoff (Vmax,q) - screener's payoff estimation (V')

screener's threshold (C)= INTEG (change in threshold (delta C),Initial threshold)

change in threshold (delta C) = IF THEN ELSE(payoff shortfall (delta V)>0, 

(observation (x) - screener's threshold (C))*k,0)

maximum payoff (Vmax,q)= Vtn+ screener's understanding of outcome*(Vtp-Vtn)

**Parameters (different values have been/can be tested in the paper)**

base rate= 0.5
d prime= 1
Initial threshold= 0.5
k=0.2
confidence in negative decisions (p)= 0.95
switch to conditional feedback= 1
Vtn= 100, Vfp= -50, Vtp= 100, Vfn= -50
TIME STEP = 1
random seed= 1
ESSAY 2:

An Examination of Complexities of Warning Issuance for Domestic Security in a Long Term Dynamic Context
Abstract:

Our daily life is full of various forms of security warnings. But do they have enough effect on people’s behavior, and how can the government improve its domestic warning system? The purpose of this paper is to examine the tradeoffs inherent in issuing security warnings and to produce a set of policies that identify the optimal timing for warning issuance. We develop a simulation model of security based warning, and conduct several simulation analyses of the model. Through the model we show that optimal solutions are sensitive to two major behavioral characteristics: people’s sensitivity to false alarms and adversaries’ perceptions of public sensitivity to false alarms. The results show that an underestimation of these effects can result in sub-optimal solutions and performance shortfall.

Keyword: security, public warning, private warning, normative decision making
2.1. Introduction

Our daily lives are full of various security warnings. Security notices in airports, subways, train stations, newspapers and TV programs are a few examples of frequently heard and observed warnings. In many public places such as banks, museums, parks, government related organizations, and universities, warning signs are frequently used to notify people about possible threats of thefts and attacks. New technologies have also helped warning issuance, and administrators can now send more frequent warnings through cell phones and emails. But do those warnings change people’s behavior? How can an organization improve its warning system’s effectiveness?

The problem of warning issuance is a behavioral problem: warnings are aimed to influence the behavior of different groups of people including employees, customers, citizens, and security officers. Effectiveness of warnings depends on whether people pay attention to them and warnings alter their behavior accordingly. If the public responds appropriately to warnings, threats are mitigated and security improves. Public’s preventive actions mitigate the level of threat and improve security. For example, in the context of homeland security, the effectiveness of warning issuance is influenced by how citizens and governmental employees such as security staffs and police officers behave and react to those warnings. Many studies have stressed the importance and complexities of effects of human cognition and behavior on security systems performance (Kivimaki and Kalimo 1993, Bazerman and Watkins 2005, Mayntz 2006, Greve 2005).

This paper addresses the public’s complex behaviors which the administrators need to take into account in warning issuance. We build a generic, dynamic model of a security system to examine warning decision making and the characteristics of optimal strategies for issuing the most effective warnings under different assumptions.
Developing a mathematical model and simulation experiments help us to examine a wide range of contexts and the effects of different assumptions that one might consider in policymaking. Our focus is on a hypothetical public place which is under the threat of possible attacks. We conduct different simulation runs under a wide range of scenarios and compare the findings. Through the analyses, we argue that optimal solutions in a long term period are sensitive to two major behavioral characteristics: the level of public elasticity to false alarms and adversaries’ reactions to the public’s sensitivity. The results show that an underestimation of these effects or a myopic optimization of warning decisions can result in biases in optimal solutions. The paper proceeds as following: First, we review the relevant literature and discuss the complex nature of warning issuance (section 2.2). Then, we introduce our simulation model in different steps and conduct several base run simulations (section 2.3). Then, we conduct optimizations under different scenarios and compare the results (section 2.4). Finally, we make concluding remarks and discuss the implications of this study for policymakers (section 2.5).

2.2. The Complex Nature of Warning Issuance

Warning as a Decision Making Problem

Studies of security in high risk and public facilities such as airports, harbors, and electric facilities reveal huge costs associated with attacks such as loss of use of critical urban infrastructures and facilities and human deaths (Gordon et al. 2007, Rosoff and von Winterfeldt 2007, Rose et al. 2007, Simonoff et al. 2007). In addition, resource limitations impose the need for smart allocation of resources to protect the most risky and costly facilities (Bier 2007, Keeney 2007, Willis 2007).
On the other hand, several studies on security policies show that this domain of policymaking is prone to decisions that may seem proper in the short run but backfire in the long term (Sagan 2004, Bunn 2004, Ghaffarzadegan 2008, Hovmand et al. 2009, Anderson 2011). Security agencies respond to continuously changing and complicated events with few or no precedents (Greve 2005). Lack of accurate information, inadequate analysis of information, and unwillingness of many agencies to share information with other agencies are properties of working in such conditions (Sutcliffe 2005, Greve 2005, Bardach 2005). The existence of various and conflicting social pressures (Davis and Silver 2004), long delays between policy implementation and results, and adversaries’ responses to the policies implemented lead to frequent shifts in security policymaking in airport screening (Weaver and Richardson 2006).

Issuing private and public warnings are two of the most common policies to increase security in public places (Keohane and Zeckhauser 2003). Private warnings are warnings that are issued for the security staff, officers, and guards. Public warnings are issued in a higher level of threat and target the whole community (including citizens and security staffs) and are communicated through the media. Warning issuance can result in more defensive reactions among the public and security staff in addition to possible deterrent effects and stopping adversaries (e.g., criminals, terrorists) from attacking.

From a behavioral decision making perspective, there are three major alternative decisions for warnings: no warning, private warning, and public warning, and security administrators take one of these actions based on the level of perceived threats. On the other hand, if for simplicity we assume that there are two possible state of the world

---

8 It is usually assumed that guards are aware of public warnings, therefore public warning issuance is equivalent to issuing public and private warnings together (Pinker 2007).
about adversaries’ plan to attack (either there is a plan to attack or there is not), six different decision outcomes emerge. Table 2.1 shows these outcomes.

<table>
<thead>
<tr>
<th>Adversaries' plan</th>
<th>No plan to attack</th>
<th>Plan to attack</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>False negatives</strong></td>
<td>- No unnecessary warning is issued.</td>
<td>- Guards and Public are not informed. - Attacks are implemented.</td>
</tr>
<tr>
<td><strong>True negatives</strong></td>
<td>- No unnecessary warning is issued.</td>
<td>- Mixed outcomes - Guards are informed. - Public is not informed. - Deterrent effect may occur due to guards’ reactions.</td>
</tr>
<tr>
<td><strong>True positives</strong></td>
<td>- Mixed outcomes - Guards are falsely alarmed. - May result in loss of sensitivity among security staff to future warnings</td>
<td>- False positives - Guards and Public are falsely alarmed. - May result in loss of sensitivity among the public and security staff to future warnings</td>
</tr>
</tbody>
</table>

Table 2.1: Six possible outcomes for warning decisions

Under the scenario that adversaries have a plan to attack, no warning results in the worst possible decision outcome (false negatives). Guards and public will not be ready to properly react and possibly attacks will result in considerable damages. If only a private warning is issued guards will be informed but the public will not be informed. Guards’ reactions might result in deterrent effects. The best condition under the scenario of possible attack is when public warnings are issued and both security staff and public are ready to react.

On the other hand, under the scenario that adversaries have no plan to attack, warning issuance is a false decision and unnecessary. In addition to social stress caused by unnecessary warning issuance, a false warning (issuance of a warning when no attack
happens) may have significant indirect effects on the society. While people are more sensitive when they have recently experienced attacks, they may lose their sensitivity as they receive more false alarms (Roulston and Smith 2004). The effectiveness of warnings is in fact not only a function of the accuracy of the warning, but also people’s responsiveness (Pate-Cornell 1986, Roulston and Smith 2004), whereby a decline in the latter effect can be considered as an indirect cost which can result in more damage in the future.

Furthermore, the accumulation of false alarms and the information feedback to adversaries about a potentially desensitized public may add to the long term costs of a warning policy. These examples reveal the complexity of making proper warning decisions. In the presence of these complexities, finding optimal time to issue security warning is not straightforward. In this study our model will represent all of the six conditions illustrated in Table 2.1.

**Optimization of Warning Decisions**

Considering a wide range of possible outcomes from a warning decision and the environmental uncertainties that makes error inevitable, it is important to finds ways to develop a warning optimization model and take the best warning decisions. Several studies have been conducted to define and examine optimization of warning decisions in a short time period.

In one of the first warning optimization studies, Pate-Cornell (1986) examines the tradeoff between false alarm issuance and warning effectiveness and develops a method to find the optimal threshold placement. In her model, when the warning threshold is placed low enough, the alarm system activates sooner, so people have enough time to
evacuate, but the system is more sensitive to noises in the atmosphere such as small changes in the pressure or the temperature. Her study is unique as the Crying wolf effect, the deterioration of sensitivity to alarms as results of past false alarms, is taken into account.

Pinker (2007) proposes a framework for optimization of warning decisions. He studies the effect of short-term responses to security policies and proposes a framework to estimate the optimal level of guard allocation and private and public issuance of warnings. Pinker (2007) offers mathematical procedures that lead to the minimization of total costs in short term.

Considering the wide range of studies on warning issuance, the literature lacks a dynamic model of security warning in public places such as subways, airports, and universities. In contrast to weather warnings or fire alarms, the source of a threat in the homeland security system (e.g., adversaries, terrorists) can react to any warnings (Sagan 2000) which makes the problem more complicated. Public warnings can result in the deterrent effect (cancelation of an attack), and people may perceive the results as false alarms. Specifically, an extension of Pinker’s (2007) study to consider long term dynamic effects of warnings in order to minimize long-term costs can be interesting and insightful. Especially, as we mentioned many of security policies have long term effects, thus myopic optimization does not necessarily prove effective in the long term.

Empirical examination of effectiveness of warnings in particular, and security policies in general, is very difficult due to the small frequency of related events such as attacks, and therefore lack of longitudinal data with enough variation. Due to the high public consequences of security policies, experimentation in the field is very risky and
rare. These characteristics of the problem leave an opportunity for a simulation approach and a counterfactual analysis of warning policies (Davis et al. 2007, Ghaffarzadegan, Lyneis and Richardson 2010).

