UNIVERSITY OF CALIFORNIA SAN DIEGO

Essays in Microeconomic Theory and Experimental Economics

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Economics

by

Erin Giffin

Committee in charge:

Professor James Andreoni, Co-Chair
Professor Isabel Trevino, Co-Chair
Professor Craig McKenzie
Professor Marta Serra-Garcia
Professor Joel Watson

2018

The dissertation of Erin Giffin is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____
Co-Chair

_____
Co-Chair

University of California San Diego

2018

DEDICATION

To my husband,

Matthew Nash,

for being a true partner throughout this journey.

To my parents,

David and Susan Giffin,

for their unwavering support and encouragement.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to thank my dissertation committee for their invaluable support and guidance throughout my graduate studies: James Andreoni, Isabel Trevino, Joel Watson, Craig McKenzie, and Marta Serra-Garcia. I would also like to thank James Andreoni, B. Douglas Bernheim, and Justin Rao for generously sharing their experimental data for use in Chapter 2 and Matthew Embrey, Guillaume Fréchette, and Sevgi Yuksel for sharing their experimental instructions, which were invaluable for the experimental design used in Chapter 3.

My most sincere gratitude to the following people for extensive comments and friendship: Alyssa Brown, Maya Duru, Elizabeth Hastings, Veena Jeevanandam Blume, Grant Johnson, Adeline Lo, Shanthi Manian, Héctor Pifarré i Arolas, Chelsea Swete, and Erin Wolcott. I would also like to thank Kate Antonovics, Julie Cullen, and seminar participants at UCSD, ESA North America 2017, and FISP Symposium 2017 for additional comments on this research.

Finally, I would like to thank my husband and my parents for supporting me at home so I could succeed at work.

I gratefully acknowledge funding provided by the Department of Economics and the Frontiers of Innovation Scholarship Program at UC San Diego.

Chapter 1, in full, is currently being prepared for submission for publication of the material. Giffin, Erin; Lillethun, Erik. The dissertation author is the co-author of this material.

Chapter 2, in full, is currently being prepared for submission for publication of the material. Giffin, Erin. The dissertation author is the sole author of this material.

Chapter 3, in full, is currently being prepared for submission for publication of the material. Giffin, Erin. The dissertation author is the sole author of this material.

VITA

2011	Bachelor of Arts in Economics and Psychology *summa cum laude*, University of Colorado, Boulder
2018	Doctor of Philosophy in Economics, University of California San Diego

ABSTRACT OF THE DISSERTATION

Essays in Microeconomic Theory and Experimental Economics

by

Erin Giffin

Doctor of Philosophy in Economics

University of California San Diego, 2018

Professor James Andreoni, Co-Chair
Professor Isabel Trevino, Co-Chair

Chapter 1 considers a theoretical model of evidence acquisition and disclosure in a legal setting. It analyzes how risk aversion affects an agent's willingness to seek out information and studies how legal rules can be used to balance the gains from evidence gathering with the costs of acquisition. Chapter 2 examines if different expectations about men and women's behavior, or stereotypes, could be responsible for observed gender differences, even in anonymous laboratory settings, and shows experimental evidence consistent with this model's predictions. Chapter 3 experimentally examines players' recall of past play in a canonical economic game.

Chapter 1

Incentives for Evidence Acquisition and Disclosure

1.1 Introduction

In civil trials, a judge or jury uses the evidence acquired and presented by biased litigants to make the best decision possible. The quantity of evidence acquired can affect both the fairness of restitution and the strength of disincentives for law breaking. However, the risk preferences of the litigant play a crucial role in deciding how much effort the litigant expends to gather evidence. Consider the example of a small, self-published author of a copyrighted work suing the author of a derivative work. The copyright holder can work to find evidence of the other author profiting from the derivative work, each instance of which would increase the magnitude of the damages. However, the copyright holder may fail to discover any new evidence, in which case the effort is wasted. Even worse, if the court expects that the copyright holder searched extensively for evidence, a lack of evidence presented at trial could result in a negative inference on the part of the court. A very risk-averse copyright holder will not value the extra benefit from higher damages enough to justify the cost of evidence gathering and the danger of sometimes inducing a stronger negative inference. In this way, risk aversion can work against evidence acquisition incentives. In particular, we can imagine that a self-published author would be more risk averse than a large publishing house and would therefore gather less evidence to present at

trial. As a result, not only would restitution for the self-published author be less accurate but the deterrent effect of the laws would be less powerful.

The main problem in the example is that the copyright holder does not reap enough of the benefits of evidence acquisition. After the evidence has been presented to the court, the judge or jury ignores the copyright holder's preferences and simply maximizes the social objective function given the information at hand. However, if the judge or jury were constrained by law in the right way, say by having to give a bonus award for any presented evidence, then the problem of low evidence acquisition by very risk-averse parties could be remedied. Designing these constraints in a socially optimal way requires rules treating very risk-averse litigants differently from nearly risk-neutral ones. This is a controversial idea, as in many countries the rules of the legal process are nearly identical for different types of litigants. For example, the 5^{th} and 14^{th} Amendments to the U.S. Constitution guarantee due process and equal protection rights. Taken together, these rights can be interpreted as preventing different treatment of litigants who are in similar legal circumstances. However, this interpretation is not universal, and there is precedent for laws treating individuals differently from corporations (e.g., 197 (1973)).

The example demonstrates that if litigants with different degrees of risk aversion are treated the same in court, they will exert different effort levels towards acquiring evidence. As a result, these litigants will have different probability distributions of trial outcomes, even when the underlying facts of their cases are identical. In particular, rulings for very risk-averse litigants will be based on little information. Therefore, if the goal is to have similar trial outcomes (and therefore similar restitution and deterrent effects), there must be some departure from nominal equal treatment in the procedures governing trials.

In this paper, we analyze a setting where one or more litigants exert costly effort to try to acquire evidence. Evidence can then be revealed or concealed at trial to influence a judge or jury, who may or may not observe the effort. One key novelty of our model is that

we allow the litigants to have varying degrees of risk aversion (e.g., big corporations vs. private individuals). The first main question we answer is how litigant risk preferences affect equilibrium evidence acquisition and trial outcomes. The court's decision depends on the evidence submitted at trial. If no one acquires evidence, then only one decision is possible: the optimal decision in the state of ignorance. If someone acquires and presents evidence, then the court's decision depends on what that evidence reveals, so multiple different decisions may result. Hence, the act of acquiring and revealing evidence is inherently riskier, because it results in more possible outcomes than not acquiring evidence at all. This does not necessarily mean that risk-averse parties gather no evidence, as acquisition and strategic revelation can still make the expected decision of the court more favorable. However, it does introduce a tradeoff between the "bias motivation"–wanting more evidence to bias the decision favorably–and the "risk motivation"–wanting less evidence, because evidence causes spreads in the resulting decision distribution. We find that high degrees of risk aversion result in poor incentives for evidence gathering, which causes the court to make worse decisions on average. Furthermore, evidence gathering disappears entirely in the infinitely risk-averse limit, as does the effect of observability of effort. Therefore, looking at this risk preference channel in isolation, civil law produces more accurate restitution and a stronger deterrent effect for large corporations than it does for private individuals.

We then address the question of what socially optimal legal rules look like and how they depend on litigant risk preferences. A social planner may commit to rules that constrain the court's decision making flexibility in order to more efficiently balance evidence acquisition incentives against the costs of acquiring evidence. We model this by restricting the allowed mappings from submitted evidence to court judgments (i.e., restricting the decision rules). When the planner may choose the court's decision rule in its entirety (full commitment), the form of the optimal rule depends on the weight the planner places on evidence acquisition costs. When this weight is low, there is "overincentivization," meaning

that omitted evidence results in a more punitive action than without commitment, and revealed evidence leads to a more rewarding action than without commitment. When the cost weight is high, there is "underincentivization," which features weaker punishments and smaller rewards than without commitment. Moreover, an overincentivization structure is always socially optimal for a sufficiently risk-averse litigant. We show that qualitatively similar results hold for a variety of more specialized design settings that correspond to common legal rules, such as admissibility of evidence, maximum penalties, and minimum penalties. In this way, social welfare could be improved by having different rules for different risk types of litigants, such as individuals and corporations.

This paper contributes to the law and economics literature by examining evidence acquisition decisions. Much of the existing literature abstracts away from the evidence-gathering step of the court process. Parties do not choose if or how to acquire evidence, but are exogenously endowed with evidence and then their only decision is whether to disclose it (e.g., Lewis and Poitevin (1997), Bull and Watson (2004), Bull (2008), and Demougin and Fluet (2008)). Some papers add the discovery process, by which evidence can be obtained from the opposing side (Cooter and Rubinfeld (1994) and Hay (1994)). In this paper, we add an initial evidence acquisition stage where parties choose how hard to work at gathering new evidence prior to discovery. Adding this phase to the model provides two key contributions to the existing literature. It first enriches the theoretical models in this area, and it helps illuminate how various laws or institutional rules could be used to manipulate evidence acquisition incentives to increase social welfare.

There exists a substantial literature in which decision makers rely on the hard evidence presented by biased parties (e.g., Milgrom and Roberts (1986) and Shin (1994)). We extend Shin's model by allowing the informed party to change the probability of learning the state by choosing a costly evidence acquisition effort, much like Henry (2009), Kim (2013), and Wong and Yang (2015). Also related are Daughety and Reinganum (2000) and

Froeb and Kobayashi (1996) in which parties acquire a body of evidence and reveal only their best pieces of evidence. Our contributions relative to these latter papers are our tying of risk preferences and legal rule design to evidence acquisition incentives.

Our paper features a mechanism design problem where the mechanism constrains the allowable decision rules. Within the field of law and economics, the relevant literature is that on legal rulemaking (e.g., Ehrlich and Posner (1974), Macey (1994), and Davis (1994)), where rules curtail the court's flexibility in handing out judgments. However, the focus in these papers is on preventing judicial error and counteracting judicial bias, which are not factors in our model. More closely related are Sanchirico (1997), Persico (2012), and Lester et al. (2009). In the first of these, the optimal decision rule exhibits a burden of proof to discourage low expected recovery cases from being brought to trial. In the latter two, it may be socially optimal to rule out some types of evidence to economize on evidence costs or focus jury attention. Stephenson (2008) has findings similar to ours, that committing to rules that make ex-post inferior decisions can be beneficial for encouraging evidence production. To this most similar strand of literature, we contribute a more general model of legal rule design along with the analysis of how risk preferences influence the optimal rules.

Although this paper formally models trials, it is compatible with the large body of literature that finds that settlement should usually be reached before going to trial, especially with risk-averse litigants (e.g., Landes (1971), Gould (1973), and Shavell (1982)). Producing favorable evidence at settlement negotiations is crucial to obtaining a good settlement, because the parties infer that this evidence would result in a better outcome for that side at trial. Therefore, policies which manipulate trial outcomes also have a similar effect on settlement outcomes.

The paper proceeds as follows. section 1.2 analyzes a simplified version of the model to serve as an example. section 1.3 formally introduces the more general model. section 1.4

contains the results on the observability of effort. section 1.5 presents one of the key results, showing that high risk aversion decreases equilibrium evidence acquisition. In section 1.6, we analyze several different types of legal rule design settings, showing how the forces identified in section 1.5 influence optimal legal rules. section 1.7 concludes.

1.2 Example

1.2.1 Model

In this example, there is a sender (the litigant) and a receiver (the judge).[1] We will refer to the sender as "she" and the receiver as "he". There is an unknown state $\omega \in [0,1]$. The players both have a uniform prior. The state represents the relevant underlying facts of the case. The state may be interpreted in several ways. For example, the state could represent the degree of liability, or it could be some measure of the belief of liability. If "liable" is represented by 1 and "not liable" is represented by 0, ω can be interpreted as the probability that the defendant is liable.

First, the sender may acquire evidence about the state as follows: The sender exerts "effort" $p \in [0,1]$ to try to acquire evidence. Then, she receives a private signal with realization $x \in [0,1] \cup \{\phi\}$. With probability p, she finds evidence proving the true state ($x = \omega$). With probability $1-p$, she finds no evidence ($x = \phi$).

Consider the play of the game after the signal arrival. First, the receiver observes the sender's effort with probability $q \in \{0,1\}$. Whether or not the receiver observes the effort is public information (e.g., it is based on something revealed in court). Then, the sender must either reveal to the receiver what she knows ($M = x$) or conceal it ($M = \phi$). Note that the receiver cannot distinguish between concealing a known state and revealing ignorance of the state. The receiver then observes M and chooses an action $a \in [0,1]$.

[1] A second competing litigant does not have much of an impact on the key results in this paper, so it is omitted for simplicity. For a more formal argument, see subsection 1.8.12

The sender only cares about the receiver's action, so her payoff function is $u^S(a)$ (for this example, there is no cost of effort). The sender always wants higher actions (e.g., a plaintiff always wants greater damages) and is risk-averse, so $u^{S'}(a) > 0, \forall a$, and $u^{S''}(a) < 0, \forall a$. The receiver cares about his action and the state. For this example, the receiver's payoff is quadratic loss: $u^R(\omega, a) = -(a - \omega)^2$. With these preferences, the receiver always best responds by choosing an action equal to the expected state given his information.

1.2.2 Beliefs

The receiver's actions depend entirely on beliefs, so it is worth focusing on those beliefs in isolation. Since a message $M = \omega$ means that the sender actually observed ω as the true state, the updated beliefs place probability 1 on ω. Therefore, the best response action must be $a = \omega$. Now, suppose that $M = \phi$. The sender's messaging strategy is always a threshold strategy (i.e., reveal for high states, conceal for low ones) with some threshold $\overline{\omega}$. After observing $M = \phi$, the receiver's posterior density is the following:

$$f(\omega | p, M = \phi) = \begin{cases} \frac{(1-p)}{(1-p)+p \cdot \overline{\omega}} & \text{if } \omega > \overline{\omega} \\ \frac{1}{(1-p)+p \cdot \overline{\omega}} & \text{if } \omega \leq \overline{\omega} \end{cases}$$

If the receiver observed the sender's effort, then p in the above expression is the observed effort. Otherwise, p is the believed effort level. After observing $M = \phi$, the receiver shifts his belief towards states that would have been concealed had the sender observed them. Moreover, this shift is increasing in p, as an increase in p means that concealment is relatively more likely than ignorance. Taking expectations, the receiver's best response action is as follows:

$$a = \mathbb{E}[\omega | p, M = \phi] = \frac{1}{2}\left[\frac{(1-p) + p \cdot \overline{\omega}^2}{(1-p) + p \cdot \overline{\omega}}\right]$$

$M = \phi$ always results in an action below $\frac{1}{2}$ (the ex-ante optimal action). However, when concealment is common ($\overline{\omega}$ is large), this difference is small.

1.2.3 Equilibrium

First, we focus on the case where the receiver always observes effort ($q = 1$). The concealment threshold $\overline{\omega}$ must be the state that makes the sender indifferent between revealing and concealing. Therefore,

$$\begin{aligned} \overline{\omega} &= \frac{1}{2}\left[\frac{(1-p) + p \cdot \overline{\omega}^2}{(1-p) + p \cdot \overline{\omega}}\right] \\ \Rightarrow \overline{\omega} &= \frac{\sqrt{1-p} - (1-p)}{p} \end{aligned}$$

Figure 1.1 shows the threshold as a function of the probability of learning the state. It is strictly decreasing and ranges from $\frac{1}{2}$ at the greatest (this is the action that the prior induces) and 0 at the lowest (there is no equilibrium concealment when the sender has complete information). This is a first sign that the sender might not desire the maximal effort level even if it is costless. Maximal effort implies no concealment, so the action always matches the state, even when the state is very low. Moreover, since the threshold $\overline{\omega}$ is the worst possible outcome for the sender, this worst-case outcome is decreasing as a function of the chosen effort level. Since an infinitely risk-averse sender only cares about the worst-case outcome, such a sender would not acquire evidence at all in equilibrium,

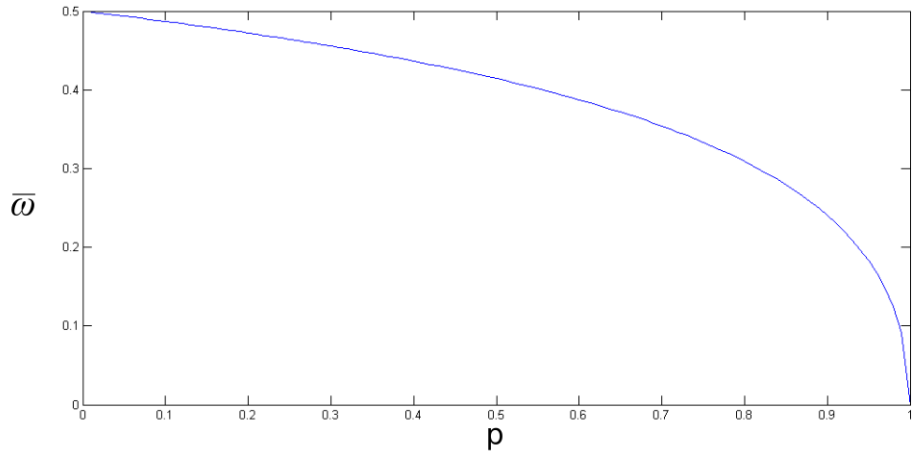


Figure 1.1: Concealment threshold as a function of effort

even when it costs nothing.[2]

In the case where effort is never observed ($q = 0$), deviating upwards from the effort that the receiver believes the sender to exert does not alter the receiver's beliefs or action following message $M = \phi$. Again, this is the worst-case outcome. Since increasing the effort does not improve the worst-case outcome, an infinitely risk-averse sender never strictly benefits from exerting effort, even when effort is costless. If effort has even minimal costs, the infinitely risk-averse sender would acquire no evidence in equilibrium.

Zero evidence acquisition in the case of an infinitely risk-averse sender is a bad outcome for the receiver. The receiver always chooses $a = \frac{1}{2}$, and the receiver's expected payoff is $-\frac{1}{12}$. What if society tries to overcome this problem by committing the receiver in advance to a maximum punishment for revealed evidence? For example, say the receiver cannot choose an action lower than $\frac{1}{2}$ if the sender provides evidence of the state to the receiver. Then the sender prefers maximal effort, because $p = 1$ and full revelation yields a payoff greater than $u^S(\frac{1}{2})$. Maximal effort then gives the receiver an expected payoff of $\int_0^{\frac{1}{2}} -(\omega - \frac{1}{2})^2 \cdot d\omega = -\frac{1}{24}$. This stylized example demonstrates that committing

[2]The infinitely risk-averse sender is an extreme case, however, we later show that a similar result holds for high degrees of risk aversion.

to decision rules in advance can improve the receiver's outcomes by providing stronger evidence acquisition incentives, and this improvement can be dramatic in cases of high risk aversion.

1.3 Model

While the preceding example illustrates the main qualitative points of the paper, there are several unrealistic assumptions: evidence acquisition is costless, the prior is uniform, effort is either always or never observed, the receiver has a quadratic loss utility function, and the sender is infinitely risk-averse. We now relax all of these assumptions in our more general model. Later, we will show (amongst other things) that the main qualitative points of the example still hold in the general model.

There are two players, a sender and a receiver also referred to as "she" and "he," respectively (the presence of multiple competing senders does not significantly change our results; a formal analysis of this can be found in subsection 1.8.12). The sender represents a litigant, and the receiver represents the relevant decision maker, such as a judge or jury. There is an unknown state $\omega \in \Omega = [\omega_0, \omega_1]$. W is the random variable with support Ω reflecting uncertainty about the state. The marginal prior belief of W is common to both players,[3] and it has a continuous probability density function (PDF) f. The corresponding cumulative distribution function (CDF) is denoted by F.

The game proceeds according to the following timeline: The sender chooses an evidence acquisition effort. The sender and receiver observe a public signal of this effort. The sender receives evidence. The sender sends a message to the receiver. Finally, the receiver takes an action $a \in \mathcal{A} \subseteq \mathbb{R}$, where $\mathcal{A}$ is convex.

[3]The common prior assumption is made only for simplicity of notation. Indeed, the key results do not even depend on the players knowing each others' priors. Private information complicates things quite a bit, although there is a certain sense in which the same main results hold. Results from a model with private information are presented in subsection 1.8.13.

The model of evidence is exactly the same as in the example. Formally, the sender's signal is $x \in X \equiv \Omega \cup \{\phi\}$. Observing $x \in \Omega$ proves that the true state equals x. Observing $x = \phi$ yields no useful information (observing it does not cause the prior to be updated). In order to control the probability of acquiring the evidence, the sender chooses an effort $p \in [0,1]$.[4]

After choosing an effort, the receiver then has an exogenous chance of observing the sender's effort. Let $q \in [0,1]$ be the probability that the receiver observes the sender's effort exactly (where otherwise the beliefs about effort must be based solely on equilibrium effort). In other words, the receiver observes a signal $y \in Y \equiv [0,1] \cup \{\psi\}$, where $y \in [0,1]$ is a perfect signal of effort, and $y = \psi$ means the receiver did not observe effort. We assume that the signal is public, so the sender always knows if the receiver observed her effort, but the other case works similarly for sufficiently high or low values of q. Since one of our motivations is to provide a robustness check for the extreme cases of $q = 0$ and $q = 1$, this assumption is innocuous.

Next, the sender observes signal x. Then, she may choose any $M \in X$ such that $x = \phi \Rightarrow M = \phi$ and $x = \omega \Rightarrow M \in \{\omega, \phi\}$ (so the sender can withhold evidence). M is called the "message", which is then observed by the receiver. The sender may choose a mixed messaging strategy $\sigma^S : X \times Y \to \Delta X$. Since the pieces of evidence in the message are costlessly verifiable, the receiver's updated beliefs will reflect the knowledge contained therein, implying that the receiver's posterior beliefs will place probability 1 on ω if $M = \omega$. Posterior beliefs following $M = \phi$ will depend on the prior, what the receiver observes about the effort, and the receiver's belief about the sender's strategy.

The sender's utility function is $u^S(a) - c(p)$. $u^S(a)$ is strictly increasing and bounded above. As an example, the sender can be a plaintiff, and the action can be the magnitude or

[4]The main insight of the paper, that high risk aversion depresses equilibrium evidence acquisition, is robust to a much more general model of evidence (not formally included in the paper). In particular, it is robust to a model where "effort" is a distribution over partitions of the state space, and the evidence proves that the state lies in the relevant partition element (not necessarily a singleton, as it is here).

probability of damages awarded. Curvature properties of u^S will eventually be important, but these will be introduced when they are needed. $c(p)$ represents the cost of evidence acquisition. Since greater acquisition effort is more costly, assume $c'(p) > 0, \forall p > 0$. Also assume that $c(0) = 0$, $c'(0) = 0$ and $c''(p) > 0$.

The receiver's utility function is $u^R(\omega, a)$, which is bounded above, twice continuously differentiable, and concave in a. Assume that for any belief G, $argmax_a \mathbb{E}_G[u^R(\omega, a)]$ is a singleton. This implies that the receiver's beliefs uniquely pin down a best response action. Furthermore, assume u^R is strictly supermodular, so that higher states warrant higher actions. For example, if ω is the degree of harm and a is the size of the damages, the supermodularity assumption holds. Define $a(E) \equiv argmax_{a \in \mathbb{R}} \mathbb{E}_F[u^R(\omega, a) | \omega \in E]$. This is the optimal action for beliefs where the receiver only knows that the true state lies in E. As a shorthand, $a(\omega)$ is taken to mean $a(\{\omega\})$. Strict supermodularity implies that $a(\omega)$ is a strictly increasing function of ω. Also assume that $\{a(\omega) | \omega \in \Omega\} \subset \mathcal{A}$, which implies that the receiver does not face binding constraints.[5] Define $a^* \equiv a(\Omega)$, which is the best response action in the absence of any additional information. Let $\mu : X \times Y \to \Delta\Omega$ be the receiver's belief system, i.e., it determines updated beliefs following any message and effort observation. A receiver's strategy is a mapping $\sigma^R : X \times Y \to \mathcal{A}$.

The supermodularity assumption yields some useful monotonicity properties with respect to optimal actions and beliefs. Firstly, it ensures that the sender must have a threshold state $\overline{\omega}$ below which she always conceals the state and above which she always reveals it (i.e., it is a threshold strategy). Secondly, it implies that empty messages result in low actions in the following way: Let $\mu(\phi, p; \overline{\omega})$ be the posterior belief induced by $M = \phi$ when effort is known to be p and the sender conceals if and only if $\omega \leq \overline{\omega}$. Then for any $\overline{\omega} \in int(\Omega)$, $argmax_a \mathbb{E}_{\mu(\phi, p; \overline{\omega})}[u^R(\omega, a)] < a^*$. This inequality holds for any supermodular

[5] The section on commitment can be seen as imposing binding constraints on the receiver by creating institutional rules. However, in the absence of institutional rules, the receiver is sufficiently free to make decisions that the action constraints are not binding.

utility function, which includes the quadratic loss utility from the example as a special case. Intuitively, this condition should usually hold, because the possibility that the sender is concealing low states means that the receiver should respond negatively to the absence of evidence.

Throughout the paper, the equilibrium concept used is similar to weak perfect Bayesian equilibrium (WPBE) but with a straightforward sequential rationality assumption. Proper subgames begin when the receiver observes the sender's effort, because then both parties have the same information. Let Γ be the overall game, and let $\Gamma(y), \forall y \in [0,1]$ be the game of incomplete information initiated when the receiver observes effort y.

Definition 1. *An **equilibrium** of the game Γ is a tuple $(p, \sigma^S, \sigma^R, \mu)$ such that*

1. *μ updates according to Bayes' Rule in Γ given (p, σ^S) whenever possible. For every $\hat{y} \in [0,1] \setminus \{p\}$, $\mu_{\hat{y}}(x) \equiv \mu(x, \hat{y})$ updates according to Bayes' Rule in $\Gamma(\hat{y})$ given $(\hat{y}, \sigma^S)$ whenever possible.*
2. *(p, σ^S) is a best response to σ^R given μ in Γ and $\Gamma(y), \forall y \in [0,1]$.*
3. *σ^R is a best response to (p, σ^S) given μ in Γ and $\Gamma(y), \forall y \in [0,1]$.*

With this definition in hand, proving the existence of an equilibrium is possible.

Proposition 1. *There exists an equilibrium of Γ.*

Proof. See subsection 1.8.1 □

1.4 Observability of Effort

In this section, we analyze the role of observability of effort on the outcome of the game. The results of this section add robustness to similar ideas found in Henry (2009) and Wong and Yang (2015), where q is restricted to $\{0,1\}$. When effort is close to unobservable

(i.e., q is close to 0), there is evidence acquisition in equilibrium, as the next proposition will show.

Proposition 2. *There exists $\underline{q} > 0$ such that for all $q < \underline{q}$, every equilibrium exhibits positive evidence acquisition. That is, $p > 0$.*

Proof. See subsection 1.8.2

□

Now, we consider the case where the receiver almost always observes the sender's effort (i.e., $q \approx 1$). For this section, we assume that the receiver has a quadratic loss utility function: $u^R(\omega, a) = -(\omega - a)^2$. Note that maximizing expected utility means that the receiver always chooses a equal to the expected state based on the posterior belief.[6] With quadratic loss preferences, we find that observability combined with risk-aversion eliminates all equilibria with positive evidence gathering.

Proposition 3. *If the sender is strictly risk-averse, and $u^R(\omega, a) = -(\omega - a)^2$, then for every $\overline{p} > 0$, there exists $\overline{q}$ such that for all $q > \overline{q}$, every equilibrium features $p < \overline{p}$. In particular, if $q = 1$, the unique equilibrium features $p = 0$.*

Proof. See subsection 1.8.3.

□

If we generalize away from a receiver with quadratic-loss preferences and instead allow u^R to be any function satisfying the assumptions in section 1.3, it is no longer the case that the observability of effort eliminates evidence acquisition. For example, the receiver may have some bias that depends on asymmetries in his beliefs. Since evidence

[6] There may be a constant bias without changing the conclusion of 3. For example, $u^R(\omega, a) = -(a - \omega - b)^2$ introduces a constant bias of b. However, when the bias depends on the asymmetries in the posterior belief, the rather extreme result of 3 does not necessarily follow. However, the results of section 1.5 still hold.

acquisition with strategic disclosure can manipulate the shape of the receiver's posterior after observing $M = \phi$, acquisition might be worthwhile.

However, we can prove a monotonicity result that indicates that observability reduces equilibrium evidence acquisition. For comparing sets of equilibrium efforts (because equilibrium is not necessarily unique), we use the strict weak set order. Define $\succ_w$ by the following: for any two sets A and B, $A \succ_w B$ if and only if for every $a \in A$ there exists $b \in B$ such that $a > b$ and for every $b \in B$, there exists $a \in A$ such that $a > b$. For any q, let $P(q)$ be the corresponding set of equilibrium effort levels. The following proposition shows that equilibrium effort is decreasing in the probability of observing effort:

Proposition 4. *Suppose $u^R(\omega, a) = -(\omega - a)^2$. Then for any $q' > q$, either $P(q) = P(q') = \{0\}$ or $P(q) = P(q') = \{1\}$ or $P(q) \succ_w P(q')$.*

Proof. See subsection 1.8.4.

□

1.5 Comparative Risk Aversion

In this section, we present our key result, which shows that if one sender is much more risk-averse than another, the former will acquire less evidence in equilibrium than the latter. Let the sender's utility be $u^S(a; r)$, where $r \geq 0$ is a risk aversion parameter.[7] Define the measure of absolute risk aversion at a as $R(u^S, a; r) \equiv -\frac{u^{S\prime\prime}(a;r)}{u^{S\prime}(a;r)}$. Assume that for all a, $R(u^S, a; 0) = 0$, $R(u^S, a; r)$ is increasing in r, and $R(u^S, a; r) \to \infty$ as $r \to \infty$ (that is, the sender ranges from risk-netural to infinitely risk-averse). Assume that $\lim_{r\to\infty} u^S(a; r)$ exists as a function to $\mathbb{R} \cup \{-\infty\}$ and is right continuous, which rules out a pathological

[7] Technically, the cost term of the Sender's utility function could change in conjunction with u^S as r changes. However, changes to the cost function will directly affect evidence acquisition incentives in a way that has nothing to do with risk preferences. Therefore, to isolate the effect of risk preferences, we make r a parameter of u^S only.

case. As an example, the constant absolute risk aversion (CARA) and constant relative risk aversion (CRRA) families of utility functions satisfy these assumptions.

Let every aspect of the game remain fixed except for r, so we can refer to "the game induced by r" and look at what happens to the set of equilibria as r changes. We can produce the following result:

Proposition 5. *For every $\overline{p} > 0$, there exists $\overline{r}$ such that for all $r \geq \overline{r}$, every equilibrium of the game induced by r features $p < \overline{p}$.*

Proof. See subsection 1.8.6.

□

Intuitively, this result follows from a very risk-averse sender's tendency to ignore good possible outcomes and focus on bad possibilities. When the receiver never observes the sender's effort, deviating to higher effort increases the probability of good outcomes (when acquisition is successful, and the state is high). However, if acquisition fails or the state is too low, the outcome is equal to the lowest outcome without the deviation. Since the lowest possible outcome is unchanged, these deviations are less appealing for more risk-averse senders. When the receiver always observes the sender's effort, an upward effort deviation results in an even lower worst outcome than without the deviation, so the effect of risk aversion is even stronger.

This result is robust to several modifications of the model, which can be found in the Appendix. subsection 1.8.12 shows that when there are two competing litigants, high degrees of risk aversion still reduce equilibrium evidence acquisition. subsection 1.8.13 analyzes a variation of the model where the litigant has private information and signaling via effort level is possible. In that section, 14 shows that there are two competing forces. One is the same force identified in this section, that high risk aversion depresses equilibrium evidence acquisition. There is also a pure signaling force that works in the opposite direction.

However, this latter countervailing force is only powerful when effort is easily observable (which is unlikely in the legal setting) and when the receiver can make inferences based directly on effort (which is of dubious legality given due process protections, since it is not evidence-based; see Friendly (1975)).

1.6 Commitment

In this section, we look at the problem of a social planner who can constrain the receiver's actions by establishing rules. Although these rules, such as evidence admissibility thresholds, force the receiver to make worse decisions given the evidence presented, they can encourage evidence acquisition enough to improve overall outcomes. Any receiver's strategy consists of a mapping $a(M)$ from messages to actions.[8] The social planner constrains the set of allowed mappings $a(M)$. These constraints commit the receiver to a set of decision rules. This commitment gives the planner some indirect control over the evidence gathering incentives of the sender. The control afforded by commitment gives the planner a tradeoff between incentivizing an ideal effort level and inducing good decisions based on the evidence presented.

Moreover, the sender's risk preferences can influence the optimal commitment policy. The best incentives for a very risk-averse sender may differ from those for a nearly risk-neutral sender, and this difference may be distinct for different types of commitment. These results will have implications for optimal legal rules for different types of disputants (e.g., civil vs. criminal trials, individuals vs. corporations, prosecution vs. defense, etc.).

[8] We are assuming that rules established by the planner treat different effort levels the same. Any effort level can be implemented by simply committing to a sufficiently severe punishment if an effort level other than the intended one is observed. However, in most settings, including the legal setting here, committing to different policies based on something as subtle as an effort observation seems impractical. On the other hand, committing to rewards and punishments depending on which pieces of hard evidence the sender submits is much more feasible in practice.

1.6.1 Full Commitment

Suppose the planner has the power to fully commit the receiver to a single decision rule $a(M)$. The sender observes the decision rule before making her effort decision. Although this setting is a bit unrealistic and extreme, it clearly demonstrates several key properties of socially optimal rules that will carry over into the more realistic settings analyzed later in this section.