This paper aims to contribute to the literature of decision and policy sciences by developing a dynamic model of warning-based security. In general, a dynamic model helps us to examine the long-term side effects of a policy and help policymakers to avoid better-before-worse patterns of behavior: the conditions under which a system is better off in short term but backlashes in long term to a worse situation (Forrester 1971). In the context of security, a dynamic model of security will help to avoid actions that may decrease short term vulnerabilities in a much more considerable long term costs.

2.3. A Formal Dynamic Model of Warning Based Security

Security is the outcome of interactions of many different players in the system. Manunta (1999) defines security as a function of protectors, assets, and the level of threat. Protectors can be both decision makers and guards who aim at ensuring an adequate level of security. It can also refer to the public, if they are aware and react properly. Assets can be financial or can be the public being protected from adversaries. The level of threat means the capacity of the adversaries to attack or commit crimes and their interest in doing so. Following this logic, a model of security should consider defenders, the public and adversaries, and the interactions among the three factors.

Figure 2.1 shows our three-sector conceptual model that we will formulate in this section. As we see, in the conceptual model, there are three main sectors. First, at top, there is the adversaries sector, where plans to attacks are made and implemented resulting
in damages. Second, at bottom – left, there is the defense decision making sector. Information about the level of threat comes as an input to this sector from the adversaries sector. Usually this information is received through an intelligence system. Information about the level of threat is not necessarily accurate and is prone to different communication errors and uncertainties. The outputs of this sector are warning decisions, which are public warning for the public and private warning to the security staff and guards. Warning decisions made in this sector enter to private and public sectors and cause defensive reactions. Warnings influence public and security staff through causing reactions. Warnings also affect adversaries through the deterrent effects, causing them to cancel or postpone their plan to attack. Finally, the sectors of public and private represent defensive reactions of citizens and guards as results of warnings. Their reaction of course depends on how sensitive they are to warnings. Such defensive reactions can decrease possible damages that might have occurred from an attack.

Figure 2.1: A Conceptual model of Public and Private Warnings in a Security Context
Formulating the continuous links shown in Figure 2.1 will give us a basic warning-based simulation model (section 3-1). Then we formulate the links shown in dash. In section 3-2, we argue that sensitivity to warnings can endogenously change in the system. As more damages occur and the words spread out, people become more frightened and defensive (the link from damage to sensitivity to warnings). Further, as they become overwhelmed with warnings and false alarms, they might lose their sensitivity to future warnings (the link from warnings to sensitivity to warnings). Finally in section 3-3 we examine the possibility that the adversaries observe the environment and receive information about how people behave and perceive threats. It can be hypothesized that adversaries might not always cancel their attack in response to warnings, but their reaction may depend on guards and citizens’ behavior (the link from sensitivity to warnings to deterrent effect). The simulation model will be formulated in a way that the elasticity of sensitivity to warnings and deterrent effects can be changed. This helps to examine and compare a wide range of conditions, such as the extreme condition of constant sensitivity (absence of dashed links), and other conditions in which these variables endogenously change in the system.

Below, we will discuss the model in more detail and formulate the model. Through building the model we examine several simulation runs.

A Numerical Model of Warning Decision Making

Since 2002, the United States has used a system of warning which is based on the Department of Homeland Security’s assessment of the likelihood of facing an attack. The Department of Homeland Security relies on information and analyses that they receive
from various intelligence agencies and decide the warning level that they want to issue (Pinker 2007). Such a judgment-based decision making process can be represented with the signal detection theory (Macmillan and Creelman 1991, Green and Swets 1966).

Signal detection theory divides events into two categories, the ones we want to detect (positive events) and others (negative events). In terms of security, we are interested in differentiating the times that we will be attacked (positive events) from the times that we will not be attacked (negative events). A perfect detection leads to proper warning decisions.

From the signal detection perspective, in order to detect positive events and issue warnings, a decision maker compares his or her perception of threat with some warning thresholds. Then the decision maker makes a decision about whether or not warnings should be issued. If the decision maker perceives the level of threat to be higher than a threshold, a warning will be issued.

Due to environmental uncertainties, it is not completely possible to distinguish positive events from negative events, so the distributions overlap. Figure 2.2 assumes a normal distribution for positive and negative events⁹ and shows how the distributions can overlap. So, wherever we position our decision threshold for public and private warnings there is always a chance that we will make a wrong decision.

---

⁹ While different distribution can be assumed for positive events and negative events, a normal distribution is consistent with the signal detection theory. Further, as the distributions are over judgment and not time we do not necessarily need to consider a Poisson distribution.
In Figure 2.2, the distance between the means of the positive and negative event distributions is labeled $d'$ and represents judgmental accuracy which may improve by improving the intelligence systems. Basically, we prefer to have as high $d'$ as possible, i.e. more accurate differentiation of positive events from negative events. The vertical lines are decision thresholds for private and public warnings. In this framework, the proportion of positive events to total events is called the base rate. For a particular case when judgment falls below both of the thresholds then no warning action will be taken. If it falls between the thresholds only private warnings will be issued, and if judgment falls above both of the thresholds, a public warning will be issued which is equivalent to issuing both public and private warnings. Mathematically we can say:

$$W_i = \begin{cases} 1 & \text{if } x > T_i \\ 0 & \text{if } x \leq T_i \end{cases}$$

Equation 2.1
whereby $i$ represents one of the two agents, i.e., $i \in \{private, public\}$. In this representation, for example if $x > T_{public}$, then $W_{public} = 1$, i.e., if the level of threat is judged to be higher than the threshold to issue public warning, a public warning will be issued.

As shown in Figure 2.1, damage is considered as the final result of the model, and we are interested in minimizing damage costs. It is reasonable to consider damage ($C$), as a cost component which is as a function of attack ($a$) and other entities’ reaction to issued warnings. Equation 2.2 displays such a relation.

$$C = f_i(a, R_{public}, R_{private})$$

Equation 2.2

Intuitively we expect if $a=0$ then $C=0$, that is, no attack results in no physical damage. Further $\frac{\partial f_1}{\partial R_{public}} \leq 0$ and $\frac{\partial f_1}{\partial R_{private}} \leq 0$, that is damage decreases as the public or guards react to warnings. Appendix 2.1 displays the detailed formulations for each functions and how reactions are defined.

Now, we can randomly create positive and negative events in the model consistent with the signal detection framework and examine effectiveness of issuing warnings as time passes. Figure 2.3.a shows a single simulation as an example of a base run of the model in a five-year time period. The model is simulated for five years under the base rate of 5 percent, i.e., in 5 percent of the weeks in five years there is a possibility of facing an attack. Later in this paper we will develop the model to consider change in different agents’ sensitivity to warnings which are illustrated in Figures 2.3.b and 2.3.c.
Figure 2.3: A set of base run simulations capturing a sample of model outputs for different developments in the model (sections 3-1, 3-2, and 3-3) in a five year time period.

Note: Darker lines show damage cost from an attack and lighter lines show maximum possible damage cost. For example in (a) we see one successful attack, four semi successful, one unsuccessful, and five canceled attacks.

(a) Dynamics of damages from adversaries attacks in a five year time period under the assumption of constant public and private sensitivity to warnings (section 3-1)
(b) Dynamics of damages from adversaries attacks in a five year time period under the assumption of inconstant social sensitivity to warnings (section 3-2)

(c): Dynamics of damages from adversaries attacks in a five year time period under the assumption of inconstant social sensitivity to warnings and inconstant reactions from adversaries (section 3-3)
Modeling social sensitivity to warnings

So far it is assumed that sensitivity to alarms is a constant, exogenous parameter. From now on, we consider the possibility of an endogenous change in sensitivity. Two main phenomena may result in change in sensitivity to warnings. The first phenomenon is rooted in the availability heuristic: individuals do not weight different information cues equally but in many instances they put more weights on the most recent and available pieces of information (Tversky and Kahneman 1974). This phenomenon can influence risk perception. In warning decision making and public reaction, it is plausible that recent attacks can make people more sensitive to the news about security and warnings. In other words, more recent damages from attacks lead to higher risk perception and more sensitivity to warnings, which in turn, everything else constant, results in more responsiveness to warnings and lower possibility of experiencing a successful attack.

The second effect is the “crying wolf” effect. When faced with many false alarms people may lose their sensitivity to alarms and not react appropriately. Weak reactions to alarms can in turn result in more successful attacks for adversaries and hence more damages. Equation 2.3 shows these arguments in a mathematical form.

$$ S_i = f_2(C_i, W_i, \alpha_i) $$

Equation 2.3

As Equation 2.3 represents, sensitivity ($S_i$) is considered as a function of recent damages ($C_i$), recent warnings ($W_i$), and elasticity to recent damage warning ratio ($\alpha_i$). Details of this formulation are presented in Appendix 2.2.

We can examine base run simulations for different values of elasticity and compare the dynamics of damages that the public organization faces. Figure 2.3.b shows a base
Modeling adversaries’ reaction to social sensitivity

It may be argued that it is simplistic to assume that the deterrent effect always exists even if the public loses its sensitivity to warnings due to extensive false alarms. In fact, terrorists can update their perceptions about public sensitivity to warnings. In such a condition, if terrorists perceive lack of public sensitivity to warnings they may not cancel their plans to attack as frequently as they would if the public was responsive to warnings. This hypothesis leads to a loop (known as a balancing loop in the systems science literature, see Sterman (2000)) in which when false alarms are issued (or perceived), public sensitivity ($S_{\text{public}}$) decreases, and that leads to a decline in adversaries’ perception of public sensitivity ($\overline{S}_{\text{public}}$). As result, the chance of canceling a possible attack decreases, and more likely the society faces an attack. We can define an elasticity parameter to test how such a hypothesis can influence the system. Equation 2.4 represents such a formulation:

$$P_{\text{canceling}} = f_3 (\overline{S}_{\text{public}}, \beta)$$

Equation 2.4

$\beta$ represents deterrence elasticity to perception of public sensitivity, that is $\frac{\partial f_3}{\partial S_{\text{public}}} = \beta$.

When $\beta = 0$, that is, canceling probability is constant (the hypothesized loop doesn’t exists), otherwise the canceling probability changes and the adversaries react to changes
in the public sensitivity to alarms. Details of this formulation are presented in Appendix 2.3.

Figure 2.3 compares the base run simulation for different scenarios to examine the effect of terrorists’ reaction to public sensitivity to warning. As we see in Figure 2.3.c, total damage increases if terrorists update their perception of public sensitivity.

So far we developed a model of warning based security and examined dynamic base runs under different conditions to describe the dynamic trend of attacks and damages. In the next section we use the model to find optimal warning thresholds that minimize total damage in long-term under different scenarios.

2.4. Optimization

There are two major cost components in this model. First is the damage cost \( (C) \) which is the result of attacks and the reaction from the public and private. The second cost component is the psychological cost associated with warnings \( (C') \). As more warnings are issued more psychological costs are imposed on society. We examine the cost components under the major scenarios for hypothetical values for the parameters, and then analyze optimal threshold placements in these scenarios. The main goal of this exercise is to examine the sensitivity of optimal thresholds to the assumptions about behavioral reactions in the society.
Examination of the cost components

First, we run a set of simple simulations to examine the cost components as threshold for warning changes. In this step, we assume $T_{public} = T_{private} + 1$, and focus on the effect of one variable. Later in this section, we will optimize both thresholds. Figure 2.4 shows total warnings and total damages from 4,000 simulation runs in three different scenarios (three different assumptions about how people react to warnings). The x-axis represents the threshold for public warning, so as we go toward the right, the thresholds increase and fewer warnings will be issued. The left side y-axis is the average of damages that occurs. The right side y-axis is the “average of warning” index, representing how much warning is issued\(^{10}\).

We examine three major scenarios representing three major assumptions about behavioral reactions in the system. The first scenario is for constant sensitivity to warnings ($\alpha_1 = \alpha_2 = \beta = 0$). The second scenario is for inconstant sensitivity to warnings among public and private ($\alpha_1 = \alpha_2 = 1$ and $\beta = 0$). In the third scenario not only sensitivity to warnings changes endogenously but also adversaries react to those changes ($\alpha_1 = \alpha_2 = \beta = 1$).

\(^{10}\) This index is defined as a weighted average of private warning and public warning in each time period, i.e. \[ \frac{\sum_{t=1}^{\tau} (W_{public} + K.W_{private})}{\tau} \]. Different values can be assumed for K, a hypothetical weight to represent the difference between psychological costs of a public warning in comparison to a private warning. We set K to be 0.2, and later tested for different values. The results are not qualitatively sensitive.
In this figure, we see that there should be an optimal threshold placement for warning under different scenarios. For example, in the first scenario (Figure 2.4.a), an increase in threshold for public warning (fewer warnings) results in more damages from attacks (the physical cost component). That happens basically as a result of less defensive reaction from public and private sectors. Further, an increase in warning thresholds results in less psychological warning costs. In sum, as the threshold for warning increases, one of the cost components increases and the other decreases. Although the graphs do not have comparable scales, the direction of effects of a change in threshold on these cost components is different which shows that a single optimal threshold exists.

In the second scenario (Figure 2.4.b), similar patterns exists; however, due to the effect of false alarms, decreasing warning thresholds will not be as influential as they were in the previous scenario. Under this scenario, issuing more warnings cause more reactions, but in the same time results in more false alarms and loss of sensitivity. Combining these effects decreases the preventive impacts of warnings. In short, very low warning thresholds are not as damage – preventive as they were under the previous scenario. This results in a shift in optimal threshold to the right.

In the third scenario (Figure 2.4.c), however, decreasing warning thresholds and issuing more warnings have side effects and result in both more psychological costs and physical costs. As we see the relation of warning threshold to average damage is a u-shaped relation, and at the extreme as we issue more warnings there will be also more damages. Loss in public sensitivity, when perceived by adversaries, results in attacks that may have been cancelled if the public was sensitive to warnings.
While in these figures we see that there should be an optimum place for the warning threshold, but without knowing the social values that a society puts on warnings in comparison to a physical damage, it is impossible to compare the graphs and find the optimum threshold placement. However, it is plausible to conclude that the optimal warning threshold increases in the second scenario in comparison to the first scenario, when we relax the assumption that sensitivity to warning is always high and constant. Further, in the third scenario, the optimal warning thresholds increase more as adversaries perceive changes in public sensitivity.

Figure 2.4: Two major cost components, physical and psychological costs versus threshold to issue public warning under three main conditions.

(a) Physical and psychological costs versus public warning threshold for constant public and private sensitivity
(b) Physical and psychological costs versus public warning threshold under the condition of inconstant public and private sensitivity

(c) Physical and psychological costs versus public warning threshold for inconstant social sensitivity to warnings and inconstant reactions from adversaries
Optimal thresholds distribution

For the purpose of optimization, we would like to minimize the total cost (physical and psychological costs) in a long term period. Using $t$ as a subscript to represent the value of a variable at time $t$, we define the utility function as following:

$$\text{utility} = \sum_{i=1}^{n} (C_i + C'_i)$$

Equation 2.5

Maximizing the utility function results in finding optimal decisions that lead to lower total costs, i.e., the sum of costs from damages and costs associated with warning issuance including social stress. In following, the goal is to find the thresholds in a way that maximizes the utility function.\(^{11}\)

Due to the existence of stochastic components in our model we conduct optimization for 100 different random seeds where each optimization is conducted from 50 different initial points. We examine the distribution of optimal thresholds in three different scenarios (total of 15,000 simulations).

Figure 2.5 shows the results of optimization of two variables, threshold for private warning, and threshold for public warning (columns), under three different scenarios (rows). In each condition, the distribution of optimal solutions is shown. What this figure displays is the shift in the distribution of optimal solution as a result of change in our assumptions about elasticity parameters. As we see the optimal threshold (especially for the threshold for public warning) changes as we change elasticity parameters. When we assume people’s sensitivity to alarms is constant the model proposes more warnings, and as we move more toward endogenous changes in sensitivity, the model proposes higher thresholds of warnings, i.e. less warning issuance. Presenting the distribution of optimal

\(^{11}\) Hypothetically, we assume the social cost from 400 public warnings is equal to a successful attack.
solutions rather than a single solution shows that our argument is independent from random seeds and is a robust finding.

The main lesson from this exercise is about the possibility of making sub-optimal decisions if we make irrelevant assumptions about how people behave when they are overwhelmed with warnings. These figures show that optimal solutions are sensitive to two major assumptions: sensitivity to false alarms and adversaries’ perceptions of public sensitivity. An underestimation of these effects results in sub-optimal situations. This is against the common observation in security contexts where decision makers think if they issue more warnings people will react more defensively and higher security levels will be achieved.

Another lesson is that for different societies or organizations with different cultures, optimal solutions can be different. Some settings are more prone to the crying wolf effect than others. This suggests the examination of behavioral characteristics of settings when deriving optimal warning decisions.

Furthermore, we see that the optimal threshold for public warning is sensitive to our assumptions about the society. That is because adversaries react to public warnings, so irrelevant assumptions about effects of public warnings become more influential in security analysis.
Assumption 1: Constant sensitivity to warnings

Assumption 2: Inconstant sensitivity to warnings

Assumption 3: Inconstant social sensitivity to warnings and inconstant reactions from adversaries

Figure 2.5: Distribution of optimal thresholds under different assumptions about how people react to warnings

Note: This figure shows the optimal thresholds are sensitive to our decision about how people and adversaries react to warnings: Irrelevant assumptions result in a bias in optimal decisions, especially in public warning threshold.
2.5. Conclusion

Although warning issuance is a common action to increase security, people’s response to warnings is complicated and is influenced by several psychological phenomena. Examination of warning issuance in the absence of a thorough understanding of how people perceive warnings, and how they react to warnings can lead to severely suboptimal public safety. Developing a model of warning that encompasses important and relevant psychological characteristics was the goal of this paper.

In this paper, we developed a simulation model based on signal detection framework to address public and private warning optimization problem. First the problem of warning was formulated using the signal detection framework. Such a formulation helped us to represent warning decision models with two thresholds, and therefore represent optimization by finding optimal warning thresholds. Then in three major steps we formulated a model of security warnings which gave us the opportunity to test the effects of different behavioral assumptions about how guards, public, and adversaries react to warnings.

This research has two major contributions. First, the extended model helps us to find numerical solutions for the framework in a dynamic context and examine optimal private and public warning decisions. Representing warning decisions with warning thresholds and clarifying the sources of error in warnings helps find optimal solutions. Once the values that the society assigns for different outcomes are estimated, the model can be used for finding optimal solutions.

Second, through the extended model we showed that optimal solutions are sensitive to two major assumptions: public sensitivity to false alarms and terrorists’ perception of
public sensitivity. Based on the assumption that as the public becomes more sensitive to false alarms, public responsiveness to alarms may decrease, we conclude that optimal solutions can be influenced by how sensitive the public is to false alarms. Furthermore, if this sensitivity is perceived by adversaries while the magnitude of damage can change, the optimal decision threshold changes. In summary, an underestimation of these effects can result in biases in optimal solutions.

There are several ways to extend this model. First, effects of change in the base rate of attacks or crimes can be examined. Further, the effect of the intelligence system’s capabilities on optimal warning decisions should be examined. Finally, this model can also be used as a descriptive model to explain effectiveness of warning decisions in different contexts. The model can also be calibrated for different cultural contexts and examined to see how well it can describe warning policy performance.

In short, this paper warns about underestimation of the importance of behavioral properties of different contexts for warning decision making. As the way that people respond to warnings can be different in different cultural settings, optimal warning thresholds can be different in different societies. Further, an optimal warning policy for a short term period can be sub-optimal in long run and result in more physical and psychological damages.
References


*Science* 185 (4157): 1124–1131.