Effort $p = 1$ can be implemented with full evidence revelation and no distortion of actions from their uncommitted socially optimal levels. This can be done by setting $a(\phi)$ so that $u^S(a(\phi)) < u^S(a(\omega_0)) - c(1)$. There cannot be $p < 1$ in equilibrium, since deviation to $p = 1$ would yield a guaranteed strictly higher payoff. Furthermore, there must be full evidence revelation, because concealing always involves a strictly lower payoff than revealing. Since the message $M = \phi$ is never sent in this equilibrium, all of the actions on the equilibrium path of play are at the uncommitted optimal levels.

However, this is an unlikely optimal policy in practice. This policy's optimality relies on the planner not caring much about the costs of the acquisition and presentation of evidence. Evidence acquisition for its own sake (i.e., without imparting better information to the receiver) is inefficient, so if the improvements in information transmission are not sufficiently valuable, the planner may prefer effort levels below 1. However, for any $p < 1$, disproportionately large punishments following $M = \phi$ are bad for the planner, since they would be executed with positive probability. Therefore, the optimal commitment policy in this setting is a bit more subtle and requires further analysis.

Example

For an illustrative example, take the model from section 1.2 but with sender costs of $c(p) = \frac{1}{2} \cdot \alpha \cdot p^2$. Let the social planner's payoff from state ω, action a, and effort p be

$-(\omega - a)^2 - \lambda \cdot c(p)$. The planner's cost weight $\lambda > 0$ reflects how much the planner cares about the sender's costs of evidence acquisition, or more generally the overall social burden of evidence acquisition.[9] Hypothetically, if there were full disclosure of evidence, and the receiver always acted optimally according to the posterior beliefs, the planner's expected utility (accounting for costs) would be $(1-p) \cdot \int_\Omega -(\omega - \mathbb{E}_F[W])^2 \cdot dF(\omega) - \lambda \cdot \frac{1}{2} \cdot \alpha \cdot p^2 = -(1-p) \cdot Var_F(W) - \lambda \cdot \frac{1}{2} \cdot \alpha \cdot p^2$. This is a strictly concave function of p, so a solution to the first-order condition (if one exists and lies in $[0,1]$) is the planner's ideal effort level. The F.O.C. yields $p = \frac{1}{\lambda \cdot \alpha} \cdot Var_F(W)$. This is always weakly positive, and it is often less than 1. Since the prior is $U(0,1)$, an assumption of $\alpha = \lambda = 1$ yields an optimal effort $p = \frac{1}{12}$. However, for the sender to always reveal evidence, the punishment action must be less than or equal to 0, which cannot be an optimal action given the receiver's beliefs. One of three things must be sacrificed: either a suboptimal effort is induced, not all of the evidence is revealed, or the actions are suboptimal given the receiver's posterior. We will show that the socially optimal policy generally features a combination of all three.

Model

We now construct the full commitment model. The model is mostly the same as in section 1.3, with only a couple modifications. Assume $\Omega = \mathcal{A} = \mathbb{R}$. The sender has utility function $u^S(a; r)$, which belongs to the type of parametric family from section 1.5. The receiver has quadratic loss utility over states and actions. In order to conveniently parameterize the marginal costs, we will assume that $c(p) = \frac{1}{2} \cdot \alpha \cdot p^2$, where $\alpha > 0$. However, the exact functional form is not crucial to the analysis. The planner's utility function is $-(\omega - a)^2 - \lambda \cdot \frac{1}{2} \cdot \alpha \cdot p^2$, where $\lambda > 0$.[10]

[9] Examples of negative consequences of evidence acquisition include the costs of workers responding to information requests, the time taken from witnesses and experts, and the opportunity costs of diverting physical pieces of evidence from their valuable normal uses.

[10] All of the commitment models in this paper have the planner internalizing the sender(s) costs to some degree. There is another important model of costs: a model of court costs. That model views effort as

At the very beginning of the game, the planner commits to a_ϕ (the "default" action taken in response to $M = \phi$), a threshold $\overline{\omega}$ below which evidence is inadmissible, and $a(\omega)$ (a schedule of bonuses for revealing evidence). The bonus is only realized when $\omega \geq \overline{\omega}$ is revealed (for $\omega < \overline{\omega}, a(\omega) = 0$). We assume that $a(\overline{\omega}) \geq 0$, which is without loss of generality, as a negative bonus must result in concealment. This can be equivalently achieved by increasing $\overline{\omega}$ instead. The policy $(a_\phi, \overline{\omega}, a(\omega))$ is equivalent to committing to an entire mapping $a(M)$ from messages to actions. Describing the commitment policy in this way emphasizes the similarity to performance incentive contracts, which usually have a guaranteed payment (analogue of a_ϕ), a performance level at which bonuses payments begin (analogue of $\overline{\omega}$), and a bonus structure (analogue of $a(\omega)$).

Optimal Policy

A policy is optimal if it maximizes the social planner's expected payoff. In this model, every policy induces a unique equilibrium, so there is no ambiguity about the planner's expected payoffs from different policies. There may be multiple optimal policies, but the results presented here apply to all optimal policies.

The solution presented in this section applies to an optimal policy that induces effort $p \in (0, 1)$. If $p = 0$ is optimal, clearly the best policy is $a_\phi = \mathbb{E}_F[W]$ and $a(\omega) = 0, \forall \omega$. If $p = 1$ is optimal, the best policy is one which gives very severe punishments for $M = \phi$ (i.e., a_ϕ is very low) and otherwise leaves the receiver free to choose the action he likes. For interior effort levels, although it is impossible to solve explicitly for the optimal policy, there is a strong characterization of several properties that an optimal policy must have.

6 divides the optimal policy into two cases depending on the value of λ. If λ is low (the planner cares little about acquisition costs), then the solution is "overincentivized,"

costly, because it results in some degree of evidence being presented to the court, which takes up extra time, paperwork, etc. Although this model is not considered here, we believe that the key observations in our models carry over into that setting.

meaning the commitment actions all differ from their uncommitted levels in the direction that increases acquisition incentives (high rewards and punishments). If λ is high, then the solution is "underincentivized" (low rewards and punishments).

Proposition 6. *Let $Q(\lambda)$ be a joint distribution of messages and states in equilibrium given an optimal policy for parameter λ. Then there exists a threshold $\overline{\lambda}(r)$ such that*

1. *If $\lambda \leq \overline{\lambda}(r)$, then every optimal policy satisfies*

   (a) $a_\phi \leq \mathbb{E}_{Q(\lambda)}[W|M=\phi]$

   (b) $a_\phi + a(\omega) \geq \omega, \forall \omega \geq \overline{\omega}$, $a'(\omega) \leq 1, \forall \omega \geq \overline{\omega}$, *and* $a_\phi + a(\omega) \to \omega$ *as* $\omega \to \infty$,

   (c) $\mathbb{E}_F[a_\phi + a(W)] \leq \mathbb{E}_F[W]$

2. *If $\lambda \geq \overline{\lambda}(r)$, then every optimal policy satisfies*

   (a) $a_\phi \geq \mathbb{E}_{Q(\lambda)}[W|M=\phi]$

   (b) $a_\phi + a(\omega) \leq \omega, \forall \omega \geq \overline{\omega}$, $a'(\omega) \geq 1, \forall \omega \geq \overline{\omega}$, *and* $a_\phi + a(\omega) \to \omega$ *as* $\omega \to \infty$,

   (c) $\mathbb{E}_F[a_\phi + a(W)] \geq \mathbb{E}_F[W]$

Proof. See subsection 1.8.7

□

The optimal policy has several key features, which can be seen in Figure 1.2. Two optimal policies are graphed: one for $\lambda < \overline{\lambda}(r)$ and action/state distribution Q, and one for $\lambda > \overline{\lambda}(r)$ and action/state distribution Q'. First, the solution depends on how λ compares to $\overline{\lambda}(r)$. If $\lambda \leq \overline{\lambda}(r)$, then effort is "over-incentivized" relative to the no commitment equilibrium. These extra incentives manifest themselves both in high bonuses (part (b)) and in an unusually punitive default action (part (a)). On the graph, this can be seen as the flat part of the policy lies below $\mathbb{E}_Q[W|M=\phi]$, and the rest of the policy lies above the dashed line representing $a(\omega) = \omega$ (the uncommitted policy). If $\lambda \geq \overline{\lambda}(r)$, then effort is

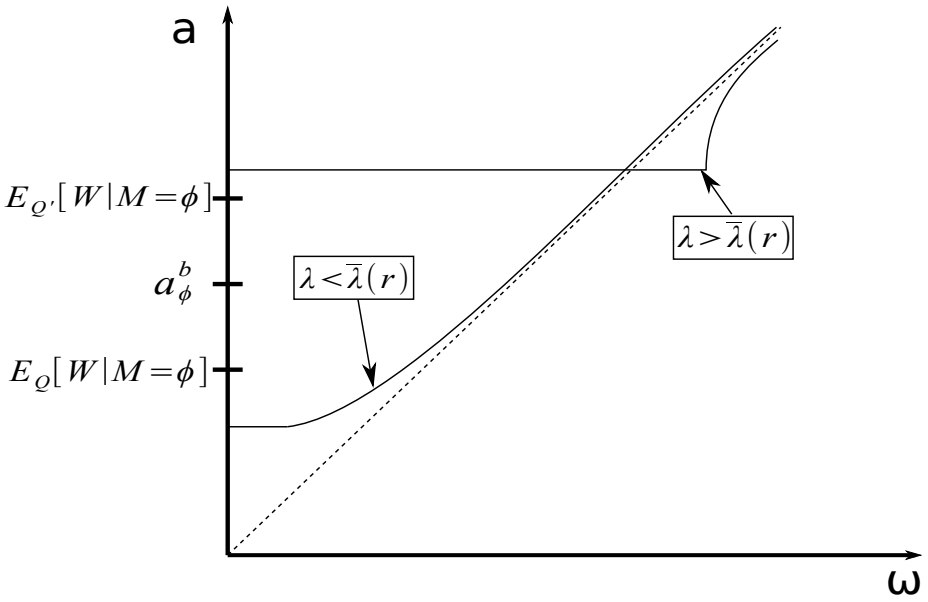


Figure 1.2: An Optimal Full Commitment Policy

"under-incentivized" relative to no commitment both in the default action and the bonuses. In the graph, this shows up as the flat part of the policy being above $\mathbb{E}_{Q'}[W|M=\phi]$ and the rest of the policy lying below the dashed line.

Overall, the incentives come disproportionately from punishments than from rewards, as the expected action in the over-incentivized case is less than the expected action in the under-incentivized case (part (c)). Moreover, the results on the slope of the bonus function (in part (b)) indicate that the planner places incentives disproportionately on lower states. Both of these results are due to the sender's risk aversion, which implies that marginal utility decreases as the action rises. A risk-neutral sender has incentives more balanced across all states, so the expected action does not depend on λ.

The optimal policy depends on the sender's degree of risk aversion, as shown in 7.

Proposition 7.

1. *$\overline{\lambda}(r) \to \infty$ as $r \to \infty$.*

2. *For all λ, every optimal policy satisfies*

(a) As $r \to 0$, $a'(\omega) \to 1, \forall \omega \geq \overline{\omega}$ pointwise, and $\mathbb{E}_F[a_\phi + a(W)] \to \mathbb{E}_F[W]$.

(b) As $r \to \infty$, if $\frac{1}{2} \cdot \lambda \cdot \alpha > Var_F(W)$ then $p^* \to 0$, and if $\frac{1}{2} \cdot \lambda \cdot \alpha < Var_F(W)$ then $p^* \to 1$.

Proof. See subsection 1.8.7. □

Part 1 of 7 shows that an optimal policy must be an overincentivized policy for high degrees of risk aversion. This makes sense given the previous results in this paper. Risk aversion has a dampening effect on evidence acquisition. The social planner combats this with higher powered incentives of the form found in 6.

Although 7 says that $p^* \to 1$ as $r \to \infty$ is possible, this is only because the default action can be made arbitrarily low. In the likely case that there is an exogenous maximum punishment (e.g., the prohibition on "excessive fines" and "cruel and unusual punishments" in the Eighth Amendment to the U.S. Constitution), maximal evidence acquisition is impossible in the limit, because there is always a positive probability that the sender receives a_ϕ. In that case, there must be no evidence acquisition in the infinitely risk-averse limit, because incentivizing positive acquisition effort becomes prohibitively costly in terms of having to arrive at ex-post suboptimal judgments.

1.6.2 Threshold Commitment

In legal settings, there is often a standard for determining what evidence is admissible in court. This is captured by our model as constraining the receiver to ignore certain types of evidence. To focus attention on this specific design problem, the planner can now only commit to a threshold which determines whether evidence is ignored (treated the same as the absence of evidence). In an example where ω measures the degree of guilt, this rule says that only evidence that shows a sufficient degree of guilt or innocence is admissible. In an example where ω measures the probability of guilt, this rule says that evidence that

does not prove a high enough probability of guilt or innocence is not admissible (this is like a requirement that evidence be probative). We assume that the default action a_ϕ is exogenous, so the observability of effort does not matter. This is often a reasonable assumption. For example, if the sender is a plaintiff suing someone for damages, a judgment of no liability is the default action.

The planner commits in advance to evidence admissibility threshold $\overline{\omega}$, where only evidence above $\overline{\omega}$ is accepted. Without loss of generality, we constrain the sender to thresholds $\overline{\omega} \geq a_\phi$ (a best responding sender will never reveal a state ω satisfying $\overline{\omega} < \omega < a_\phi$). The following proposition describes the optimal policy and how it relates to the risk aversion of the sender.

Proposition 8. *The optimal threshold commitment policy satisfies the following:*

1. *For sufficiently low λ, $\overline{\omega} = a_\phi$. For sufficiently high λ, $\overline{\omega} > a_\phi$.*

2. *As $\lambda \to \infty$, $\overline{\omega} \to \infty$.*

3. *As $r \to \infty$, $p^* \to 0$, and $\overline{\omega} \to a_\phi$.*

Proof. See subsection 1.8.8. □

This proposition follows the same pattern as 6 and 7. Low λ leads to strong acquisition incentives, and high λ leads to weak ones. Moreover, high risk aversion justifies stronger incentives, but in the limit they cannot be strong enough to result in acquisition.

1.6.3 Maximum Punishment

In many cases, there is a maximum punishment that may be inflicted on the sender. For example, if the sender is the defendant in a simple property damage case, the maximum punishment is usually the replacement value of the property in question. To establish a

maximum punishment, the planner commits to a_ϕ, but everything else is freely chosen by the receiver. Note that observability of effort does not matter here, since the action is fixed in every event where the posterior is non-degenerate.

Proposition 9. *Let $Q(\lambda)$ be a joint distribution of messages and states in equilibrium given an optimal policy for parameter λ. Then in an optimal maximum punishment commitment policy a_ϕ there exists $\overline{\lambda}(r)$ such that:*

1. *If $\lambda < \overline{\lambda}(r)$, then $a_\phi < \mathbb{E}_{Q(\lambda)}[W|M = \phi]$.*
2. *If $\lambda > \overline{\lambda}(r)$, then $a_\phi > \mathbb{E}_{Q(\lambda)}[W|M = \phi]$.*
3. *$\overline{\lambda}(r) \to \infty$ as $r \to \infty$.*
4. *If $\lambda \cdot \frac{1}{2} \cdot \alpha < Var_F(W)$, then as $r \to \infty$, $p^* \to 1$, and $a_\phi \to -\infty$.*
5. *If $\lambda \cdot \frac{1}{2} \cdot \alpha > Var_F(W)$, then as $r \to \infty$, $p^* \to 0$, and $a_\phi \to \mathbb{E}_F[W]$.*

Proof. See subsection 1.8.9. □

Parts 1 and 2 of 9 show that a low cost weight results in overincentivization (low a_ϕ) and a high cost weight results in underincentivization (high a_ϕ). Part 3 shows that overincentivization must hold for sufficiently high risk aversion. The final two parts show that the limiting outcome is either maximal effort or minimal effort, depending on the cost parameters and the underlying uncertainty in the prior. However, maximal effort in the limit again relies on arbitrarily large penalties being allowed ($a_\phi \to -\infty$).