Appendix 2.1: Formulation of different entities’ reaction to warnings

This appendix presents details of formulation of Equation 2.2, how damage \( C \) is formulated as a function of attack \( (a) \) and public and private reactions \( R_{\text{public}}, R_{\text{private}} \).

\[
C = f_1(a, R_{\text{public}}, R_{\text{private}}) \quad \text{Equation 2.2}
\]

In order to formulate attacks \( (a) \), we should consider deterrence possibilities. The deterrent effect may result in canceling a possible attack \( (\bar{a}) \). If \( e \) represents the deterrent effect, which picks one of the two values of 0 and 1, attacks can be represented as following:

\[
a = \bar{a}.(1- e) \quad \text{Equation 2.6}
\]

whereby, \( \bar{a} \) is 1 when there is a positive event (there is possibility of facing an attack) and is zero when there is a negative event (there is no possibility of facing an attack).

Let’s assume that adversaries cancel a possible attack (positive event) when a public warning is issued with the probability of \( p_{\text{canceling}} \). Conceptually that can happen when a true positive warning is issued \( (\bar{a}.W_{\text{public}} = 1) \). Equation 2.7 represents the probability of deterrent effect:

\[
Pr(e = 1) = 1 - Pr(e = 0) = \bar{a}.W_{\text{public}} \cdot p_{\text{canceling}} \quad \text{Equation 2.7}
\]

So, as \( e = \{0,1\} \), in each time period, \( a \) will be either 1 or zero.

Then, we can formulate public and private reaction and thus the total damage cost \( (C) \) from an attack. We expect public reaction \( (R_{\text{public}}) \) and private reaction \( (R_{\text{private}}) \) to result in reducing costs from an attack.

\[
C = a.C_{\text{max}} \cdot g_1(R_{\text{public}}) \cdot g_2(R_{\text{private}}) \quad \text{Equation 2.8}
\]

\( C_{\text{max}} \) is the maximum costs associated with an attack. For simplicity we assume \( C_{\text{max}} = 1 \). \( g_1 \) and \( g_2 \) represent two functions for how public and private reactions decrease.
possible damages from an attack. Both functions should be defined in a way that as $R_{\text{public}}$ and $R_{\text{private}}$ increase, damage cost from a potential attack decreases. We formulate $g_1$ and $g_2$ as: 

$$g_i(x) = 1 - k_i x,$$

where $0 < k_i < 1$. As it appears from Equation 2.8, damage cost takes the maximum value of 1 when an attack happens ($a = 1$) and neither public nor private react ($R_{\text{public}} = R_{\text{private}} = 0$).

Equations 2.9 and 2.10 represent reaction to warnings. $S_{\text{public}}$ and $S_{\text{private}}$ are public and private sensitivities to warnings, and both are less than or equal to one.

$$R_{\text{public}} = W_{\text{public}} \cdot S_{\text{public}} \quad \text{Equation 2.9}$$

$$R_{\text{private}} = W_{\text{private}} \cdot S_{\text{private}} \quad \text{Equation 2.10}$$

As we see in Equations 2.9 and 2.10, no warning result in no reactions, and higher sensitivity to warnings result in higher defensive reactions.
Appendix 2.2: Formulation of change in sensitivity to warnings

This appendix presents details of formulation of Equation 2.3, how sensitivity \((S_i)\) in section 3-2 of the paper is defined as an endogenous variable in the system.

\[ S_i = f_2(C, W, \alpha_i) \quad \text{Equation 2.3} \]

In order to formulate sensitivity \((S_i)\) as a function of recent alarms and recent physical damages, we define CWR which is the ratio of recent damages to recent alarms. We expect a higher ratio to result in higher sensitivity to future alarms. Equations 2.11 and 2.12 represent such an effect:

\[ CWR_i = \frac{C}{W_i} \quad \text{Equation 2.11} \]

\[ S_i = (1 - \alpha_i)S_{\text{max}} + \alpha_i g_3(CWR_i) \quad \text{Equation 2.12} \]

As it appears from Equation 2.12, \(S_i\) is the weighted average of \(S_{\text{max}}\) and \(g_3(CWR_i)\). When \(\alpha_i=0\) (the elasticity parameter), we have \(S_i = S_{\text{max}} = \text{constant}\). Otherwise, sensitivity will be influenced by the ratio of recent damages to recent warnings. \(g_3\) represents such a relation, whereby \(g_3(0) = 0\), and \(g_3(.) > 0\). For simplicity, we assume that \(g_3(x) = \min\ (S_{\text{max}}x^{0.2}, S_{\text{max}})\). Changing \(\alpha\) allows us to examine different assumptions about how people react to false alarms and previous attacks: \(\alpha = 0\) will represent constant sensitivity to alarms, and higher \(\alpha\) represents higher sensitivity to the history of alarms and attacks.
Appendix 2.3: Formulation of adversaries’ reaction to social sensitivity

This appendix presents details of formulation of Equation 2.4, how probability of canceling an attack ($P_{\text{canceling}}$) in section 3-3 of the paper is defined as an endogenous variable in the system.

$$P_{\text{canceling}} = f_3(\bar{S}_{\text{public}}, \beta)$$  \hspace{1cm} \text{Equation 2.4}

Adversaries’ reaction to warnings is formulated as a function of public sensitivity. It is assumed that adversaries have a perception of public sensitivity ($\bar{S}_{\text{public}}$) which lags public sensitivity to warnings ($S_{\text{public}}$) with some delays (in our model after 1 month).

$$P_{\text{canceling}} = (1 - \beta)P_{\text{canceling, max}} + \beta g_4(\bar{S}_{\text{public}})$$  \hspace{1cm} \text{Equation 2.13}

As it appears from Equation 2.13, $P_{\text{canceling}}$ is defined as the weighted average of $P_{\text{canceling, max}}$ and $g_4(\bar{S}_{\text{public}})$. We should define $g_4(x)$ in a way that $g_4(1) = P_{\text{canceling, max}}$, $g_4(0) = 0$, and $g_4(.) > 0$. For simplicity, we assume $g_4(x) = P_{\text{canceling, max}} \cdot x$.

In this representation, as perception of public sensitivity to alarms among adversaries ($\bar{S}_{\text{public}}$) decreases, $P_{\text{canceling}}$ gets further from $P_{\text{canceling, max}}$ representing less deterrent effect.
ESSAY 3:

Beyond Personality Traits and Financial Incentives: Bias and Variation in Medical Practices as a Systematic Result of Experiential Learning
Abstract

There are several indicators of abundance sub-optimal decisions in medicine. Two of the common ones are overuse of defensive medical practices such as medical tests (bias toward more tests), and variation of medical diagnoses and treatments for medically similar patients (practice variation). Besides patients’ characteristics and preferences, and the regional characteristics (such as culture), at the individual level, the most common explanations for practice variation and bias are linked to physicians’ personality traits (e.g., the level of risk aversion) or their financial incentives. We develop a theory that offers a new explanation for variation and bias in practice. With the help of a simulation model, we show that practice variation and bias does not have to be caused by personality traits and financial incentives, but can endogenously emerge through daily practices and outcome learning even for physicians with similar trainings working in the same region. In other words, the characteristics of medical tasks themselves can result in practice variation and bias. Specifically, a physician’s exposure to outcome feedback, a physician’s ability to evaluate different forms of practice and a physician’s accumulated experience with a given approach all contribute to practice variation and bias. A preliminary validation of the results is achieved by comparing projected results with the actual data from cesarean section surgery in the states of New York and Florida.
3.1. Introduction

There are several indicators of abundance sub-optimal decisions in medicine. Two of the common ones are overuse of defensive medical practices such as medical tests, and variation of medical diagnoses and treatments for medically similar patients (Fisher, Bynum, & Skinner, 2009; Institute Of Medicine, 2000, 2003; Wennberg, Fisher, & Skinner, 2002; Wennberg, Freeman, & Culp, 1987; Wennberg & Gittelsohn, 1973). The Institute of Medicine (2000) estimates that 98,000 people die in the United States hospitals every year as a result of preventable mistakes, and the Dartmouth Atlas of Health Care argues that decision making factors are central in health disparities in the country (Fisher et al., 2009). Medical errors and inefficient decisions, in aggregate, contribute to an inefficient healthcare system, a major concern in the current US healthcare reform.

Two of the most common indicators for sub–optimal practices in medicine are practice variation and over utilization bias. First, it is shown that different doctors do not make similar decisions for medically similar patients, but in fact in many cases they disagree. For example, controlling for patients’ health risks, different obstetricians have different rates of c-section surgeries (Epstein & Nicholson, 2009). Similar patterns of disagreement across different physicians happen in prescribing cancer diagnostic tests and treatments (Bynum, Song, & Fisher, 2010), pediatric services (Sorum et al., 2002), and psychiatric services (Way, Allen, Mumpower, Stewart, & Banks, 1998). Practice variation for medically similar patients has been argued to be an indicator of sub-optimal healthcare system (Fisher et al., 2009).
Second, in average physicians prescribe more tests than it is needed and they incorporate more surgeries than necessary (bias toward over utilization of resources). The current discussion around the optimal frequency of mammography is an example of when a standard of efficiency (as determined by an expert panel) may not be used in practice (Welch, 2010). Similar arguments have been made for frequency of other medical tests and over utilization of medical resources (Bynum et al., 2010).

Regional characteristics have been argued to result in variation in medical expenditure (Fisher et al., 2009; Sutherland, Fisher, & Skinner, 2009). The Dartmouth atlas of healthcare offers a lot of evidence that in some regions there is more health expenditure than others. In these studies, they find that only 30% of the excess spending in the highest cost regions can be related to income and health and the rest are regional factors. In other words, some regions have higher level of health expenditure due to the cultural and organizational factors in these regions. Such a higher level of expenditure does not result in a better outcome (quality of healthcare services), and therefore is an indicator for an inefficient healthcare system (Sutherland et al., 2009). Although they offer explanation for across-region variation, they leave the question of why in the same region practice variation can exists, i.e. why doctors that are performing in the same region differ significantly (Epstein & Nicholson, 2009).

Besides patients’ preferences and regional and organizational factors, two of the most common explanations that are offered for sub-optimal medical practices are physicians’ financial incentives and/or their risk avoiding behavior. Financial incentives have been argued to be a reason for over-utilizing resources. For example fee for service models are argued to create more incentives to perform more services in contrast to other financial
systems (Bodenheimer T. S. & K., 2005). It is expected that the financial systems that give more incentives to physicians to overprescribe medical tests and treatments lead to inefficiencies, and the difference in financial systems and physicians’ personal financial interests cause practice variation.