1.6.4 Minimum Overall Punishment

This section analyzes the situation where the planner can commit to a minimum punishment (alternatively, a maximum award) that must be faced by the litigant simply by virtue of appearing in court. In this model, the planner commits to a cap a_H on the

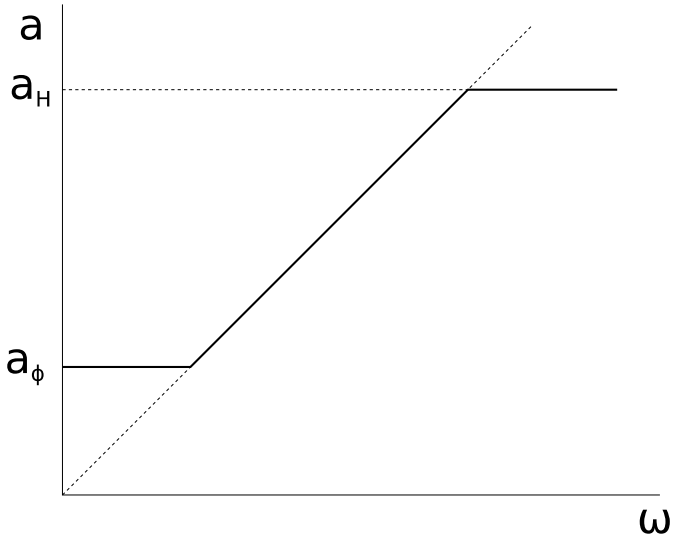


Figure 1.3: Minimum Overall Punishment

action, which may be infinite (no cap). We will assume that $a_\phi \leq \mathbb{E}_F[W]$ is exogenous, so the observability of effort is not an issue. This also prevents $a_H < a_\phi$ from ever being a solution ($a_H = \mathbb{E}_F[W]$ is always better). The receiver's strategy in equilibrium must be to choose a_ϕ following an empty message, ω following $M = \omega$ whenever $\omega \leq a_H$, and a_H whenever $M = \omega$ for $\omega > a_H$ is observed. The sender will conceal if and only if $\omega < a_\phi$. This type of policy is illustrated in Figure 1.3.

Proposition 10. *There exists $\overline{\lambda}(r)$ such that the optimal minimum overall punishment commitment policy a_H satisfies the following:*

1. $\frac{da_H}{d\lambda} \leq 0$.
2. *If $\lambda \leq \overline{\lambda}(r)$, then there is no minimum overall punishment ($a_H = \infty$). Otherwise, there is one ($a_H < \infty$).*
3. $\overline{\lambda}(r) \to \infty$ *as* $r \to \infty$.

Proof. See subsection 1.8.10. □

As before, for low λ or high r, there are high powered acquisition incentives. In this case, high powered incentives means high a_H (low minimum punishment).

1.6.5 Minimum Conditional Punishment

This section focuses on a model where the planner can commit to a minimum punishment magnitude required by passing a fixed evidence threshold (e.g., a requirement of guilt being proven beyond any reasonable doubt). This can be interpreted as a model of a trial where there is a minimum penalty if the defendant is found guilty or liable. Alternatively, this can be interpreted as establishing a maximum award that can be given to a plaintiff who has not met the evidence threshold. The default action a_ϕ is taken as given. There is also an exogenous evidence threshold $\overline{\nu} \geq a_\phi$ (this is different from the concealment threshold $\overline{\omega}$). The planner commits to $B \geq 0$, which is the minimum jump in the action that occurs when state $\overline{\nu}$ is revealed. If state $\overline{\nu}$ or lower is revealed, then the receiver's action must be at most $\overline{\nu} - B$. This is lower than the unconstrained action for all states between $\overline{\nu} - B$ and $\overline{\nu}$, so the receiver will choose action $\overline{\nu} - B$. For all states $\omega > \overline{\nu}$, he will choose action ω. There is a natural upper bound on B, since $\overline{\nu} - B \geq a_\phi$, which implies $B \leq \frac{\overline{\nu}}{a_\phi}$. This type of policy is illustrated in Figure 1.4.

Proposition 11. *In the optimal minimum conditional punishment commitment policy B, there exist $\overline{\lambda}(r) > \underline{\lambda}(r)$ such that:*

1. *If $\lambda \leq \underline{\lambda}(r)$, $B = 0$ (i.e., no minimum conditional punishment).*
2. *If $\lambda \geq \overline{\lambda}(r)$, $B = \frac{\overline{\nu}}{a_\phi}$.*
3. *If $\lambda \in (\underline{\lambda}(r), \overline{\lambda}(r))$, then B is interior.*
4. *As $r \to \infty$, $\underline{\lambda}(r) \to \infty$.*

Proof. See subsection 1.8.11. □

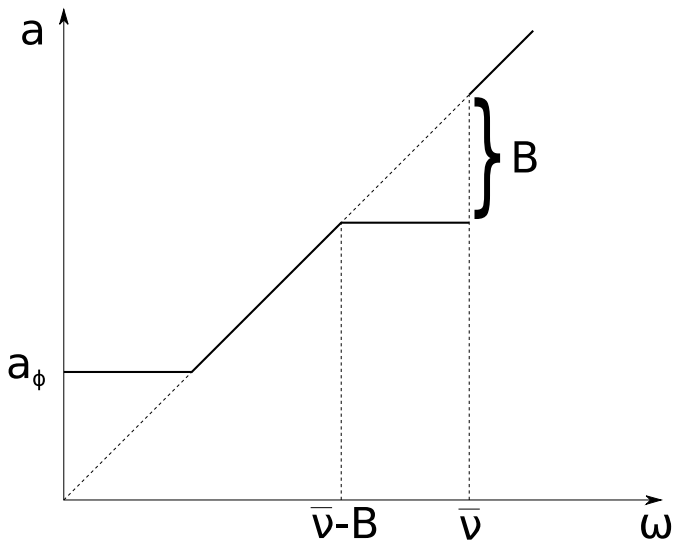


Figure 1.4: Minimum Conditional Punishment

Minimum conditional punishments force the receiver to give lower actions for revealed evidence, which lessens evidence gathering incentives. Therefore, when λ is low and high effort is desirable, the incentives are maximized at $B = 0$. When effort is less desirable, there may be a minimum conditional punishment to reduce costly evidence gathering activities. As in earlier results, high risk aversion leads to high powered incentives.

1.7 Conclusion

In this paper, we study evidence acquisition incentives in litigation and how they relate to risk preferences. We first analyze how risk aversion interacts with the observability of effort. A greater probability of observing effort leads to lower equilibrium effort levels. However, for very risk-averse litigants, this effect is minimal, and effort is always near zero. In this way, risk aversion on the part of the litigants causes the court to make poor decisions on average, which leads to unfair restitution and a weak deterrent effect. In the Appendix, we show that this result is robust to multiple competing litigants and to a certain extent to the case where litigants have private information.

Legal institutions can counteract weak incentives for evidence acquisition by constraining the court's freedom to assign rewards and punishments. To this end, we examine several different types of commitment policies and how they interact with the litigant's risk preferences. Specifically, we analyze full commitment, evidence admissibility thresholds, maximum punishments, minimum overall punishments, and minimum conditional punishments. We find that the socially optimal policies of various types of commitment are heavily dependent on the litigant's degree of risk aversion. In general, high risk aversion warrants high powered acquisition incentives built into these legal rules.

Since different types of litigants may have different degrees of risk aversion, for example individuals may be more risk averse than corporations, these results have implications for optimal legal rules for the admissibility of evidence and bounds on the magnitude of monetary awards. In particular, the socially optimal policy differentiates between litigants based on personal characteristics that are independent from the facts of the case, because heterogeneity in these characteristics tends to produce different outcomes at trial. This demonstrates a conflict between nominal equal protection and de facto equal protection, the latter of which requires equal trial outcomes and equal deterrent effect. Because of different underlying tendencies regarding evidence acquisition, unequal rules may be required to have de facto equal protection.

1.8 Appendix

1.8.1 Proof of 1

Proof. First, consider the subgames $\Gamma(y)$. We will show that there exists $\overline{\omega}(y)$ such that there is a "threshold equilibrium" on the subgame $\Gamma(y)$ with threshold $\overline{\omega}(y)$. That is, $\sigma^S(x, y)$ places probability 1 on $M = \phi$ if $x \leq \overline{\omega}(y)$ or $x = \phi$, and it places probability 1 on $M = x$ otherwise. Specifically, $a(\overline{\omega}(y)) = \sigma^R(\phi, y)$ implies that the sender is best responding

(this is a slight abuse of notation, since $\sigma^R(\phi, y)$ is technically a degenerate distribution). The receiver's posterior belief is

$$F(\omega|p=y,\overline{\omega},M=\phi)=\begin{cases}\frac{(1-y)\cdot F(\omega)+y\cdot F(\overline{\omega})}{(1-y)+y\cdot F(\overline{\omega})} & \text{if } \omega>\overline{\omega}\\ \frac{F(\omega)}{(1-y)+y\cdot F(\overline{\omega})} & \text{if } \omega\leq\overline{\omega}\end{cases}$$

Note that $F(\omega|p=y,\overline{\omega},M=\phi)$ is continuous in $\overline{\omega}$ in the uniform metric, so there is a continuous function $a_\phi(\overline{\omega};y)$ giving the receiver's best response when $M=\phi$ (in fact, $a_\phi(\overline{\omega};y)$ is continuously differentiable in both $\overline{\omega}$ and y). When $y=1$, setting $\overline{\omega}(y)=\omega_0$ results in an equilibrium, since $a(\omega_0)=\sigma^R(\phi,1)$ (this is a standard "unraveling" equilibrium). Otherwise, note that $a_\phi(\overline{\omega}(y);y)\to a^*$ as $\overline{\omega}(y)\to\omega_0$, and $a_\phi(\overline{\omega}(y);y)\to a^*$ as $\overline{\omega}(y)\to\infty$ (in both cases, the posterior belief following $M=\phi$ converges pointwise to the prior belief F). Since $a(\omega_0)\leq a^*$ and $\lim_{\omega\to\infty}a(\omega)\geq a^*$ and $a(\omega)$ is a continuous function of ω, there must be a crossing where $a(\overline{\omega}(y))=a_\phi(\overline{\omega}(y);y)$. The supermodularity of u^R ensures that this threshold messaging strategy is a best response for the sender. Moreover, $\overline{\omega}(y)$ defined in this way is continuously differentiable.

Let $U^S(y)$ be the sender's equilibrium expected utility on $\Gamma(y)$. Then,

$$\begin{aligned}U^S(y)=&\,y\cdot[1-F(\overline{\omega}(y))]\cdot\mathbb{E}[u^S(a(\omega))|\omega>\overline{\omega}(y)]\\&+[(1-y)+y\cdot F(\overline{\omega}(y))]\cdot u^S(a_\phi(\overline{\omega}(y);y))\end{aligned}$$

Given that $\overline{\omega}(y)$ and $a_\phi(\overline{\omega};y)$ are both continuously differentiable, $U^{S\prime}(y)$ exists and is continuous.

Now, we consider the game Γ. Let $\hat{p}$ be the effort the receiver believes the sender to exert, and let $\hat{a}_\phi$ be the receiver's optimal action given the posterior induced by $\hat{p}$, the

believed messaging threshold $\hat{\overline{\omega}}$, and $M = \phi$. The sender's expected utility in Γ is

$$q \cdot U^S(p) + (1-q) \cdot \{p \cdot [1 - F(\overline{\omega})] \cdot \mathbb{E}[u^S(a(\omega)) | \omega > \overline{\omega}] \\ + [(1-p) + p \cdot F(\overline{\omega})] \cdot u^S(\hat{a}_\phi)\} - c(p)$$

Since the sender can always guarantee $\hat{a}_\phi$ by sending $M = \phi$, only states inducing higher actions will be revealed. Therefore, in any best response, $\overline{\omega} = a^{-1}(\hat{a}_\phi)$ (convexity of Ω and continuity and supermodularity assumptions guarantee $a^{-1}(\hat{a}_\phi)$ exists and is unique). The sender's first-order condition (FOC) for an interior optimum in p is

$$c'(p) = q \cdot U^{S'}(p) + (1-q) \cdot [1 - F(a^{-1}(\hat{a}_\phi))] \cdot \{\mathbb{E}[u^S(a(\omega)) | \omega > a^{-1}(\hat{a}_\phi)] - u^S(\hat{a}_\phi)\}$$

The receiver takes in believed effort $\hat{p}$, believed threshold $\hat{\overline{\omega}}$, and message $M = \phi$ and maximizes $\mathbb{E}[u^R(W, a) | \hat{p}, \hat{\overline{\omega}}, M = \phi]$. The receiver's best response evaluated at $M = \phi$ (call this a_ϕ) always satisfies the FOC $\mathbb{E}[u^{R'}(W, a_\phi) | \hat{p}, \hat{\overline{\omega}}, M = \phi] = 0$. Therefore, the necessary interior equilibrium conditions are

$$c'(p) = q \cdot U^{S'}(p) + (1-q) \cdot [1 - F(\overline{\omega})] \cdot \{\mathbb{E}[u^S(a(\omega)) | \omega > \overline{\omega}] - u^S(a_\phi)\} \tag{1.1}$$

$$\overline{\omega} = a^{-1}(a_\phi) \tag{1.2}$$

$$\mathbb{E}[u^{R'}(W, a_\phi) | p, \overline{\omega}, M = \phi] = 0 \tag{1.3}$$

Define $a_0 \equiv argmax_a u^R(\omega_0, a)$. Taking $p < 1$ as fixed, there is always a solution to conditions 1.2 and 1.3 by the following argument: In the limits as $\overline{\omega} \to \omega_0$ and as $\overline{\omega} \to \infty$, the posterior beliefs following $M = \phi$ converge to the prior F. In the former limit, $\mathbb{E}[u^{R'}(W, a(\overline{\omega}))|p, \overline{\omega}, M = \phi]$ converges to something ≥ 0, and in the latter limit, $\mathbb{E}[u^{R'}(W, a(\overline{\omega}))|p, \overline{\omega}, M = \phi]$ converges to something ≤ 0 or diverges to $-\infty$. Since $a(\overline{\omega})$ and the posterior beliefs are continuous in $\overline{\omega}$, and $u^{R'}(\omega, a)$ is continuous, a solution exists. In the $p = 1$ case, setting $\overline{\omega} = \omega_0$ and $a_\phi = a_0$ satisfies both conditions. Furthermore, there exist $\overline{\omega}(p)$ and $a_\phi(p)$ such that $(\overline{\omega}(p), a_\phi(p))$ solves conditions 1.2 and 1.3, and $\overline{\omega}(p)$ and $a_\phi(p)$ are continuous in p (this follows from various continuity assumptions).

If conditions 1.1 - 1.3 never hold, and $q \cdot U^{S'}(0) + (1 - q) \cdot [1 - F(a^{-1}(a^*))] \cdot \{\mathbb{E}[u^S(a(\omega))|\omega > a^{-1}(a^*)] - u^S(a^*)\} \leq 0$, then there is an equilibrium where $p = 0, \overline{\omega} = a^{-1}(a^*)$, and $a_\phi = a^*$. If conditions 1.1 - 1.3 never hold, and $q \cdot U^{S'}(1) + (1 - q) \cdot \{\mathbb{E}[u^S(a(\omega))] - u^S(a_0)\} \geq c'(1)$, then there is an equilibrium where $p = 1, \overline{\omega} = \omega_0$, and $a_\phi = a_0$. If neither of these inequalities hold, then the RHS of condition 1.1 starts off less than the LHS at $p = 0$ and ends up greater than the LHS at $p = 1$ (using $\overline{\omega} = \overline{\omega}(p)$ and $a_\phi = a_\phi(p)$ to adjust for best response changes). Since both sides are continuous in p, there must be a solution to conditions 1.1 - 1.3.

□

1.8.2 Proof of 2

Proof. Suppose to the contrary that $p = 0$ in equilibrium for all small q. Let $a_u(M)$ be the receiver's equilibrium action after observing message M when effort is unobserved and let $a_o(M)$ be the same for when effort is observed. Let G be the equilibrium distribution of messages when effort is unobserved. We will show that a marginal increase in p is strictly beneficial to the sender as long as q is low enough.

In the deviation messaging strategy, if the receiver succeeds in observing the sender's

effort, the sender reveals everything (when x is acquired, $M = \omega$). If the receiver fails to observe the sender's effort, the sender conceals the state when optimal and reveals the state (if x is acquired) when optimal. When x is successfully acquired, the sender sends message $M = \phi$ whenever $a_u(\phi) \geq a_u(\omega)$ and message $M = \omega$ whenever $a_u(\phi) < a_u(\omega)$. This induces a new distribution of messages, which we call H. This latter "high state" case must occur with positive probability, so $\mathbb{E}_H[u^S(a_u(M))] > \mathbb{E}_G[u^S(a_u(M))]$.

The marginal effect on the sender's utility from using this strategy and increasing p (starting from $p = 0$) is the following:

$$
\begin{aligned}
& q \cdot \{[\mathbb{E}_F[u^S(a_o(\omega))] - u^S(a_o(\phi)]\} \\
& + (1-q) \cdot \{\mathbb{E}_H[u^S(a_u(M))] - \mathbb{E}_G[u^S(a_u(M))]\} \\
& - c'(0)
\end{aligned}
$$

The first term in $\{\}$ may be positive or negative, but it is finite and does not depend on q. The second term in $\{\}$ is strictly positive by construction and does not depend on q. Finally, the marginal cost term is 0 by assumption. Therefore, for sufficiently small q, the deviation is strictly profitable, so $p = 0$ could not have been an equilibrium for all small q. □

1.8.3 Proof of 3

Proof. Let Q be any joint probability measure of states and messages sent (when there is only one argument, it is taken to be the corresponding marginal probability measure), and let G be the corresponding distribution of the actions taken by the receiver (A is the random variable representing actions). The expected action is the following:

$$\mathbb{E}_G[A] = \int_{M \in X} \mathbb{E}_Q[W|M] \cdot dQ(M)$$

This is itself an expectation of expectations conditional on the message being sent. Thus, using the Law of Iterated Expectations,

$$\begin{aligned}\mathbb{E}_G[A] &= \mathbb{E}_Q[\mathbb{E}_Q[W|M]] \\ &= \mathbb{E}_Q[W] = \mathbb{E}_F[W]\end{aligned}$$

This means that any sender strategy induces the same expected action in equilibrium. For now, assume that $q = 1$. For any $\overline{p} > 0$, if $p \geq \overline{p}$, then there is a positive probability of acquiring evidence of the state. This means that the action distribution induced by p is a mean-preserving spread of that induced by $p = 0$. Since the sender is risk-averse and costs are increasing, the sender could strictly benefit from deviating to $p = 0$ and always sending messages $M = \phi$. As long as q is high enough, this deviation is still strictly beneficial, meaning p is not an equilibrium effort. Moreover, $p = \overline{p}$ induces the least spread of these distributions, so q high enough for $p = \overline{p}$ to be incompatible with equilibrium is also high enough for all $p > \overline{p}$. This establishes the existence of $\overline{q}$.

Moreover, if $q = 1$, there is an equilibrium where $p = 0$. The messaging strategy must always send message $M = \phi$ (this is the only feasible message). The receiver's beliefs on receiving M place probability 1 on ω if $M = \omega$ and are equal to the prior if $M = \phi$. The receiver always picks the unique optimal action given these beliefs. Any deviation results

in an expected receiver action of $\mathbb{E}_F[W]$, so this is not a strictly beneficial deviation for a strictly risk-averse sender, and this construction is in fact an equilibrium. □

1.8.4 Proof of 4

Proof. Take any fixed q and any $p \in P(q)$. The marginal effect of increasing p depends on the observability parameter q. Let a_ϕ be the receiver's response to the empty message in equilibrium (this is exactly the same in the observable and unobservable effort cases). In the unobservable effort case, the marginal effect of increasing effort at p (ignoring costs) is $[1-F(a_\phi)] \cdot \{\mathbb{E}_F[u^S(W;r)|W > a_\phi] - u^S(a_\phi;r)\}$.

In the observable effort case, the marginal effect is more complicated. Firstly, the best response default action a_ϕ changes. The receiver's posterior following an empty message must be the following:

$$F(\omega|p, M=\phi) = \begin{cases} \frac{(1-p)\cdot F(\omega)+p\cdot F(\overline{\omega})}{(1-p)+p\cdot F(\overline{\omega})} & \text{if } \omega > \overline{\omega} \\ \frac{F(\omega)}{(1-p)+p\cdot F(\overline{\omega})} & \text{if } \omega \leq \overline{\omega} \end{cases}$$

Therefore,

$$\begin{aligned} a_\phi &= \frac{1}{(1-p)+p\cdot F(\overline{\omega})} \cdot \int_{-\infty}^{\overline{\omega}} \omega \cdot dF(\omega) + \frac{1-p}{(1-p)+p\cdot F(\overline{\omega})} \cdot \int_{\overline{\omega}}^{\infty} \omega \cdot dF(\omega) \\ &= \frac{E_F[W] - p\cdot[1-F(\overline{\omega})]\cdot \mathbb{E}_F[W|W>\overline{\omega}]}{(1-p)+p\cdot F(\overline{\omega})} \end{aligned}$$

The optimal concealment threshold is $\overline{\omega} = a_\phi$, so if everyone is best responding given p,

$$a_\phi = \frac{E_F[W] - p \cdot [1 - F(a_\phi)] \cdot \mathbb{E}_F[W|W > a_\phi]}{(1-p) + p \cdot F(a_\phi)}$$

The implicit effect of a_ϕ on itself in the preceding equation turns out to be zero, so

$$\begin{aligned}\frac{da_\phi}{dp} &= \frac{-[(1-p) + p \cdot F(a_\phi)] \cdot [1 - F(a_\phi)] \cdot \mathbb{E}_F[W|W > a_\phi]}{[(1-p) + p \cdot F(a_\phi)]^2} \\ &\quad - \frac{-[1 - F(a_\phi)] \cdot \{E_F[W] - p \cdot [1 - F(a_\phi)] \cdot \mathbb{E}_F[W|W > a_\phi]\}}{[(1-p) + p \cdot F(a_\phi)]^2} \\ &= \frac{-[1 - F(a_\phi)] \cdot \{\mathbb{E}_F[W|W > a_\phi] - E_F[W]\}}{[(1-p) + p \cdot F(a_\phi)]^2} < 0\end{aligned}$$

The sender's equilibrium expected utility (ignoring costs) in the observable effort case is

$$[(1-p) + p \cdot F(a_\phi)] \cdot u^S(a_\phi; r) + p \cdot [1 - F(a_\phi)] \cdot \mathbb{E}_F[u^S(W; r)|W > a_\phi]$$

Then, the marginal effect of increasing p is

$$\begin{aligned}&[1 - F(a_\phi)] \cdot \{\mathbb{E}_F[u^S(W; r)|W > a_\phi] - u^S(a_\phi; r)\} \\ &+ [(1-p) + p \cdot F(a_\phi)] \cdot u^{S'}(a_\phi; r) \cdot \frac{da_\phi}{dp}\end{aligned}$$

The overall marginal effect of increasing p for fixed q is

$$\begin{aligned}&[1-F(a_\phi)]\cdot\{\mathbb{E}_F[u^S(W;r)|W>a_\phi]-u^S(a_\phi;r)\}\\&+q\cdot[(1-p)+p\cdot F(a_\phi)]\cdot u^{S'}(a_\phi;r)\cdot\frac{da_\phi}{dp}\end{aligned}$$

Since the second term is strictly negative for all p, increasing q to q' brings down the entire marginal benefit curve (as a function of p). Since the marginal cost curve is staying fixed, this results in strictly lower equilibrium efforts (unless it hits the lower bound of 0). □

1.8.5 Proof of 1

Lemma 1. *There exist* $C\in\mathbb{R},\overline{a}\in\mathbb{R}$ *such that*

$$\lim_{r\to\infty}u^S(a;r)=\begin{cases}C & \text{if } a\geq\overline{a}\\ -\infty & \text{if } a<\overline{a}\end{cases}$$

Proof. Let $R(u^S,a;r)\equiv-\frac{u^{S''}(a;r)}{u^{S'}(a;r)}$. Take any $a_H>a_L$. Then

$$\begin{aligned}u^{S'}(a_H;r)&=u_n'(a_L;r)+\int_{a_L}^{a_H}u^{S''}(a;r)da\\&=u^{S'}(a_L;r)-\int_{a_L}^{a_H}u^{S'}(a;r)\cdot R(u^S,a;r)da\\&\leq u^{S'}(a_L;r)-u^{S'}(a_H;r)\cdot\int_{a_L}^{a_H}R(u^S,a;r)da\\\Leftrightarrow u^{S'}(a_H;r)&\leq u^{S'}(a_L;r)\cdot\left[1+\int_{a_L}^{a_H}R(u^S,a;r)da\right]^{-1}\end{aligned}$$

Since this holds for all $a_H > a_L$, this inequality can be integrated to produce

$$\begin{aligned}
&\int_{a_L}^{a_H} u^{S'}(b;r)\cdot db \leq u^{S'}(a_L;r)\cdot \int_{a_L}^{a_H}\left[1+\int_{a_L}^{b} R(u^S,a;r)da\right]^{-1}\cdot db \\
\Leftrightarrow & u^S(a_L;r)+\int_{a_L}^{a_H} u^{S'}(b;r)\cdot db \leq u^S(a_L;r) \\
&+u^{S'}(a_L;r)\cdot \int_{a_L}^{a_H}\left[1+\int_{a_L}^{b} R(u^S,a;r)da\right]^{-1}\cdot db \\
\Leftrightarrow & u^S(a_H;r)-u^S(a_L;r) \leq u^{S'}(a_L;r)\cdot \int_{a_L}^{a_H}\left[1+\int_{a_L}^{b} R(u^S,a;r)da\right]^{-1}\cdot db
\end{aligned}$$

Since $\int_{a_L}^{b} R(u^S,a;r)da \to \infty$ as $r\to\infty$ for all $b > a_L$, one of two results must hold: either $u^{S'}(a_L;r)$ is bounded and $u^S(a_H)-u_n(a_L)\to 0$ or $u^{S'}(a_L;r)\to\infty$. The former case indicates that in the limit utility is flat for all $a \geq a_L$. The latter case implies that the limiting utility is $-\infty$ for all $a < a_L$. Right continuity gives the functional form from the statement of the Lemma.

□

1.8.6 Proof of 5

Proof. The proof proceeds by considering the cases where effort is observed and unobserved separately. In both the $q=0$ and $q=1$ cases, sufficiently high risk-aversion pushes the equilibrium effort below $\overline{p}$. Therefore, the higher of these two levels of risk-aversion is enough to prove the result for any $q\in(0,1)$.

Before splitting into separate cases, note that $u^S(a;r)$ must be converging to a step function of the form

$$\lim_{r\to\infty} u^S(a;r) = \begin{cases} C & \text{if } a \geq \overline{a} \\ -\infty & \text{if } a < \overline{a} \end{cases}$$

Here, $C \in \mathbb{R}$ and $\overline{a} \in \mathbb{R}$. For a proof of this fact, see 1 in subsection 1.8.5.

Case: q = 0

Suppose to the contrary that for some $\overline{p}$ and for every $\overline{r}$, there exists $r \geq \overline{r}$ such that some equilibrium of the game induced by r has $p \geq \overline{p}$. This implies that in the limit as $r \to \infty$, equilibrium effort does not always converge to 0. In equilibrium, the marginal benefit of increasing p must equal the marginal cost. Note that any change in p is not observed, so the receiver's action in response to any given message is the same after the deviation. Since the receiver's best responses are the same, the sender's optimal mapping from pieces of evidence to messages must be the same. Let G be the action distribution induced by successfully acquiring evidence and using the equilibrium disclosure strategy. Define a_ϕ as the equilibrium action taken following $M = \phi$. Hence, a_ϕ is the lowest action in the support of G. Then, the equation asserting that marginal benefit equals marginal cost is the following:

$$\mathbb{E}_G[u^S(A;r)] - u^S(a_\phi;r) = c'(p)$$

Following 1, if $a_\phi \geq \overline{a}$, the left hand side clearly converges to 0, so $p \to 0$, which is a contradiction. If $a_\phi < \overline{a}$, the utility from a_ϕ eventually dominates in $\mathbb{E}_G[u^S(A;r)]$ (a_ϕ always occurs with strictly positive probability), so again the left hand side converges to 0.

Case: q = 1

Again, suppose that p does not always converge to 0 as $r \to \infty$. Since u^R is

supermodular, the optimal disclosure strategy is always a threshold strategy, implying that the action following $M = \phi$ is the lowest possible action. Also, as p decreases, the belief following $M = \phi$ increases (FOSD), so this lowest action must increase. By 1 the distribution with the higher minimum action must be weakly preferred in the limit as $r \to \infty$. Since the cost also decreases as p falls, every equilibrium effort must eventually fall below $\overline{p}$, which is a contradiction. □

1.8.7 Proof of 6 and 7

Proof. The sender's expected utility is

$$p \cdot [1 - F(\overline{\omega})] \cdot \mathbb{E}_F[u^S(a_\phi + a(W); r) | W \geq \overline{\omega}]$$
$$+ [(1-p) + p \cdot F(\overline{\omega})] \cdot u^S(a_\phi; r) - \frac{1}{2} \cdot \alpha \cdot p^2$$

The F.O.C. yields

$$p^* = \frac{1}{\alpha} \cdot [1 - F(\overline{\omega})] \cdot \{\mathbb{E}_F[u^S(a_\phi + a(W); r) | W \geq \overline{\omega}] - u^S(a_\phi; r)\} \geq 0$$

The relevant derivatives[11] are

$$\begin{aligned}
\frac{dp^*}{da_\phi} &= \frac{1}{\alpha} \cdot \{ \int_{\overline{\omega}}^{\infty} u^{S'}(a_\phi + a(\omega); r) \cdot dF(\omega) - [1 - F(a_\phi)] \cdot u^{S'}(a_\phi; r) \} \\
&= \frac{1}{\alpha} \cdot [1 - F(a_\phi)] \cdot \{ \mathbb{E}_F[u^{S'}(a_\phi + a(W); r) | W > a_\phi] - u^{S'}(a_\phi; r) \} \leq 0 \\
\frac{dp^*}{d\overline{\omega}} &= -\frac{1}{\alpha} \cdot f(\overline{\omega}) \cdot \{ u^S(a_\phi + a(\overline{\omega}); r) - u^S(a_\phi; r)]) \} \leq 0 \\
\frac{dp^*}{da(\omega)} &= \frac{1}{\alpha} \cdot u^{S'}(a_\phi + a(\omega); r) \cdot f(\omega) \geq 0, \forall \omega \geq \overline{\omega}
\end{aligned}$$

It is important to allow p^* to vary depending on the exact policy, and the policy itself may depend on parameter values. Therefore, these can be thought of as $p^*(\alpha, r, \lambda)$, $a_\phi(\alpha, r, \lambda)$, $\overline{\omega}(\alpha, r, \lambda)$, and $a(\omega; \alpha, r, \lambda)$. However, these additional arguments will usually be suppressed in the notation.

The planner's expected utility is

$$\begin{aligned}
&p^* \cdot [\int_{\overline{\omega}}^{\infty} -(\omega - [a_\phi + a(\omega)])^2 \cdot dF(\omega) + \int_{-\infty}^{\overline{\omega}} -(\omega - a_\phi)^2 \cdot dF(\omega)] \\
&+ (1 - p^*) \cdot \int_{-\infty}^{\infty} -(\omega - a_\phi)^2 \cdot dF(\omega) - \lambda \cdot \frac{1}{2} \cdot \alpha \cdot p^{*2}
\end{aligned}$$

Define

[11] The derivative $\frac{dp^*}{da(\omega)}$ is an abuse of notation meant to capture the effect of a marginal increase in the bonus structure in a single location. In reality, all of these derivatives are zero. However, the planner will be deciding on the entire function $a(\omega)$, and in the planner's combined F.O.C.'s this derivative helps capture the overall effect of changing the entire function $a(\omega)$.

$$\Delta(a_\phi,\overline{\omega},a) \equiv [\int_{\overline{\omega}}^{\infty} -(\omega-[a_\phi+a(\omega)])^2\cdot dF(\omega) + \int_{-\infty}^{\overline{\omega}} -(\omega-a_\phi)^2\cdot dF(\omega) - \int_{-\infty}^{\infty} -(\omega-a_\phi)^2\cdot dF(\omega)$$

The F.O.C.'s are

$$\begin{aligned}
0 &= p^*\cdot[\int_{\overline{\omega}}^{\infty} 2\cdot(\omega-[a_\phi+a(\omega)])\cdot dF(\omega) + \int_{-\infty}^{\overline{\omega}} 2\cdot(\omega-a_\phi)\cdot dF(\omega)] \\
&\quad + (1-p^*)\cdot\int_{-\infty}^{\infty} 2\cdot(\omega-a_\phi)\cdot dF(\omega) + [\Delta-\lambda\cdot\alpha\cdot p^*]\cdot\frac{dp^*}{da_\phi} \\
&= \int_{-\infty}^{\infty} 2\cdot(\omega-a_\phi)\cdot dF(\omega) + p^*\cdot[\int_{\overline{\omega}}^{\infty} -2\cdot a(\omega)\cdot dF(\omega)] \\
&\quad + [\Delta-\lambda\cdot\alpha\cdot p^*]\cdot\frac{dp^*}{da_\phi} \equiv MU_{a_\phi}(a_\phi,\overline{\omega},a) \\
0 &= p^*\cdot f(\overline{\omega})\cdot[(\overline{\omega}-[a_\phi+a(\overline{\omega})])^2-(\overline{\omega}-a_\phi)^2] \\
&\quad + [\Delta-\lambda\cdot\alpha\cdot p^*]\cdot\frac{dp^*}{d\overline{\omega}} \equiv MU_{\overline{\omega}}(a_\phi,\overline{\omega},a) \\
0 &= p^*\cdot 2\cdot(\omega-[a_\phi+a(\omega)])\cdot f(\omega) \\
&\quad + [\Delta-\lambda\cdot\alpha\cdot p^*]\cdot\frac{dp^*}{da(\omega)} \equiv MU_{a(\omega)}(a_\phi,\overline{\omega},a),\forall\omega
\end{aligned}$$

Several aspects of the solution depend on the sign of $\Delta-\lambda\cdot\alpha\cdot p^*$, which is the planner's marginal utility with respect to p^* at the solution. For this reason, it is worthwhile to analyze the knife edge solution where $\Delta=\lambda\cdot\alpha\cdot p^*$. This borderline policy is denoted $(a_\phi^b,\overline{\omega}^b,a^b(\omega))$. The third F.O.C. implies that $a^b(\omega)=\omega-a_\phi^b,\forall\omega\geq\overline{\omega}^b$. Then, the second F.O.C. implies that $\overline{\omega}^b=a_\phi^b$. Let p^{*b} be the borderline effort level. Then, the first F.O.C. implicitly characterizes a_ϕ^b as follows:

$$
\begin{aligned}
0 &= 2 \cdot p^{*b} \cdot F(a_\phi^b) \cdot [\mathbb{E}_F[W|W < a_\phi^b] - a_\phi^b] + 2 \cdot (1 - p^{*b}) \cdot [\mathbb{E}_F[W] - a_\phi^b] \\
\Leftrightarrow a_\phi^b &= \frac{p^{*b} \cdot F(a_\phi^b) \cdot \mathbb{E}_F[W|W < a_\phi^b] + (1 - p^{*b}) \cdot \mathbb{E}_F[W]}{p^{*b} \cdot F(a_\phi^b) + (1 - p^{*b})} \\
\Leftrightarrow a_\phi^b &= \mathbb{E}_{F,p^{*b}}[W|M = \phi]
\end{aligned}
$$

In other words, a_ϕ^b is exactly the receiver's optimal action given the concealment that a_ϕ^b induces. In fact, the borderline policy always chooses a best response action to each observed message, so it is the same as the unobservable effort equilibrium in the game without commitment.[12]

Now, we will find the conditions under which the borderline policy is the solution. First, notice that the borderline policy does not depend on λ. This policy is optimal when $\Delta = \lambda \cdot \alpha \cdot p^{*b}$, which can be simplified to $\lambda = \frac{\mathbb{E}_F[(W - a_\phi^b)^2 | W \geq a_\phi^b]}{\mathbb{E}_F[u^S(W;r)|W \geq a_\phi^b] - u^S(a_\phi^b;r)} \equiv \overline{\lambda}(r) > 0$. Note that $\overline{\lambda}(0)$ is finite. Since $a_\phi^b \to \mathbb{E}_F[W]$ as $r \to \infty$, it follows that $\overline{\lambda}(r) \to \infty$ as $r \to \infty$. If $\lambda > \overline{\lambda}(r)$, then evidence acquisition should be discouraged relative to the borderline level, so $\Delta < \lambda \cdot \alpha \cdot p^*$. If $\lambda < \overline{\lambda}(r)$, then acquisition has extra incentives, so $\Delta > \lambda \cdot \alpha \cdot p^*$ (this case tends to hold for very risk-averse senders). We will later confirm these previous two statements. This divides the solution into two cases: high λ and low λ.