Non-financial reasons are also offered to explain the sources of sub-optimal behaviors. One of the common explanations for practice variation is the physician-specific factors such as risk aversion. These factors are known to persist over time, and have been argued to be difficult to measure outside of the laboratory (Epstein & Nicholson, 2009). For example, one way to avoid any risk of making a wrong decision is to prescribe medical tests for a larger population of patients. In such cases, physicians who are more risk averse would prescribe more tests. Furthermore, uncertainties have been argued to contribute to imperfect decisions and disagreements across doctors. Uncertainties can make it difficult to come up with a consistent decision, and it can result in more risk aversion. For example, in a high uncertain situation, higher risk aversion can lead to more defensive practices through abundant prescription of medical tests.

This study offers a new explanation for practice variation beyond the current arguments. Controlling for most of the already discussed factors in the literature, we argue that the suboptimal medical practices in the form of heterogeneity or bias can appear through daily practices due to the characteristics of medical tasks. We hypothesize that outcome feedback, the process of judging effectiveness of different styles of practice, and the process of experience accumulation when combined with environmental uncertainties lead to heterogeneity in medical practice and bias toward overutilization of defensive practices. Through a simulation experiment, we show that for (mathematically)
similar physicians visiting a similar population of patients, bias and disagreement emerge as physicians practice, receive information, and gain experience. In such cases, it is not necessary to assume different doctors have different financial incentives and preferences, or different personality traits. Instead, our findings indicate that sub-optimal decisions in the form of bias and variation can emerge as a result of task characteristics and daily practices. A preliminary validation of the results is achieved by comparing them with the data from cesarean section surgery in the states of New York and Florida. In the next sections, we review the research background, the model, and simulation runs.

3.2. Background

This study builds on the literature of medical decision making and experiential learning. We review the relevant studies in both domains, and show, first, how the problem of bias and variation in medical practices is been studied, and second, how learning from outcome feedback is studied in the literature of decision sciences. We argue that experiential learning in medicine has unique characteristics. Then, we discuss how we can shed more lights on the problem of practice variation and bias in medicine through understanding the experimental learning processes and experience accumulation.

Sub-optimality in medical decision making

Medical decision making is prone to different forms of inaccuracies. Ghaffarzadegan and Martin (2011) review a lot of evidence for unreliability, disagreement and bias in medical practices (Ghaffarzadegan & Martin, 2011). For example, unreliability is a common problem in medical decisions, and a physician can make different and
inconsistent diagnoses or decisions for the same case or medically similar cases in
different time periods (Einhorn, 1974a, 1974b; Institute Of Medicine, 2003; Kirwan,
Chaput De Saintonge, Joyce, & Currey, 1983; Koran, 1975; Levi, 1989; Little, 1961;
Millimet & Greenberg, 1973). Such unreliability in decision making can happen as
results of uncertainties in the environment, which can affect the way that information
acquired or interpreted (Stewart, 2001). It is shown that a physician’s initial belief and
diagnosis about a patient can influence the way that further information is acquired. In
turn, biased information acquisition limits possibilities of correcting the initial diagnosis
if it is necessary (Kostopoulou, Mousoulis, & Delaney, 2009).

Sub-optimal decisions are also found in cross-physician comparisons, usually referred
to as practice variation. Practice variation across physicians happens when different
physicians disagree and make different decisions for medically similar patients, a
common phenomenon in medicine (Gonzalez-Vallejo, Sorum, Stewart, Chessare, &
Mumpower, 1998; Sirovich, Gaiiagher, Wennberg, & Fisher, 2008; Sorum et al., 2002;
Way et al., 1998). Evidence on practice variation includes a wide range of studies on
different diseases. For example, Gonzalez-Vallejo et al. (1998) and Sorum el al. (2002)
conduct experimental studies of practice variation in pediatric services. They give
vignettes to pediatricians representing kids with different health indicators and ask them
to make judgments about the probability of acute otitis media. They find that there is a
considerable variation in the way information cues are used and weighted to make
diagnosis, and in the kind of antibiotics prescribed. Way et al. (1998) conduct an
experiment with psychiatrists. They ask a psychiatrist to make a diagnosis based on
several taped interviews of patients at psychiatric emergency services. They also find a
wide range of variability in diagnosis in different diseases. Sirovich et al. (2008) investigate the prescription of cancer screening with a national sample of U.S. physicians. They send survey-based vignettes to doctors and ask about their decisions on different tests such as a prostate screening or mammography. They also find evidence of practice variation across doctors. However, the magnitude of variation usually depends upon the level of uncertainties in the disease or information presentation. These studies in general show that physicians make sub-optimal decisions due to characteristics that are not easily observable (characteristics other than gender and race) which sometimes account for more than 47% of variation (Grytten & Sorensen, 2003). Such characteristics are usually captured in a random effect model with a dummy variable for each physician (DeSalvo K.B., Block J.P., Muntner P., & Merrill W.W., 2003; Epstein & Nicholson, 2009).

The studies of sub-optimal medical practices have a strong tie with health policy research due to the considerable policy implications of understanding the source of unnecessary expenditure and inefficiency in medicine. The Dartmouth Atlas Project has claimed for a long time that a significant portion of regional variation in health service expenditures relates to individual and organizational decision making characteristics rather than technological, socioeconomic, health and income differences. For example, they show that around 70% of the excess spending in the highest cost regions are not attributable to different income, health, or technological factors, but are related to other regional factors (Fisher et al., 2009; Sutherland et al., 2009). They stress the importance of understanding decision making sources of variation in practice to offer proper policy implications. The current study aims in a similar direction and offers an in-depth
understanding of how decision making characteristics can result in variation and bias in practice.

**Experiential learning in medical decision making**

A person’s decision model is usually affected by feedback that they receive about the outcomes of their previous decisions. Such a learning process is referred as experiential learning (Cyert & March, 1963; Levitt & March, 1988; Nelson & Winter, 1982) and exists in many natural settings. Outcome feedback in general can help people to learn. When people receive clear information about the results of their past actions, they can improve their decision making. However, there also a lot of other evidence that outcome feedback may lead to sub-optimal decisions rather than the best possible decisions (Huber, 1991; Lant, 1992; Levinthal & March, 1993; Miner & Mezias, 1996; Rahmandad, 2008). For example it is argued that when outcome feedback is delayed, it is very difficult for decision makers to interpret the results of their decisions and learn (Rahmandad, 2008).

Understanding of the process of experiential learning in a task was improved when Denrell and March (2001) showed that this process can result in systematic bias toward more certain choices. Based on a simulation model, they show that the reproduction of successful choices can result in a bias against risky alternatives. When people have initial unsuccessful experiences, they may avoid trying them again which in turn limits the opportunity of getting to know the true value of a choice in an experiential learning context – i.e., the hot stove effect (Denrell & March, 2001).
In the context of medicine, outcome feedback can play an important role as well. Physicians accumulate knowledge and experience through a kind of feedback that is usually contingent upon decisions, what is referred as conditional feedback. For example, if a physician decides to prescribe a mammography test, she will know if her judgment about a patient was correct or not. Otherwise (not prescribing a test) she may not know if her judgment was correct until much later if at all. Such conditionality in feedback is argued to cause overconfidence and underestimation of the base rates (Dalgleish & Smillie, 2006; Einhorn & Hogarth, 1978; Elwin, Juslin, Olsson, & Enkvist, 2007; Fischer & Budescu, 2005; Stewart, Mumpower, & Holzworth, 2011; Tindale, 1989). Even when information is easily available for physicians, they may still have a problem with information acquisition and interpretation (Kostopoulou et al., 2009). For example, it is argued that psychiatrists gain more confidence as they receive more information without necessarily changing their initial belief (Oskamp, 1965).

The current study

This study is a systematic investigation of effects of experiential learning on medical practices and on sub-optimal practices in the forms of bias and variation. We focus on the case of obstetricians and study bias toward too many cesarean section surgeries and variation in surgery rates across obstetricians. Based on a simulation model, we show that it is not necessary to assume heterogeneity in personality traits and financial incentives of physicians to have practice variation, but variation can endogenously emerge through daily practices for physicians that even visit similar populations of patients. Then we
generalize the model to a wider range of conditions and derive a theory of bias and variation in medical practice.

While this study does not reject previous studies, it offers a feedback–based theory of sub–optimal medical practices. The new theory can have unique policy implications. Our modeling practice shows that for (mathematically) similar physicians, practice variation and bias can emerge as results of the processes through which they receive information, interpret the results, and accumulate experiences. Based on previous experimental and simulation studies (Elwin et al., 2007; Fischer & Budescu, 2005; Ghaffarzadegan & Stewart, 2011a, 2011b; Stewart et al., 2011), we believe that the type of feedback influences accuracy and reliability in decision making. In the medical context, the literature lacks understanding of the effects of outcome feedback on physicians. We expect such an investigation to be a significant step to improve agreement, reliability, and accuracy in medical decision making. In the next section, we discuss sub-optimality across obstetricians as a case of practice variation and bias in medicine. Then we build a simulation model of the case and later generalize the results to a wider range of medical domains.

3.3. Sub-Optimality Across Obstetricians

There are several reasons for Obstetrics to be a proper and an important case of sub-optimal decisions in medicine. Obstetrics is prone to suboptimal decisions in the forms of bias and variation in practice. Cesarean section surgery have been argued to be over-performed (in 2010, 31% of birth cases in US), and in many cases for nonmedical reasons (O'Callaghan, 2010). Patients get admitted to c-section surgery either through a pre-
scheduled process where their doctor suggests the surgery before the due date, or during the vaginal delivery when the doctor decides to switch to a c-section surgery as a result of a new diagnosis. In additions to the costs of surgery and longer stays in hospitals, cesarean section surgery can cause more risks for healthy mothers and babies (O'Callaghan, 2010). While it is difficult to find optimal c-section rate, the 31% rate in US, both intuitively and based on comparison with other countries, is considered as a high c-section surgery rate.