Rearranging the first F.O.C.,

[12] There may be no finite solution for the borderline policy, and it instead it exists only as the limit $a_\phi \to -\infty$ (intuitively, if there were a minimum action a_L that induced $p^* = 1$, then as $a_\phi^b \to a_L$, both sides converge to a_L). The other limit is never the solution, as the right hand side converges to $\mathbb{E}_F[W]$ as $a_\phi^b \to \infty$. Otherwise, there is a finite solution. Uniqueness of the optimal policy cannot be ensured without more assumptions, but uniqueness is not essential.

$$p^* \cdot [1 - F(\overline{\omega})] \cdot \mathbb{E}_F[a_\phi + a(W) | W \geq \overline{\omega}] + [(1 - p^*) + p^* \cdot F(\overline{\omega})] \cdot a_\phi$$
$$= \mathbb{E}_F[W] + \frac{1}{2} \cdot [\Delta - \lambda \cdot \alpha \cdot p^*] \cdot \frac{dp^*}{da_\phi}$$

The left hand side of this equation is the expected action. If the sender were risk-neutral, $\frac{dp^*}{da_\phi} = 0$, so the expected action is exactly the prior expected state. If the sender is risk-averse, the expected action depends on λ and r. If $\lambda > \overline{\lambda}(r)$, the second term on the right is positive, so the expected action exceeds the expected state. If $\lambda < \overline{\lambda}(r)$, the second term on the right is negative, so the expected action is less than the expected state.

$$0 = p^* \cdot 2 \cdot (\omega - [a_\phi + a(\omega)])$$
$$+ [\Delta - \lambda \cdot \alpha \cdot p^*] \cdot \frac{1}{\alpha} \cdot u^{S'}(a_\phi + a(\omega); r)$$

Evaluating this at the leftmost extreme results in $a_\phi + a(\overline{\omega}) = \overline{\omega} + \frac{1}{2} \cdot \frac{1}{p^*} \cdot [\Delta - \lambda \cdot \alpha \cdot p^*] \cdot \frac{1}{\alpha} \cdot u^{S'}(a_\phi + a(\overline{\omega}); r)$. This pins down the relationship between $a_\phi + a(\overline{\omega})$ and $\overline{\omega}$. If $\lambda > \overline{\lambda}(r)$, this equation implies that $a_\phi + a(\overline{\omega}) < \overline{\omega}$, and if $\lambda < \overline{\lambda}(r)$, it follows that $a_\phi + a(\overline{\omega}) > \overline{\omega}$.

Now, we will look at the implications for more general ω. First, note that $a(\omega)$ is unbounded, because otherwise, the right hand side diverges to ∞. Because of bounded utility, as $\omega \to \infty$, the second term vanishes to 0. Therefore, $a(\omega) \to (\omega - a_\phi)$. That is, the reward function converges to the no commitment reward function. Second, since this equation holds for all ω, it can be differentiated with respect to ω, producing the following equation:

$$0 = p^* \cdot 2 \cdot [1 - a'(\omega)] + [\Delta - \lambda \cdot \alpha \cdot p^*] \cdot \frac{1}{\alpha} \cdot u^{S''}(a_\phi + a(\omega); r) \cdot a'(\omega) \tag{1.4}$$

Since u^S is concave, the second term depends on the sign of $[\Delta - \lambda \cdot \alpha \cdot p^*]$. If $\lambda > \overline{\lambda}(r)$, the second term is positive, so for all $\omega, a'(\omega) \geq 1$. This also implies that $a(\omega) \geq 0, \forall \omega$. If $\lambda < \overline{\lambda}(r)$, the second term is negative, so for all $\omega, a'(\omega) \leq 1$. However, it is not consistent with $a'(\omega) < 0$, as this would make both terms positive. Therefore, this case also satisfies $a(\omega) \geq 0, \forall \omega \geq \overline{\omega}$.

Combining this with the results on the overall expected action allows us to compare a_ϕ with a_ϕ^b in each case. When $\lambda < \overline{\lambda}(r)$, the expected action is lower than in the borderline case, yet the reward function is higher and revelation occurs relatively more often ($a_\phi > \overline{\omega}$). This means that $a_\phi < a_\phi^b$. By a symmetric argument, when $\lambda > \overline{\lambda}$, the expected action is higher despite the reward function being lower and revelation occuring less frequently, so $a_\phi > a_\phi^b$.

The first F.O.C. implies

$$[p^* \cdot F(\overline{\omega}) + (1 - p^*)] \cdot a_\phi = \mathbb{E}_F[W] - p^* \cdot [1 - F(\overline{\omega})] \cdot \mathbb{E}_F[a_\phi + a(W) | W \geq \overline{\omega}] + \frac{1}{2} \cdot [\Delta - \lambda \cdot \alpha \cdot p^*] \cdot \frac{dp^*}{da_\phi}$$

Suppose $\lambda > \overline{\lambda}(r)$, so $\Delta - \lambda \cdot \alpha \cdot p^* < 0$ and $a_\phi + a(\omega) < \omega, \forall \omega \geq \overline{\omega}$. Then

$$
\begin{aligned}
[p^* \cdot F(\overline{\omega}) + (1-p^*)] \cdot a_\phi &> p^* \cdot F(\overline{\omega}) \cdot \mathbb{E}_F[W|W < \overline{\omega}] + (1-p^*) \cdot \mathbb{E}_F[W] \\
&\quad + \frac{1}{2} \cdot [\Delta - \lambda \cdot \alpha \cdot p^*] \cdot \frac{dp^*}{da_\phi} \\
&> p^* \cdot F(\overline{\omega}) \cdot \mathbb{E}_F[W|W < \overline{\omega}] + (1-p^*) \cdot \mathbb{E}_F[W] \\
\Rightarrow a_\phi &> \frac{p^* \cdot F(\overline{\omega}) \cdot \mathbb{E}_F[W|W < \overline{\omega}] + (1-p^*) \cdot \mathbb{E}_F[W]}{p^* \cdot F(\overline{\omega}) + (1-p^*)}
\end{aligned}
$$

Hence, not only is $a_\phi > a_\phi^b$, but a_ϕ is larger than the unconstrained optimal action following an empty message given this sender's equilibrium strategy (which is also higher than a_ϕ^b). Now, suppose $\lambda < \overline{\lambda}(r)$, so $\Delta - \lambda \cdot \alpha \cdot p^* > 0$ and $a_\phi + a(\omega) > \omega, \forall \omega \geq \overline{\omega}$. Then

$$
\begin{aligned}
[p^* \cdot F(\overline{\omega}) + (1-p^*)] \cdot a_\phi &< p^* \cdot F(\overline{\omega}) \cdot \mathbb{E}_F[W|W < \overline{\omega}] + (1-p^*) \cdot \mathbb{E}_F[W] \\
&\quad + \frac{1}{2} \cdot [\Delta - \lambda \cdot \alpha \cdot p^*] \cdot \frac{dp^*}{da_\phi} \\
&< p^* \cdot F(\overline{\omega}) \cdot \mathbb{E}_F[W|W < \overline{\omega}] + (1-p^*) \cdot \mathbb{E}_F[W] \\
\Rightarrow a_\phi &< \frac{p^* \cdot F(\overline{\omega}) \cdot \mathbb{E}_F[W|W < \overline{\omega}] + (1-p^*) \cdot \mathbb{E}_F[W]}{p^* \cdot F(\overline{\omega}) + (1-p^*)}
\end{aligned}
$$

So, we get the opposite result for low λ. If $\lambda < \overline{\lambda}(r)$, then not only is $a_\phi < a_\phi^b$, but a_ϕ is less than the unconstrained optimal action conditional on an empty message for this sender's equilibrium strategy (which is also lower than a_ϕ^b).

Since the low λ solution has a lower a_ϕ and $\overline{\omega}$ and higher rewards $a(\omega)$ than the borderline solution, it must induce a higher p^* than the borderline solution. By a symmetric argument, when λ is high, the solution features p^* lower than the borderline solution. In

order to confirm the two facts we asserted earlier, we must analyze how p^* varies with λ. Implicitly differentiating the F.O.C.s with respect to λ:

$$
\begin{aligned}
0 &= -\alpha \cdot p^* \cdot \frac{dp^*}{da_\phi} + \frac{dMU_{a_\phi}}{da_\phi} \cdot \frac{da_\phi}{d\lambda} + \frac{dMU_{a_\phi}}{d\overline{\omega}} \cdot \frac{d\overline{\omega}}{d\lambda} + \frac{dMU_{a_\phi}}{da(\omega)} \cdot \frac{da(\omega)}{d\lambda}, \forall \omega \\
0 &= -\alpha \cdot p^* \cdot \frac{dp^*}{d\overline{\omega}} + \frac{dMU_{\overline{\omega}}}{da_\phi} \cdot \frac{da_\phi}{d\lambda} + \frac{dMU_{\overline{\omega}}}{d\overline{\omega}} \cdot \frac{d\overline{\omega}}{d\lambda} + \frac{dMU_{\overline{\omega}}}{da(\omega)} \cdot \frac{da(\omega)}{d\lambda}, \forall \omega \\
0 &= -\alpha \cdot p^* \cdot \frac{dp^*}{da(\omega)} + \frac{dMU_{a(\omega)}}{da_\phi} \cdot \frac{da_\phi}{d\lambda} + \frac{dMU_{a(\omega)}}{d\overline{\omega}} \cdot \frac{d\overline{\omega}}{d\lambda} + \frac{dMU_{a(\omega)}}{da(\omega)} \cdot \frac{da(\omega)}{d\lambda}, \forall \omega \\
\Rightarrow & D_\lambda a = \alpha \cdot p^* \cdot H^{-1}(\nabla_a p^*), \forall \omega \\
\Rightarrow & (\nabla_a p^*)^T D_\lambda a = \alpha \cdot p^* \cdot (\nabla_a p^*)^T H^{-1} \nabla_a p^* \leq 0, \forall \omega \\
\Rightarrow & \frac{dp^*}{d\lambda} \leq 0
\end{aligned}
$$

Here, the vector $D_\lambda a$ is the vector of policy derivatives with respect to λ, the vector $\nabla_a p^*$ is the gradient of p^* with respect to the policy variables (holding parameters constant), and the matrix H is the Hessian of the planner's expected utility function, which is negative semidefinite at any interior solution.

Since the sign of $\Delta - \lambda \cdot \alpha \cdot p^*$ at the solution divides the solutions into high incentives with $p^* > p^{*b}$ and low incentives with $p^* < p^{*b}$, the sign of $\frac{dp^*}{d\lambda}$ means that for all $\lambda > \overline{\lambda}(r), \Delta - \lambda \cdot \alpha \cdot p^* < 0$ and for all $\lambda < \overline{\lambda}(r), \Delta - \lambda \cdot \alpha \cdot p^* > 0$. This confirms the earlier assumption that $\overline{\lambda}(r)$ splits the solutions into two categories (high incentive and low incentive) based solely on λ.

Now, we will analyze the role of risk-aversion. Returning to Equation 1.4, the slope of the bonus function converges to 1 everywhere as the sender's utility converges to risk-neutral ($r \to 0$) (Δ does not diverge, because this would imply that the planner's utility diverges, which is either suboptimal or impossible). Otherwise, the solution is qualitatively the same as for a risk-averse sender. Note that $\frac{dp^*}{da_\phi} = 0$ for any policy with a risk-neutral

sender. Looking at the first F.O.C.,

$$\begin{aligned}
0 &= \int_{-\infty}^{\infty} 2 \cdot (\omega - a_\phi) \cdot dF(\omega) + p^* \cdot [\int_{\overline{\omega}}^{\infty} -2 \cdot a(\omega) \cdot dF(\omega)] \\
&\quad + [\Delta - \lambda \cdot \alpha \cdot p^*] \cdot \frac{dp^*}{da_\phi} \\
&= \int_{-\infty}^{\infty} 2 \cdot (\omega - a_\phi) \cdot dF(\omega) + p^* \cdot [\int_{\overline{\omega}}^{\infty} -2 \cdot a(\omega) \cdot dF(\omega)] \\
\Leftrightarrow \mathbb{E}_F[W] &= p^* \cdot [1 - F(\overline{\omega})] \cdot \mathbb{E}_F[a_\phi + a(W) | W \geq \overline{\omega}] \\
&\quad + [(1 - p^*) + p^* \cdot F(\overline{\omega})] \cdot a_\phi
\end{aligned}$$

In other words, the expected action equals the expected state.

As $r \to \infty$, $\overline{\lambda}(r) \to \infty$, so the $\lambda < \overline{\lambda}(r)$ solution holds. Suppose that $p^* > 0$ in the limit. Since the increasing risk-aversion is making the incentives ineffective, a_ϕ must be diverging to $-\infty$. If $\lim_{r \to \infty} p^* < 1$, the planner is getting infinitely low utility, so this is not the limit solution. If $\lim_{r \to \infty} p^* = 1$, then the solution satisfies $a_\phi \to -\infty, \overline{\omega} \to -\infty, a(\omega) \to \omega - a_\phi$. This is the perfect information solution. The planner's expected utility in the limit is $-\frac{1}{2} \cdot \lambda \cdot \alpha$. However, the zero evidence solution, where $a_\phi = \mathbb{E}_F[W], \overline{\omega} = a_\phi$, and $a(\omega) = \omega - a_\phi$, might be better in the limit, giving a utility of $-Var_F(W)$. So, the infinitely risk-averse solution is maximal evidence acquisition when $\frac{1}{2} \cdot \lambda \cdot \alpha < Var_F(W)$, and it is zero evidence acquisition when $\frac{1}{2} \cdot \lambda \cdot \alpha > Var_F(W)$.

□

1.8.8 Proof of 8

Proof. The sender's expected payoff is

$$(1-p\cdot[1-F(\overline{\omega})])\cdot u^S(a_\phi;r)+p\cdot[1-F(\overline{\omega})]\cdot\mathbb{E}_F[u^S(W;r)|W\geq\overline{\omega}]-\frac{1}{2}\cdot\alpha\cdot p^2$$

The sender's F.O.C. yields

$$\begin{aligned} p^* &= \frac{1}{\alpha}\cdot[1-F(\overline{\omega})]\cdot\{\mathbb{E}_F[u^S(W;r)|W\geq\overline{\omega}]-u^S(a_\phi;r)\}\geq 0 \\ \Rightarrow \frac{dp^*}{d\overline{\omega}} &= -\frac{1}{\alpha}\cdot f(\overline{\omega})\cdot\{u^S(\overline{\omega};r)-u^S(a_\phi;r)\}\leq 0 \end{aligned}$$

The planner's problem is

$$\begin{aligned} &\max_{\overline{\omega}}(1-p^*)\cdot\int_{-\infty}^{\infty}-(\omega-a_\phi)^2\cdot dF(\omega) \\ &+p^*\cdot\int_{-\infty}^{\overline{\omega}}-(\omega-a_\phi)^2\cdot dF(\omega)-\lambda\cdot\frac{1}{2}\cdot\alpha\cdot p^{*2} \\ &\text{subject to} \\ &\overline{\omega}\geq a_\phi \end{aligned}$$

Let μ be the Karush-Kuhn-Tucker (KKT) multiplier for the constraint. The KKT conditions are

$$
\begin{aligned}
0 &= p^* \cdot -(\overline{\omega} - a_\phi)^2 \cdot f(\overline{\omega}) \\
&\quad + [\int_{\overline{\omega}}^{\infty} (\omega - a_\phi)^2 \cdot dF(\omega) - \lambda \cdot \alpha \cdot p^*] \cdot \frac{dp^*}{d\overline{\omega}} + \mu \\
0 &= \mu \cdot (\overline{\omega} - a_\phi) \\
\overline{\omega} &\geq a_\phi \\
\mu &\geq 0
\end{aligned}
$$

If $\overline{\omega} = a_\phi$, the first condition implies $\mu = 0$. Therefore, constraint never strictly binds, and we can safely assume that $\mu = 0$.