In addition it is found that physicians differ in their tendency to schedule a cesarean section surgery. The observation has been robust in many studies that control for patient’s health status with different indicators showing that variation is more a practice style issue. Epstein and Nicholson (2009) investigate variation in cesarean surgery rate in New York and Florida. Controlling for patients’ risk factors, they show that within and across regions there is a considerable variation in caesarean section surgeries and some physicians are more inclined to conduct surgery than others. They show that variation within a region is two times more than variation across regions. This is an indictor that in contrast to many other studies that claim regional differences can result in variation in practice, they show that there is also something about each individual physician that can result in practice variation. Interestingly, the variation in the style of practice persists over time and they claim that physicians do not converge to a community standard.

Epstein and Nicholson (2009) shows within-region variation in c-section rate for 1992–2006 for the states of New York and Florida. They first calculate the probability that each physician will perform a c-section, controlling for observed characteristics of mother’s health and the status of pregnancy. This is called risk-adjusted c-section rate an
indicator for a physician’s treatment style. Then they calculate the difference between a physician’s treatment style and the mean treatment style for the physician’s region at the same time period. Looking at the distribution of these data points can show variation in treatment styles (if everybody is practicing similarly, then there is no deviation from mean). They find that standard deviation of the distribution is 6.5 percentage points. They also find that in 24 percent of the cases, the physician’s risk adjusted c-section rate is statistically different from the regional mean at the five percent level.

An obstetrician may perform more than 100 surgeries a year and one may expect that through these practices, she should learn about optimal decisions and make more accurate judgments about her patients. However, there are many complexities in the obstetric practices which make it difficult to learn.

One of the major sources of complexities is about the way that an obstetrician observes her decision outcome and the way the outcome is interpreted. In fact, in the context of obstetrics outcome feedback is asymmetric and contingent upon the decision, what is usually referred as conditional feedback in the literature of behavioral decision making. Let’s assume that patients can be categorized into two groups, the ones with higher health risks that should go under c-section surgeries and the ones with lower risks that can deliver through vaginal birth. A doctor is not necessarily able to differentiate patients based on their true status, but she makes a judgment based on available data and she might make a wrong decision.

Based on an obstetrician’s decision, which can be to conduct a vaginal birth or a c-section surgery, four decision outcomes as shown in Table 3.1 can happen which are true positive, false positive, true negative, and false negative. In addition, as shown in the
Outcome feedback in these four conditions is not similar. In the case of vaginal birth when the decision is a false negative, a doctor can observe her decision result during the practice and may even change her decision and conduct a c-section surgery for a patient that is in labor. But under false positive decisions, when a surgery is scheduled, c-section surgery will be conducted any way, and there is little clear outcome feedback. In sum, an immediate and clear feedback exists on vaginal birth, but in c-section surgery, a high portion of poor outcomes will not be observed and may be attributed to patients’ health risks.

<table>
<thead>
<tr>
<th>True status of patient</th>
<th>Obstetrician’s initial decision</th>
<th>Patients that should have c-section (High risk patients)</th>
<th>Patients that should have Vaginal birth (Low risk patients)</th>
</tr>
</thead>
<tbody>
<tr>
<td>False negative</td>
<td>Vaginal birth</td>
<td>In most cases, the obstetrician will observe the decision outcome immediately and sometimes may change the decision to a c-section surgery</td>
<td>A correct decision, and if the decision is performed well in practice should have a proper outcome.</td>
</tr>
<tr>
<td>True negative</td>
<td>C-section surgery</td>
<td>A correct decision, and if the decision is performed well in practice should have a proper outcome.</td>
<td>An unnecessary surgery, with side effects. However, feedback is unclear, delayed and can be attributed to other factors than a wrong decision.</td>
</tr>
<tr>
<td>False positive</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3.1: Four possible outcomes for an obstetrician’s decision about vaginal birth vs. c-section surgery for different patients

Table 3.1 represents four different decision outcomes and conditionality of outcome feedback in the context of obstetrics. Feedback asymmetries exist in many other medical contexts. In the following, we focus on the cesarean section surgery as our case for
modeling to make more sense of different concepts and variables in the model. We will later discuss how a general theory of practice variation can be developed based on the lessons from this modeling practice.

3.4. Modeling

Our unit of analysis (an agent) is a physician. Our model represents mathematically similar physicians through their daily practices as they make decisions, learn and accumulate experience.

The model has three state variables including: the threshold for c-section (C), experience of c-section (E₁), and experience of vaginal birth (E₂). An overview of the model is depicted in Figure 3.1. The basic logic of the model is as follows. In each time period, a doctor visits a patient with a level of health risk (x), and makes a judgment about her health risk (x). Then the doctor compares her perception of the health risk with her threshold to perform a c-section (C). If the perceived risk is higher than the threshold, the doctor makes and performs a c-section decision, otherwise a vaginal birth. The outcome of the practice (y) depends on how good the doctor’s decision was for this specific patient (dₓ), and how well the surgery was performed. Then, through a specific form of outcome feedback that was explained in table 3.1, the doctor perceives the result of her decision, perception of practice outcome (y). If the perceived result is not good enough, the doctor tries to correct her threshold to conduct c-section (C) in a way that if the outcome from a c-section surgery was not positive she increases her threshold, and if the result from a vaginal birth was not positive she decreases her threshold to perform more c-section surgeries. Through the whole process, the doctor accumulates experience
of vaginal birth and c-section surgery ($E_i$), which in turn affects her skill of performing ($S_i$).

Figure 3.1: A simplified causal diagram of the medical practice model for obstetrics practice for a single obstetrician

Note: Equation signs on the graph help to track the model formulations in this section (eq. 1-8 represent Equations 3.1 – 3.8 in the text). The dashed links will be activated under conditional feedback. The variables with indices (e.g., Experience) are array variables and take different values for c-section and vaginal delivery.

The depicted conceptual model in Figure 3.1 is formulated in three steps: decision making process, learning process, and gaining skills.
**Decision Making Process**

Decisions to perform a surgery are assumed to be based on diagnoses. A physician decides to perform a c-section surgery if she diagnoses the level of risk to be high enough. Equation 3.1 represents this decision making principle. For each doctor, the model considers a threshold for performing c-section (C), and if the doctor perceives her patient’s health risk (\( \bar{x} \)) to be higher than the threshold, she makes a c-section decision (d=1). In Equation 3.1, d represents the physician’s decision, and is 1 for c-section decisions and zero for vaginal delivery decisions.

\[
\begin{align*}
    d = 0 & \quad \text{if} \quad \bar{x} < C \\
    d = 1 & \quad \text{if} \quad \bar{x} = C
\end{align*}
\]  

(Equation 3.1)

Equation 3.2 represents a physician’s judgment (diagnosis) of her patient’s health risk. We assume no systematic bias in judgment, but a simple random error in diagnosis (\( \varepsilon \)).

\[
\bar{x} = x + \varepsilon
\]

(Equation 3.2)

where x is the true health risk of the patient.

As stated, in our model there are two alternatives for a doctor on each case: to perform a c-section or to perform a vaginal delivery. The outcome of the practice can depend on many factors, two of the most important ones are a) how effective the decision was (i.e., does the decision fit the patient’s health characteristics), and b) how well it was performed by the physician (i.e., a physician’s skill in performing the decision). For example, for a patient who is in an extremely high risk situation, an effective decision is a c-section decision, and the result of the practice also depends on the doctor’s skill in performing c-section. We formulate the practice outcome (y) as a Cobb–Douglas function.
of how effective a decision is \( (d_e) \) and how skilled the physician is in performing that decision \( (S_d) \). Equations 3.3 and 3.4 represent these relations.

\[
d_e = f(d, x) \quad \text{(Equation 3.3)}
\]

\[
y = S_d^\alpha d_e^\beta \quad \text{(Equation 3.4)}
\]

where \( f \) should be defined in a way that \( f(0,x) \) represents how proper a vaginal delivery decision is for a patient with the health risk of \( x \), and \( f(1,x) \) represents how proper a c-section surgery is for a patient with the health risk of \( x \). We expect that a vaginal delivery decision to be more appropriate for the lower levels of health risk, and a c-section surgery to be more appropriate for the higher level of health risk. In other words, for \( d=0 \) we have \( \frac{\partial f}{\partial x} < 0 \), and for \( d=1 \) we have \( \frac{\partial f}{\partial x} > 0 \).

**Learning Process**

We assume a physician’s interpretation of her decision outcome \( (\bar{y}) \), can help her learn and correct her decision threshold. In simple words, if a physician sees that her patient with a health risk lower than her threshold did not perform well under vaginal delivery, the physician understands that her threshold was too high, and for next patients she should decrease her decision making threshold. Equivalently, if she sees that a c-section surgery was not good for a patient, she increases her threshold in next practices. The change in threshold, therefore, will be in the direction of the observed risk. The size of change in threshold \( (\Delta C) \) is assumed to be driven by the size of the mistake \( (\bar{x} - C) \), the magnitude of shortfall in decision outcome \( (y_{\text{max}} - \bar{y}) \), and a constant parameter,
threshold adjustment speed \( (k) \), representing how fast the physician changes her decision model. Equation 3.5 shows this relation.

\[
\Delta C = k(\bar{x} - C)(y_{\text{max}} - \bar{y}) \quad \text{(Equation 3.5)}
\]

As it is apparent from equation 3.5, if a physician perceives her decision outcome to be the best possible outcome \( (y_{\text{max}} = \bar{y}) \), she does not change her threshold.

We expect a physician’s interpretation of the results of her practice to depend on the availability of outcome feedback. In case, outcome feedback exists \( (F=1) \) she will perceive it, and if not \( (F=0) \), we assume she judges the outcome of her practice based on her confidence on her own practice \( (p) \). Equation 3.6 represents this condition.

\[
\bar{y} = y \quad \text{if } F = 1 \\
\bar{y} = p \quad \text{if } F = 0 
\quad \text{(Equation 3.6)}
\]

In our model, we assume under vaginal delivery, feedback is provided \( p(F=1) = 1 \), but under c-section, only some times feedback is provided. Mathematically, if a c-section surgery is performed we have \( p(F=1) = q < 1 \), that is in \( q \) portion of c-section surgeries feedback is provided.