As $\lambda \to \infty$, it must follow that $p^* \to 0$ (otherwise, the planner's payoff diverges to $-\infty$, which can be improved upon). According to the sender's F.O.C., this is only possible if $\overline{\omega} \to \infty$.

If $\int_{\overline{\omega}}^{\infty} (\omega - a_\phi)^2 \cdot dF(\omega) - \lambda \cdot \alpha \cdot p^* \geq 0$, the marginal utility of $\overline{\omega}$ is negative, so that part of the solution must be $\overline{\omega} = a_\phi$. This condition can be rewritten as

$$
\lambda \leq \frac{\int_{a_\phi}^{\infty} (\omega - a_\phi)^2 \cdot dF(\omega)}{[1 - F(a_\phi)] \cdot \{\mathbb{E}_F[u^S(W; r) | W \geq a_\phi] - u^S(a_\phi; r)\}} \equiv \overline{\lambda}(r)
$$

If the above condition does not hold, then the planner benefits from increasing $\overline{\omega}$ above a_ϕ, making some evidence inadmissible.

Note that $\overline{\lambda}(r) \to \infty$ as $r \to \infty$. So, if the sender is sufficiently risk-averse, restrictions on admissibility should be removed. However, note that the sender's F.O.C. implies that $p^* \to 0$. On the contrary, for lower degrees of risk-aversion, this threshold is finite.

□

1.8.9 Proof of 9

Proof. The sender's solution is

$$p^* = \frac{1}{\alpha} \cdot [1 - F(a_\phi)] \cdot \{\mathbb{E}_F[u^S(W;r)|W \geq a_\phi] - u^S(a_\phi;r)\} > 0$$
$$\frac{dp^*}{da_\phi} = -\frac{1}{\alpha} \cdot [1 - F(a_\phi)] \cdot u^{S\prime}(a_\phi;r) \leq 0$$

The planner's expected utility is

$$(1-p^*) \cdot \int_{-\infty}^{\infty} -(\omega - a_\phi)^2 \cdot dF(\omega) + p^* \cdot \int_{-\infty}^{a_\phi} -(\omega - a_\phi)^2 \cdot dF(\omega) - \lambda \cdot \frac{1}{2} \cdot \alpha \cdot p^{*2}$$

The F.O.C. is

$$\begin{aligned} 0 &= (1-p^*) \cdot 2 \cdot \int_{-\infty}^{\infty} (\omega - a_\phi) \cdot dF(\omega) + p^* \cdot 2 \cdot \int_{-\infty}^{a_\phi} (\omega - a_\phi) \cdot dF(\omega) \\ &\quad + [\int_{a_\phi}^{\infty} (\omega - a_\phi)^2 \cdot dF(\omega) - \lambda \cdot \alpha \cdot p^*] \cdot \frac{dp^*}{da_\phi} \\ \Leftrightarrow a_\phi &= \Big\{(1-p^*) \cdot \mathbb{E}_F[W] + p^* \cdot F(a_\phi) \cdot \mathbb{E}_F[W|W \leq a_\phi] \\ &\quad + \frac{1}{2} \cdot [\int_{a_\phi}^{\infty} (\omega - a_\phi)^2 \cdot dF(\omega) - \lambda \cdot \alpha \cdot p^*] \cdot \frac{dp^*}{da_\phi}\Big\} / [(1-p^*) + p^* \cdot F(a_\phi)] \end{aligned}$$

Given equilibrium message/state distribution $Q(\lambda)$, it follows that $\mathbb{E}_{Q(\lambda)}[W|M = \phi] = \frac{(1-p^*) \cdot \mathbb{E}_F[W] + p^* \cdot F(a_\phi) \cdot \mathbb{E}_F[W|W \leq a_\phi]}{(1-p^*) + p^* \cdot F(a_\phi)}$. If $\lambda > \overline{\lambda}(r) \equiv \frac{\int_{a_\phi}^{\infty} (\omega - a_\phi)^2 \cdot dF(\omega)}{[1-F(a_\phi)] \cdot \{\mathbb{E}_F[u^S(W;r)|W \geq a_\phi] - u^S(a_\phi;r)\}}$, then $a_\phi > \mathbb{E}_{Q(\lambda)}[W|M = \phi]$ (underincentivized). If $\lambda < \overline{\lambda}(r)$, then $a_\phi < \mathbb{E}_{Q(\lambda)}[W|M = \phi]$

(overincentivized). As $r \to \infty$, $\overline{\lambda}(r) \to \infty$, so for high enough r the solution is overincentivized.

With extreme risk-aversion ($r \to \infty$), there are two possibilities to consider: the planner gives up on incentives and sets $a_\phi \to \mathbb{E}_F[W]$, or the planner sends $a_\phi \to -\infty$ to keep effort positive. If the latter is the solution, it must send $p^* \to 1$, or otherwise there is a strictly positive probability of choosing a_ϕ when it is infinitely bad for the planner. In the limit, this solution would yield planner utility $-\lambda \cdot \frac{1}{2} \cdot \alpha < 0$, whereas the zero effort solution would give $-Var_F(W)$.

□

1.8.10 Proof of 10

Proof. The sender's expected utility is

$$[(1-p)+p\cdot F(a_\phi)]\cdot u^S(a_\phi;r)+p\cdot\{[F(a_H)-F(a_\phi)]\cdot\mathbb{E}_F[u^S(W;r)|W\in[a_\phi,a_H)]$$
$$+[1-F(a_H)]\cdot u^S(a_H;r)\}-\frac{1}{2}\cdot\alpha\cdot p^2$$

The (interior) solution is

$$
\begin{aligned}
p^* &= \frac{1}{\alpha} \cdot \{-[1-F(a_\phi)] \cdot u^S(a_\phi; r) + [[F(a_H) - F(a_\phi)] \cdot \mathbb{E}_F[u^S(W; r) | W \in [a_\phi, a_H)] \\
&+ [1-F(a_H)] \cdot u^S(a_H; r)]\} \\
&= \frac{1}{\alpha} \cdot \{[1-F(a_\phi)] \cdot [\mathbb{E}_F[u^S(W; r) | W \in [a_\phi, a_H)] - u^S(a_\phi; r)] \\
&+ [1-F(a_H)] \cdot [u^S(a_H; r) - \mathbb{E}_F[u^S(W; r) | W \in [a_\phi, a_H)]]\} > 0
\end{aligned}
$$

It changes with a_H as follows:

$$
\frac{dp^*}{da_H} = \frac{1}{\alpha} \cdot [1-F(a_H)] \cdot u^{S'}(a_H; r) \geq 0
$$

The planner's expected utility is

$$
\begin{aligned}
&(1-p^*) \cdot \int_{-\infty}^{\infty} -(\omega - a_\phi)^2 \cdot dF(\omega) + p^* \cdot \left[\int_{-\infty}^{a_\phi} -(\omega - a_\phi)^2 \cdot dF(\omega) \right. \\
&\left. + \int_{a_H}^{\infty} -(\omega - a_H)^2 \cdot dF(\omega) \right] - \lambda \cdot \frac{1}{2} \cdot \alpha \cdot p^{*2}
\end{aligned}
$$

Define $\Delta \equiv \int_{a_\phi}^{\infty} (\omega - a_\phi)^2 \cdot dF(\omega) - \int_{a_H}^{\infty} (\omega - a_H)^2 \cdot dF(\omega) \geq 0$. The F.O.C. is

$$
0 = p^* \cdot \int_{a_H}^{\infty} 2 \cdot (\omega - a_H) \cdot dF(\omega) + [\Delta - \lambda \cdot \alpha \cdot p^*] \cdot \frac{dp^*}{da_H} \equiv MU_{a_H}
$$

Note that if $\Delta \geq \lambda \cdot \alpha \cdot p^*$, the F.O.C. can never be satisfied, instead yielding a solution where there is no minimum overall punishment ($a_H \to \infty$). In this solution $p^* = \frac{1}{\alpha} \cdot [1 - F(a_\phi)] \cdot \{\mathbb{E}_F[u^S(W;r)|W \geq a_\phi] - u^S(a_\phi;r)\}$ and $\Delta = \int_{a_\phi}^{\infty} (\omega - a_\phi)^2 \cdot dF(\omega)$. So, the condition for this to be the solution is

$$\lambda \leq \overline{\lambda}(r) \equiv \frac{\int_{a_\phi}^{\infty} (\omega - a_\phi)^2 \cdot dF(\omega)}{[1 - F(a_\phi)] \cdot \{\mathbb{E}_F[u^S(W;r)|W \geq a_\phi] - u^S(a_\phi;r)\}}$$

If λ exceeds this threshold, then $\Delta - \lambda \cdot \alpha \cdot p^* < 0$, and there must be a minimum punishment. As $r \to \infty$, the threshold goes to infinity, so for large enough r the solution must be to have no minimum overall punishment. As $r \to 0$, the threshold converges to a positive, finite number. Therefore, the solution may or may not have a minimum overall punishment, depending on λ.

Implicitly differentiating the F.O.C. with respect to λ,

$$\begin{aligned} 0 &= -\alpha \cdot p^* \cdot \frac{dp^*}{da_H} + \frac{dMU_{a_H}}{da_H} \cdot \frac{da_H}{d\lambda} \\ \Rightarrow \frac{da_H}{d\lambda} &= \alpha \cdot p^* \cdot \frac{dp^*}{da_H} \Big/ \frac{dMU_{a_H}}{da_H} \leq 0 \end{aligned}$$

In words, the more the planner dislikes acquisition costs, the more severe a minimum overall punishment he should commit to.

□

1.8.11 Proof of 11

Proof. Given the sender's equilibrium strategy, the sender's expected payoff is

$$
\begin{aligned}
&[(1-p)+p\cdot F(a_\phi)]\cdot u^S(a_\phi;r)\\
&+p\cdot\{[F(\overline{\nu}-B)-F(a_\phi)]\cdot\mathbb{E}_F[u^S(W;r)|W\in[a_\phi,\overline{\nu}-B)]\\
&+[F(\overline{\nu})-F(\overline{\nu}-B)]\cdot u^S(\overline{\nu}-B;r)\\
&+[1-F(\overline{\nu})]\cdot\mathbb{E}_F[u^S(W;r)|W\geq\overline{\nu}]\}-\frac{1}{2}\cdot\alpha\cdot p^2
\end{aligned}
$$

The F.O.C. is

$$
\begin{aligned}
p^*=\frac{1}{\alpha}\cdot\{&-[1-F(a_\phi)]\cdot u^S(a_\phi;r)\\
&+\{[F(\overline{\nu}-B)-F(a_\phi)]\cdot\mathbb{E}_F[u^S(W;r)|W\in[a_\phi,\overline{\nu}-B)]\\
&+[F(\overline{\nu})-F(\overline{\nu}-B)]\cdot u^S(\overline{\nu}-B;r)+[1-F(\overline{\nu})]\cdot\mathbb{E}_F[u^S(W;r)|W\geq\overline{\nu}]\}\}\geq 0
\end{aligned}
$$

Differentiating with respect to B:

$$
\frac{dp^*}{dB}=-\frac{1}{\alpha}\cdot[F(\overline{\nu})-F(\overline{\nu}-B)]\cdot u^{S'}(\overline{\nu}-B;r)\leq 0
$$

The planner's expected payoff is

$$(1-p^*)\cdot\int_{-\infty}^{\infty}-(\omega-a_\phi)^2\cdot dF(\omega)+p^*\cdot\{\int_{-\infty}^{a_\phi}-(\omega-a_\phi)^2\cdot dF(\omega)$$
$$+\int_{\overline{\nu}-B}^{\overline{\nu}}-(\omega-\overline{\nu}+B)^2\cdot dF(\omega)\}-\lambda\cdot\frac{1}{2}\cdot\alpha\cdot p^{*2}$$

Define $\Delta\equiv\int_{a_\phi}^{\infty}(\omega-a_\phi)^2\cdot dF(\omega)-\int_{\overline{\nu}-B}^{\overline{\nu}}(\omega-\overline{\nu}+B)^2\cdot dF(\omega)$. For small B, Δ is positive, but for large enough B, it is negative. The F.O.C. is

$$0=-p^*\cdot\int_{\overline{\nu}-B}^{\overline{\nu}}2\cdot(\omega-\overline{\nu}+B)\cdot dF(\omega)+[\Delta-\lambda\cdot\alpha\cdot p^*]\cdot\frac{dp^*}{dB}$$

The first term on the right hand side is always negative. If $\Delta-\lambda\cdot\alpha\cdot p^*\geq 0$, then the right hand side is always negative. This means the solution is at the boundary of $B=0$ (no minimum conditional punishment). Conditional on $B=0$, $\Delta-\lambda\cdot\alpha\cdot p^*\geq 0$ is equivalent to

$$\lambda\leq\frac{\int_{a_\phi}^{\infty}(\omega-a_\phi)^2\cdot dF(\omega)}{[1-F(a_\phi)]\cdot\{\mathbb{E}_F[u^S(W;r)|W\geq a_\phi]-u^S(a_\phi;r)\}}\equiv\underline{\lambda}(r)$$

This is the same threshold for λ as in the section on minimum overall punishments. If λ lies below this threshold, the planner wants to maximize acquisition incentives as much as possible, meaning $B=0$. Otherwise, $\Delta-\lambda\cdot\alpha\cdot p^*<0$ at the solution.

Setting $B=\frac{\overline{\nu}}{a_\phi}$, there must exist λ large enough that the right hand side is positive. Call the smallest such value $\overline{\lambda}(r)$.

As $r \to \infty$, the $\underline{\lambda}(r) \to \infty$, so for sufficiently risk-averse senders, there should be no minimum conditional punishment.

□

1.8.12 Competing Senders

In order to comment on heterogeneous risk aversion within a single trial, we consider the case of two competing senders: H and L. The former has strictly increasing utility, and the latter has strictly decreasing utility. Both are strictly risk-averse. Their payoffs over receiver actions are $u^H(a; r^H)$ and $u^L(a; r^L)$. They simultaneously choose efforts p^H and p^L, and they simultaneously receive signals x^H and x^L, which are independent conditional on the state. Each sender i faces a cost function $c^i(p^i)$ with the same properties as in the single sender case (increasing, convex, etc.). Whenever a state ω is revealed by one of the parties, the receiver chooses $a = \omega$. When neither party reveals evidence, the receiver takes action a_ϕ, which in equilibrium depends on the strategies of the senders.

The main objective of this section is to show that a version of 5 holds with two senders. This result drives the most important secondary results in the commitment section, so the conclusions of that section will also apply to the case of competing senders. We will consider only the case of $r^H \to \infty$; the other case follows from symmetry.

Proposition 12. *For every $\overline{p} > 0$ and every r^L, there exists $\overline{r}$ such that for all $r^H \geq \overline{r}$, every equilibrium of the game induced by (r^H, r^L) features $p^H < \overline{p}$.*

Proof. Consider any unboundedly increasing sequence of risk aversion parameters $\{r_n^H\}_{n=1}^\infty$. Consider also any corresponding sequence of equilibria, where $\{(p_n^L, \sigma_n^L)\}_{n=1}^\infty$ is the sequence of L's equilibrium strategies. Let M_n^L be L's message (a random variable) in the n'th equilibrium. If $M_n^L = \omega$, then H cannot change the receiver's action by presenting evidence. If $M_n^L = \phi$, then evidence can still have an effect.

Case: $q = 0$

Let G_n be the joint distribution of the receiver's action and L's message in the n'th equilibrium, conditional on H succeeding in acquiring evidence (i.e., $x^H = \omega$). Let $\hat{a}_n$ be the action taken when $M_n^L = M_n^H = \phi$. Then in the n'th equilibrium,

$$Pr(M_n^L = \phi) \cdot \left\{ \mathbb{E}_{G_n}[u^H(A; r_n^H) | M_n^L = \phi] - u^H(\hat{a}_n; r_n^H) \right\} = c^{H'}(p_n^H)$$

Since $Pr(M_n^L = \phi) \leq 1$ and the benefit of acquisition is greatest when $M_n^H = \phi$ receives the maximal punishment, the left hand side is bounded above by

$$\mathbb{E}[u^H(a(W); r_n^H)] - u^H(a(\omega_0); r_n^H)$$

The same argument as in the proof of 5 shows that this upper bound $\to 0$ as $n \to \infty$, and thus $p^H \to 0$ in equilibrium.