**Gaining Practice Skills**

Finally, physicians can gain skills as they practice and accumulate experience. We assume they accumulate experience \( (E_i) \) with the rate of \( e \) for each practice while their forgetting time constant is \( \tau \). \( E_i \) is the experience of c-section and \( E_0 \) is the experience of vaginal delivery. Equations 3.7 and 3.8 represent experience and skill in our model.

\[
\Delta E_i = e - E_i / \tau \quad \text{if } d = 1 \\
\Delta E_0 = e - E_0 / \tau \quad \text{if } d = 0 
\quad \text{(Equation 3.7)}
\]
\[ S_d = g(E_d) \]  
(Equation 3.8)

where \( g \) should be defined in such a way that as physicians perform more practice they gain more skill in a declining rate, that is \( \dot{g}(.) \geq 0 \) and \( \ddot{g}(.) \leq 0 \). In the next step we simulate the model. The model parameterization is reported in Appendix 3.1. In short, the parameters are set in a scale that the optimal threshold to conduct c-section is on health risk of 0.5, and the distribution of health risk among patients is in a way that 20% of the patients have health risk above 0.5 and are better off to be treated by a c-section surgery. Scaling the parameters is hypothetical, and the results of the analysis are not qualitatively sensitive to those values. A Vensim snapshot of the model is shown in Appendix 3.2.

3.5. Simulations

Base Run

First we present the base run of the model. As stated, in our model, doctors are mathematically similar agents, and the only difference is in their random seeds. Each doctor visits one patient a week for 30 years, and makes a decision about the patient’s baby delivery (vaginal or c-section). During the years of practice, the doctors try to improve their practice performance through learning and accumulating experiences. Figure 3.2 shows the results, each graph representing one doctor. The x-axis is the time horizon, i.e., years of practice, and the y-axis is the threshold to make a c-section decision (how much health risk should be perceived to perform a c-section). Lower threshold would mean higher percentage of c-section deliveries, of which some would not be justified based on true risk. Therefore, each graph on Figure 3.2 shows how the threshold of a doctor changes during her practice. As we mentioned the model is parameterized in a
way that 5/10 is the optimal threshold (a patient whose health risk is above 0.5 is in
gen general better off with vaginal birth, vice versa). All physicians are assumed to start from
an identical condition with the initial threshold of 5/10.

As we see the thresholds of the doctors go below the optimal threshold meaning that
they perform c-sections more frequently than what is set in the model to be optimal.
Further, as we see in the Figure, threshold to perform a c-section is different across
different physicians, and they do not converge. In other words there is a variation of c-
section threshold.

![Figure 3.2: Threshold dynamics](image)

We can also look at the cross sectional synthetic data, output of the simulation model.
In the real world, not all doctors start practicing together. For example, in the year 2011,
some doctors are at their first year of practice, some at the second, and so on. To count
for the variability of starting time, we assign a random number to each of our doctors as their years of experience on year 2011, and read their threshold on time equal to that random number. Therefore, we have one observation from each doctor, and looking at the distribution of those observations we can represent how a cross sectional synthetic data looks like. Figure 3.3 shows the distribution and compares that with the optimal threshold.

Figure 3.3: distribution of practice threshold and the optimal threshold.

Note: the graph shows there is a variation in threshold and there is bias from optimal threshold in ways that physicians’ threshold to perform c-section is smaller than the optimal threshold

As we see in Figure 3.3, the average of the distribution of threshold is smaller than the optimal threshold, showing that there is a bias in c-section rate toward performing
more c-sections. In addition, the figure shows that there is a considerable variation across physicians. While there are physicians who set their thresholds around 3 (they perform c-section for patients with the health risk of 3/10), some set their thresholds around 5. It is important to mention that all of these physicians started from a similar initial condition including medial trainings and visited the same population of patients, and what was different across them was just a random seed that was influencing their errors and the order of the patients they visited. In short the model demonstrates that bias and variation can arise solely from experiential learning and in the absence of financial incentives or personality traits. Next we compare the results of the model with the data from Epstein and Nicholson (2009).

A Comparison of the Model Results with the Data

As mentioned, Epstein and Nicholson (2009) find a considerable level of practice variation in the states of New York and Florida. A comparison of our synthetic data from simulation with their data, although a weak test, can be interesting. We calculate the c-section rate for each data point and find the deviation from the mean for each data point, in a same way that is done in Epstein and Nicholson (2009). The results are shown in Figure 3.4. The resulted distribution is similar to Epstein and Nicholson’s (2009) figure, and has a 10.9 percentage point standard deviation fairly higher than Epstein and Nicholson’s observation (6.5 percentage point).
Figure 3.4: Distribution of deviation of c-section rate from mean. The standard deviation is 10.9 percentage points.

We believe although this practice is a weak test, but still the results are promising, and further detailed investigations can prove useful. There are a few parameters that need to be calibrated, and in such a case we can get a better fit. For example, the physicians’ confidence on their negative decisions (p) was unknown for us, and we used estimations from Elwin et al. (2007) and Ghaffarzadegan and Stewart (2011b) to set p. Those estimations were based on laboratory experiments and can be updated for our modeling purposes. Access to the detailed data on practice performance can be helpful.

**Behavioral Analysis**

We would like to further analyze the results in Figure 3.2 and find what parts of the structure result in practice variation and what parts are contributing to bias. This method
is known as partial model testing in the system dynamics literature (Morecroft, 1983; Sterman, 2000). We divide the model into two major sub-structures. The first sub-structure, which we call the skill sub-structure consists the model excluding the conditionality of feedback (providing full feedback after any decision whether it is vaginal or c-section surgery). The second sub-structure, which we call the conditional feedback sub-structure, is the entire model excluding the effects of skill ($\alpha=0$). These experiments can give us a clearer explanation about “why” we have the results shown in Figures 3.2-3.4 (bias and variation), and what are the effects of each sub-structure on the final behavior.

**Effects of Skill**

First, we focus on the effects of skill, and assume a full feedback condition (feedback is always provided and there is no effect from confidence). To conduct this experiment, we provide feedback to physicians, after any kind of practice, whether it is a c-section or a vaginal birth.

Figure 3.5 shows different simulation runs for different random seeds. Each line represents one doctor’s decision threshold for performing c-section. The deviation shows that physicians are diverging as they gain more experience and some are more likely to perform c-section surgery than others. The graph shows that the stated rules are adequate to generate disagreements across doctors, even if they have similar financial incentives and similar initial training and even if clear feedback is provided after practice.
Figure 3.5: Dynamics of decision threshold for performing c-section for different doctors when feedback is complete, but performance depends on experience and skill.

Figure 3.6 in another representation of the results and demonstrates the distribution of threshold for c-section in a cross sectional observation of the synthetic data. Each data point is one physician’s threshold to perform c-section in year 2011 considering the fact that not all of the physicians are in the same cohort. As we see, having only the skill substructure leads to a three-modal distribution, where most physicians are polarized at two extremes, some perform too many c-sections and some perform too few. There are also a group of physicians who are at the middle where most of them are in early years of practice, before being pushed to one of the two extremes.
Figure 3.6: The distribution of the synthetic *threshold to conduct a c-section* in a cross sectional observation of physicians when there is a full and clear feedback.

Figures 3.5 and 3.6 show that practice variation can emerge as results of the skill sub-structure even when there is no assumption about feedback asymmetries (feedback is clear and always provided). The simulated doctors’ disagreement on what is the best decision for a single patient is due to the effect of past decisions on experience, and the causal connections between experience, perception and next decisions. As doctors perform their practice, they gain more experience, however in an unbalanced way where some have more experience in c-section and some have more experience in vaginal delivery. The unbalanced experience affects their performance, their perception of how effective each style can be, and their next decisions.
It is also important to mention that the skill substructure does not create bias in practice. As we see in Figure 3.6 the average of the distribution is roughly around the optimal threshold.

**Effects of Conditionality of Feedback**

Now we focus on the second substructure and investigate the effect of conditionality in feedback on the final results. As it is shown in Figure 3.7, when skill has no effect on the physicians’ performance, their threshold do not diverge but decreases as they perform more practices, creating bias toward more c-section surgeries (i.e., lower threshold for c-section surgery).

![Graph showing dynamics of decision threshold for performing c-section for different doctors under the conditional feedback situation, controlled for the effects of experience and skill.](image)

Figure 3.7: Dynamics of decision threshold for performing c-section for different doctors under the conditional feedback situation, controlled for the effects of experience and skill.
Figure 3.8 shows a cross sectional demonstration of the synthetic data when we have different cohorts of physicians. As we see the result does not show a considerable variation, but a considerable bias from the optimal threshold. Asymmetries in feedback (more information about errors after vaginal deliveries, than information about errors after c-section surgeries) result in more correction forces to decrease the threshold. In turn, such unbalanced feedback causes physicians to perform more c-section surgeries than what is the optimal ratio.

Figure 3.8: The distribution of synthetic threshold to conduct a c-section in a cross sectional observation of physicians when we control for the effects of skill and the results are mainly driven by the conditionality of feedback.
In summery, the analysis shows that the skill sub-structure results in practice variation and the conditional feedback sub-structure results in bias in practice.

Finally, a comparison of Figures 3.8 and 3.6 with 3.4 is also interesting. As we see the bias in Figure 3.8 is less than the bias in Figure 3.4 when we have both effects (interactions between skill and conditional feedback). Further the magnitude of variation in Figure 3.6 is more than variation in Figure 3.4. In other words, when we have both effects of skill and effects of conditionality of feedback we have higher level of bias and lower level of variation. The main reason is that while conditionality of feedback triggers bias toward more c-section, such an effect in interaction with experience results in more unbalanced experience and skill toward c-section, increasing the bias toward more c-section.

Next we examine the model for a more generic case.

**Simulation for a Generic Case**

C-section is an illustrative example of bias and variation in medical decision making because like in many other domain, decisions are repetitive (frequent cases of patients), skill and experience are important on how medical decisions are performed, and there is an unbalanced outcome feedback. Many other medical domains have similar characteristics. However, in some domains there can be more (or less) elasticity to skill and some domains may provide more (or less) balanced outcome feedback. Having the mathematical model, we can generalize the arguments by changing the parameters and investigating the magnitude of bias and variation in different domains.
We define 10 (5×2) different hypothetical medical domains: 5 conditions on the elasticity of practice performance to skill (α= 0, 0.25, 0.5, 0.75, 1) times 2 feedback conditions (full feedback (FF), and conditional feedback (CF)). Although we have set these conditions hypothetically, they can represent different domains of medicine. For example, the condition of α=0 and CF can represent anti-biotic perception for otitis media in pediatric services. While a doctor should make a proper diagnosis and decision in prescribing anti-biotic, after the decision is made (the pills are prescribed) the doctor’s skill has no effect on how anti-biotic pills affect the patient. In contrast, in a c-section surgery a doctor should make a good decision, but also should perform it well where her skill to perform a c-section surgery comes to play. Another example, can be in dental health and the practice of root channel, where a doctor’s skill in performing a root channel is important (larger α), and feedback is conditional (usually a teeth with a removed nerve will not ache, even if the decision to remove the nerve was unnecessary). Practices that require frequent follow-ups and can control for the effects of decision can present a full feedback condition. For example a heart surgery can be considered as a case of a large α but a full feedback condition (FF), usually a physician will know if the surgery was performed well or not.

We simulated the model under the described 10 conditions (5 values of α times 2 feedback conditions). In each condition we have 100 agents (physicians). We then calculate the amount of bias and variance in each of those conditions. Figure 3.9 compares the magnitude of bias and variation across those 10 conditions.
Figure 3.9: Bias and variation under 10 different conditions of medical practices: 5 different values for elasticity to skill ($\alpha$) times 2 feedback conditions (CF: conditional feedback, FF: full feedback).

Note: As we see, some conditions have more variation than others and some have more bias. For example the bias in practice is greater for ($\alpha=0$, conditional feedback) than for ($\alpha=0$, full feedback), while the latter has more practice variation.

Figure 3.9 shows the magnitude of bias and variance under these conditions vary. In this figure, the x-axis is the magnitude of variation and the y-axis is magnitude of bias, and each point represents on the 10 hypothetical domains of medicine. So, if a domain is closer to the origin of the graph (0,0), then there is less bias and variation there. For example the bias in practice is greater for ($\alpha=0$, conditional feedback) than for ($\alpha=0$, full feedback).
feedback), while the latter shows more practice variation. The graph shows that under the full feedback conditions, larger elasticity to experience ($\alpha$) results in larger practice variation. It also shows that under the conditional feedback conditions, larger elasticity to experience ($\alpha$) results in larger bias in practice. Interestingly, in a constant level of elasticity to experience ($\alpha$), providing full feedback decreases bias but results in more variation (compare $\alpha=1$ and CF with $\alpha=1$ and FF). An empirical investigation of the findings can be interesting.

### 3.6. Conclusion

We develop a model of medical practice specifically tailored for obstetrics to study bias and variation in the practice of baby delivery when physicians should make a decision between a vaginal delivery and a c-section surgery for each of their patients. While the most common explanations for practice variation and bias are linked to patients characteristics, regional characteristics, physicians’ personality traits (e.g., the level of risk aversion) or their financial incentives, we offered a new theory of how medical task characteristics especially outcome feedback characteristics and the elasticity of the results to physicians’ skills can results in practice variation and bias.

Our simulation model controlled for all of the common previous explanations offered in the literature and still created practice variation and bias. In our model, we do not impose any regional variation to our agents, and the personality characteristics and financial incentives of the agents are the same. In addition, doctors are visiting similar populations of patients with the same level of health risk. Therefore, the simulation results show that practice variation and bias does not have to be caused by patients’
characteristics, regional characteristics, and physicians’ personality traits and financial incentives, but can endogenously emerge through daily practices even across physicians with totally similar characteristics. In other words, the structure of medical tasks, and specifically physician’s exposure to outcome feedback, and the experience accumulation processes through repetitive medical decisions can contribute to practice variation and bias.

We also analyzed the final results of the model through controlling different sub-structure of the model. The experiments revealed that accumulation of experience and skill can result in variation if one’s performance is highly depended on one’s skill. We also showed that the conditional feedback sub-structure drives bias in practice through forcing physicians to perform a kind of practice for which they receive less negative feedback. Further analyses revealed that the interactions of skill and conditionality of feedback exacerbate the bias, leading people to accumulate more experience on the kind of practice for which there is less negative feedback.

A preliminary validation of the results is achieved by comparing projected results with the actual data from cesarean section surgery in the states of New York and Florida. More empirical investigation is needed to test this dynamic theory.

The study contributes to the literatures of decision and policy sciences and medical decision making on different levels. First the study develops a new explanation for practice variation and bias in medicine. The new explanation is structurally different from the previous theories of practice variation and, therefore, has different policy implications. It is important to mention that this study does not reject any previous explanations offered for practice variation and bias in medicine, but it develops a new
coherent theory of sub-optimal decisions in medicine. In other words, the theory is a new layer to existing understandings of the factors that contribute to bias and variation in medicine. This new theory needs to be empirically investigated. Second, our study is one of the first ones to apply the concepts of experiential learning into the studies of medical decision making. We believe such an approach is important as doctors perform repetitive tasks where for some portion of them they receive incomplete feedback. Third, the study has methodological contributions as it is the only study which has modeled physicians’ decision making processes in a feedback-loop based approach. The model can be applied to a wide range of medical practices and can be used as a platform for further empirical investigations.

In short, we argue that practice variation and bias can dynamically emerge as physicians perform practices due to the learning characteristics of medical tasks. Such a structure can result in sub-optimal practices even if the financial incentives, the personality characteristics of the doctors, the regional characteristics, and the distribution of patients are totally similar.
References


Appendix 3.1: The Model Detailed Formulation

\[ \text{Beta} = 0.5 \]

"c-section skill (S2)" = \( g(\text{"experience of c-section (E2)"}/\text{"normal c-section exp"}) \)

"chng in threshold (DeltaC)" = IF THEN ELSE("outcome shortfall (y_{\text{max}}-y_{\text{bar}})"\rangle0,

\( (\text{"Perception of health risk (x-bar)" } - \text{"threshold for c-section (C)"}) \times \text{"threshold adjustment speed (k)"}, 0 ) \times \text{"outcome shortfall (y_{\text{max}}-y_{\text{bar}})"} \)

"confidence in c-section decisions (p)" = 0.95

"decision (d)" = IF THEN ELSE("Perception of health risk (x-bar)"\rangle"threshold for c-section (C)", 1, 0 )

"decision effectiveness (de)" = \( f(\text{abs}(\text{"decision (d)" } - \text{"health risk (x)"})/10)) \)

"experience of c-section (E2)" = \( \text{INTEG} (\text{"gaining experience of c-section" } - \text{"forgetting exp of c-section"}, 1) \)

"experience of natural birth (E1)" = \( \text{INTEG} (\text{gaining experience of NB} - \text{forgetting exp of NB}, 1) \)

\( f([(0,0)-(1,1)], (0,1), (0.25,0.95), (0.4,0.75), (0.5,0.5), (0.6,0.25), (0.75,0.05), (1,0)) \)

"forgetting exp of c-section" = "experience of c-section (E2)"/time to forget

forgetting exp of NB = "experience of natural birth (E1)"/time to forget

\( g([(0,0)-(2,1)], (0,0.2), (0.35,0.25), (0.5,0.6), (0.65,0.95), (1,1), (2,1)) \)

"gaining experience of c-section" = "decision (d)" \times \text{normal exp per practice}

gaining experience of NB = (1-"decision (d)") \times \text{normal exp per practice}
"health risk (x)"=RANDOM NORMAL(0, 10, mean health risk, Stdev Dist of Patients, seed)

"Initial threshold (C0)"= 5

"natural birth skill (S1)"= g("experience of natural birth (E1)"/normal nb exp)

"normal c-section exp"= 50

normal exp per practice= 1

normal nb exp = 50

"outcome feedback availability (F)"= IF THEN ELSE(switch to conditional feedback=0,1,IF THEN ELSE("decision (d)"=0,1, IF THEN ELSE(RANDOM UNIFORM(0,1 , seed)<"prob of getting feedback in CF on c-section (q)"",1,0)))

"outcome shortfall (ymax-ybar)"= Max("perception of the best possible outcome (ymax)"-"perception of practice outcome (ybar)",0)

"Perception of health risk (x-bar)"= Max(Min(RANDOM NORMAL(+"health risk (x)-10, 10+"health risk (x)" , "health risk (x)" , "Stdev (Error of Physicians)" , seed),10),0)

"perception of practice outcome (ybar)"= IF THEN ELSE("outcome feedback availability (F)"=1,"practice outcome (y)" , "confidence in c-section decisions (p)")

"perception of the best possible outcome (ymax)"= 1

"practice outcome (y)"="decision effectiveness (de)^Beta*"("Skill (Si)"^(1-Beta))

"prob of getting feedback in CF on c-section (q)"= 0.4

seed= 33
"Skill (Si)" = "decision (d)" \times "c-section skill (S2)" + (1 - "decision (d)") \times "natural birth skill (S1)"

switch to conditional feedback = 1

"threshold adjustment speed (k)" = 0.05

"threshold for c-section (C)" = \text{INTEG} ("chng in threshold (DeltaC)", "Initial threshold (C0)"

time to forget = 50

I = "decision (d)"

mean health risk = 3

O = positive decisions \times \text{IF THEN ELSE(} \text{integer(Time/time interval)}=\text{Time/time interval}, 1, 0) \times \text{IF THEN ELSE(O=0,0,positive decisions/time interval)}

positive decisions = \text{INTEG} (I-O, 0)

selection rate = \text{IF THEN ELSE(O=0,0,positive decisions/time interval)}

"Stdev (Error of Physicians)" = 2

\text{Stdev Dist of Patients} = 2

time interval = 50
Appendix 3.2: Snapshots of the model in Vensim

Threshold Model: This sector creates the endogenous learning model

Learning Model

Experience sector

Figure 3.A1: Threshold Model
Enviornmental Model: This sector only creates random patients, errors, and calculates the ratio of positive decisions.

1. Test the effects of skill: change "seed" and see how the behavior changes.
2. Test effects of conditional feedback: change Alpha to 0, and turn on the sw to cond. Feedback.
3. Complete model: mean health risk = 3, Alpha=0.5, SW to cond feedback = 1.