Case: $q = 1$

Sender H always has a threshold disclosure strategy in equilibrium. When $M_n^L = \omega$, the receiver will always choose the same action, regardless of H's message. When $M_n^L = \phi$, the lowest possible action is that resulting from $M_n^H = \phi$. Regardless of (p_n^L, σ_n^L), decreasing p_n^H increases the lowest possible receiver action. Since decreasing p_n^H also lowers acquisition costs, and in the limit, the lowest possible action determines H's payoff (see 1), equilibrium effort $p_n^H \to 0$ as $n \to \infty$.

□

To see the intuition of 12, look at the problem of effort choice for sender H. If L presents evidence, then H's effort is wasted. If L does not present evidence, then the source of value of effort is the same as with only one sender: it increases the probability that H reveals states leading to actions higher than the action taken after concealment. For any strategy L chooses, H faces the same problem as a single sender, restricted to a subspace of the joint state/signal space. Competition may affect the magnitude of H's effort in equilibrium, but the qualitative features of equilibrium are unchanged.

1.8.13 Private Information

Consider a simple example where there is a $\frac{1}{2}$ probability that the state is $\omega = 0$ and a $\frac{1}{2}$ probability that the state has some continuous, full support distribution F_H on $[0, 1]$. The receiver has quadratic loss utility, so he always chooses action equal to the expected state. The added complication is that the sender now observes a signal at the very beginning of the timeline. This signal has two possible realizations, a low one and a high one. The signal realization is called the "type" of the sender. The low type learns that the state is $\omega = 0$. The high type learns that the state is distributed according to F_H. These strong distributional assumptions make the example very tractable, but they are not essential. The sender's choice of effort is a function of the observed signal: p_H and p_L. For simplicity, we will also assume that $c'(p) \to \infty$ as $p \to 1$, so there is no chance of a corner solution with $p = 1$.

When the effort is unobserved, there is an equilibrium default action a_ϕ. In a pooling equilibrium, this is the same as the default action after observing the pooling effort level. In a separating equilibrium, observing the effort level informs the receiver of the sender's private signal. This adjusts the receiver's beliefs up or down, implying a higher (a_ϕ^H) or lower (a_ϕ^L) default action.

Lemma 2. *In any separating equilibrium, $p_L = 0$ and $a_\phi^L = 0$.*

Proof. Suppose in equilibrium $p_L > 0$. In the event that effort is observed, the receiver knows that the state is $\omega = 0$, resulting in the lowest possible action. A deviation to $p_L = 0$ cannot induce a worse action and saves on acquisition costs. In the event that effort is not observed, the sender will always conceal, because she knows that any evidence will reveal $\omega = 0$. Since the sender always sends the empty message, the acquisition effort is wasted, so a deviation to $p_L = 0$ is strictly beneficial. Since the deviation is beneficial in either event, this is not an equilibrium. □

Lemma 3. *In any separating equilibrium,* $p_H \geq p^{min}(r) \equiv c^{-1}\left(q \cdot \left[u^S(a_\phi^H; r) - u^S(0; r)\right]\right)$.

Proof. In a separating equilibrium, p_H must be high enough that the low type sender does not want to pose as the high type sender. The low type sender gets payoff

$$q \cdot u^S(0; r) + (1 - q) \cdot u^S(a_\phi; r)$$

By choosing effort p_H, the low type sender gets payoff

$$q \cdot u^S(a_\phi^H; r) + (1 - q) \cdot u^S(a_\phi; r) - c(p_H)$$

The sender will not have an incentive to deviate in the low state if

$$p_H \geq c^{-1}\left(q \cdot \left[u^S(a_\phi^H; r) - u^S(0; r)\right]\right)$$

□

As is typical of signaling games, there is a huge number of equilibria. We will restrict

attention to the more realistic low observability case and use the Intuitive Criterion (Cho and Kreps (1987)) to select an equilibrium. In this case, only separating equilibria survive.

Proposition 13. *No pooling equilibrium survives the Intuitive Criterion when q is low.*

Proof. Consider an equilibrium where both types of sender choose effort level p^*. The equilibrium payoff for the low type is

$$u^S\left(a_\phi; r\right) - c(p^*)$$

An optimistic deviation (receiver believes the sender is a high type after observing) to higher effort $p > p^*$ gives the low type payoff

$$q \cdot u^S(a_\phi^H; r) + (1-q) \cdot u^S\left(a_\phi; r\right) - c(p)$$

This is higher than the equilibrium payoff for small p, but the sender must be indifferent to the deviation at sufficiently high p. Define p' as the level of effort that achieves this indifference:

$$u^S\left(a_\phi; r\right) - c(p^*) = q \cdot u^S(a_\phi^H; r) + (1-q) \cdot u^S\left(a_\phi; r\right) - c(p') \tag{1.5}$$

For the high type, the pooling equilibrium payoff is

$$p^* \cdot [1 - a_\phi] \cdot \mathbb{E}_{F_H}[u^S(W; r) | W > a_\phi] + [(1-p^*) + p^* \cdot a_\phi] \cdot u^S\left(a_\phi; r\right) - c(p^*)$$

If the high type deviates to effort p', the receiver after observing this will know that the sender is a high type. The payoff from the high type deviating to p' is

$$
\begin{aligned}
&q \cdot \left\{ p' \cdot [1 - a_\phi^H] \cdot \mathbb{E}_{F_H}[u^S(W;r) | W > a_\phi^H] + [(1-p') + p' \cdot a_\phi^H] \cdot u^S\left(a_\phi^H; r\right) \right\} \\
&+ (1-q) \cdot \left\{ p' \cdot [1 - a_\phi] \cdot \mathbb{E}_{F_H}[u^S(W;r) | W > a_\phi] + [(1-p') + p' \cdot a_\phi] \cdot u^S\left(a_\phi; r\right) \right\} \\
&- c(p')
\end{aligned}
$$

Substituting using Equation 1.5:

$$
\begin{aligned}
&q \cdot u^S\left(a_\phi; r\right) \\
&+ q \cdot p' \cdot [1 - a_\phi^H] \cdot \left[\mathbb{E}_{F_H}[u^S(W;r) | W > a_\phi^H] - u^S(a_\phi^H; r) \right] \\
&+ (1-q) \cdot \left[p' \cdot [1 - a_\phi] \cdot \mathbb{E}_{F_H}[u^S(W;r) | W > a_\phi] + [(1-p') + p' \cdot a_\phi] \cdot u^S\left(a_\phi; r\right) \right] \\
&- c(p^*)
\end{aligned}
$$

For small q, the first two terms are insignificant, and the last two terms exceed the high type's pooling equilibrium payoff. Therefore, no pooling equilibrium survives the Intuitive Criterion when q is low.

□

Now, we analyze separating equilibria. For any separating equilibrium, there is a separation constraint requiring that the high type exert enough effort that the low type will not want to imitate the high type. In the notation that is to follow, the separation constraint is written as $p_H \geq p'$.

Proposition 14. *Any separating equilibrium satisfies the Intuitive Criterion and the following statements hold:*

1. *Conditional on the separation constraint not binding, for every $\overline{p} > 0$, there exists $\overline{r}$ such that for all $r \geq \overline{r}$, every separating equilibrium of the game induced by r features $p < \overline{p}$,*

2. *There exists $\overline{r}$ such that for all $r \geq \overline{r}$, the separation constraint binds in every separating equilibrium. Moreover, let $\mathcal{P}'$ be the set of all equilibrium values of p' across all r. Then, 1 is a limit point of $\mathcal{P}'$.*

14 contains two conflicting parts. Part 1 is essentially the same as the main result of this paper, 5. It says that risk aversion decreases evidence acquisition in the limit. However, it is conditional on the separation constraint not binding (i.e., signaling concerns not dominating). Part 2 is a contrary result, showing that the separation constraint does eventually bind, and that high risk aversion can sometimes result in very high equilibrium effort levels. The proof is presented below.

Proof. In a separating equilibrium, $p_L = 0$. For the low type, the equilibrium payoff is $q \cdot u^S(0;r) + (1-q) \cdot u^S\left(a_\phi; r\right)$. An optimistic deviation to effort p for the low type yields payoff $q \cdot u^S(a_\phi^H; r) + (1-q) \cdot u^S\left(a_\phi; r\right) - c(p)$. Define p' by

$$\begin{aligned} & q \cdot u^S(0;r) + (1-q) \cdot u^S\left(a_\phi; r\right) = q \cdot u^S(a_\phi^H; r) + (1-q) \cdot u^S\left(a_\phi; r\right) - c(p') \\ \Leftrightarrow & c(p') = q \cdot [u^S(a_\phi^H; r) - u^S(0;r)] \end{aligned}$$

In order to dissuade the low type sender from imitating the high type sender, it must be that $p_H \geq p'$ (and this also satisfies the Intuitive Criterion). Effort p_H gives high type payoff

$$
\begin{aligned}
&q \cdot \left\{ p_H \cdot [1 - a_\phi^H] \cdot \mathbb{E}_{F_H}[u^S(W;r) | W > a_\phi^H] + [(1-p_H) + p_H \cdot a_\phi^H] \cdot u^S\left(a_\phi^H; r\right) \right\} \\
&+ (1-q) \cdot \left\{ p_H \cdot [1 - a_\phi] \cdot \mathbb{E}_{F_H}[u^S(W;r) | W > a_\phi] + [(1-p_H) + p_H \cdot a_\phi] \cdot u^S\left(a_\phi; r\right) \right\} \\
&- c(p_H)
\end{aligned}
$$

Any equilibrium effort level $p_H > p'$ is given by the first-order condition:

$$
\begin{aligned}
c'(p_H) =& q \cdot [1 - a_\phi^H] \cdot \left\{ \mathbb{E}_{F_H}[u^S(W;r) | W > a_\phi^H] - u^S\left(a_\phi^H; r\right) \right\} \\
&+ (1-q) \cdot [1 - a_\phi] \cdot \left\{ \mathbb{E}_{F_H}[u^S(W;r) | W > a_\phi] - u^S\left(a_\phi; r\right) \right\} \\
&+ q \cdot [(1-p_H) + p_H \cdot a_\phi^H] \cdot u^{S\prime}(a_\phi^H; r) \cdot \frac{da_\phi^H}{dp_H}
\end{aligned}
$$

Note that there are upper bounds for a_ϕ and a_ϕ^H (corresponding to effort 0). This implies that the infimum of the measure of absolute risk aversion evaluated across all possible values of a_ϕ and a_ϕ^H is diverging to infinity as $r \to \infty$. Using 1, the first two terms converge to 0 as $r \to \infty$. The third term is always ≤ 0 (default actions are always decreasing in observed effort). Therefore, $p_H \to 0$ as $r \to \infty$.

However, in the limit, p_H must eventually bind at p' by the following argument. Recall that the equation for p' is $c(p') = q \cdot [u^S(a_\phi^H; r) - u^S(0; r)]$. Suppose the separation constraint does not bind for all large r: for all $\overline{r}$, there exists $r > \overline{r}$ such that $p_H > p'$ in equilibrium. Then it is possible to construct an unboundedly increasing sequence of values for r where the constraint does not bind. In this sequence, $p_H \to 0$, and a_ϕ^H converges to its upper bound. Using 1, $q \cdot [u^S(a_\phi^H; r) - u^S(0; r)] \to \infty$, so $p' \to \infty$. This is a contradiction to $p_H > p'$, so there exists $\overline{r}$ such that for all $r > \overline{r}$, $p_H = p'$ in the separating equilibrium.

Suppose 1 is not a limit point of p' as $r \to \infty$. Then, 0 is not a limit point of a_ϕ^H.

In other words, there exists $\epsilon > 0$ such that $a_\phi^H > \epsilon$ for all sufficiently high r. This implies that $q \cdot [u^S(a_\phi^H; r) - u^S(0; r)] \to \infty$ as $r \to \infty$, so $p' \to 1$. This is a contradiction, so 1 is a limit point of p'.

□

Chapter 1, in full, is currently being prepared for submission for publication of the material. Giffin, Erin; Lillethun, Erik. The dissertation author is the co-author of this material.

Chapter 2

Identity Formation, Gender Differences, and the Perpetuation of Stereotypes

2.1 Introduction

Gender differences in economic decisions are well-documented and span many important dimensions of economic choices. Men and women differ in their consumption and savings behaviors (LIMRA, 2016), human capital investments (Ceci et al., 2014), choice of college major (National Center for Education Statistics, 2016), and occupational choice (Sapienza et al., 2009). At the household level, there are gender differences in the division of labor within the household (Bertrand et al., 2015) and expenditure on children (Thomas, 1990). Gender differences have also been documented in economic outcomes, with the greatest attention on the gender wage gap (e.g., Bertrand et al., 2010). While the results from observational data have established empirical facts, they frequently do not provide an explanation for these differences. However, understanding the mechanisms behind these differences is important both to interpret and understand results from observational data as well as to inform the optimal policy response. Experimentalists have studied gender differences in laboratory settings to examine potential mechanisms in a controlled environment. Experimental results that find that women are less competitive than men

have been used to explain gender differences in occupational choice, namely why there are not more women in top executive positions (Saccardo et al., 2017). Results on how women are less likely to negotiate have been used to explain part of the gender gap in earnings (Babcock and Laschever, 2003). Results that women are more likely to accept requests have been used to explain why women are more likely than men to complete non-promotional tasks at work (Babcock et al., 2017).

There are two potential explanations for these results: either men and women face different constraints, or men and women differ in terms of fundamentals (i.e., men and women have different preferences or are different types). In anonymous laboratory settings, differences in observable choices are most often attributed to differences in preferences, because constraints that individuals face outside the laboratory should not apply (e.g., Croson and Gneezy, 2009).

In this paper, I propose a novel mechanism that can explain gender differences in anonymous lab settings without assuming different preferences. I propose that external costs, which are different for men and women, become internalized over time through habit formation. As a result, individuals will adhere to behaviors dictated by social norms even when no one is watching. In this paper, I focus on altruistic choices, as this is the focus of a large portion of experimental papers on gender differences (see, for example, Bolton and Katok, 1995; Eckel and Grossman, 1998; Andreoni and Vesterlund, 2001). I then test the model in two ways: I first conduct an empirical analysis to test the model's predictions using existing experimental datasets where gender data were collected but never analyzed, and second I design and implement a new experiment as a direct test of the model's mechanism. Using both of these methods, I find empirical support for the model's validity.

My theoretical model begins with the assumption that men and women are identical in both their preferences (utility functions) and types. In the model, a decision-maker,

who is either a man or a woman, chooses to act either selfishly or fairly, and this choice and their gender is observed. Individuals care about their own consumption as well as how others view them. Based on the decision-maker's choice, observers make inferences about the decision maker's character. I show that if there is at least one observer who draws harsher inferences about a female decision-maker's character if she chooses to act selfishly, then women will be more likely to behave generously in equilibrium.[1] The model thus predicts that stereotypes will perpetuate: men and women behave differently ex post because of the stereotype, even though they were identical ex ante.

I then extend this to a multi-period model and allow individuals to endogenously form gender identities. Through habit formation, as decision-makers behave in a way that is stereotypical of their gender, the association between those behaviors and their own gender strengthens. Through identity formation, individuals internalize external constraints. I show that after gender identities are formed, gender differences in behaviors will persist, even after choices are no longer observed. So although initial group differences were driven by observers' beliefs about men and women, these differences will be perpetuated in the long-run through identity formation.

I then conduct an empirical analysis using existing experimental data where gender data were collected but never analyzed and find evidence that is consistent with the model's predictions. Specifically, I find that women are more generous than men when their decision is observed, even when given the opportunity to hide selfish actions, and that women are significantly more generous than men in an anonymous dictator game where they are asked to give a particular allocation. I also find an interesting secondary result: that although these datasets were not collected with the intention of examining gender differences, the results of the papers that originally used these datasets were partly or entirely driven by only one gender (men in one and women in the other). That is, I find that the results

[1] I show that this is also true even if all observers draw identical inferences for both genders, but women anticipate that there is at least one observer who will draw harsher inferences against them.

of the original papers were only statistically significant because either only men or only women were responsive to the experimental treatment and the result was strong enough to make the pooled result statistically significant.

I finally design and implement an experimental test of the model's mechanism. In the experiment, subjects made a series of decisions on how to allocate $30 between themselves and their partner. To generate external constraints, in some decisions subjects' choices were perfectly observed by others in the experiment, while in others, subjects had plausible deniability. For these decisions, there was a chance that subjects could not make a choice and an allocation where they kept everything was made for them. This offered subjects an opportunity to hide selfish actions, because if others in the experimental session saw that the subject was allocated everything, they could not be sure if the subject made this choice or if this choice was made for them.

Experimental treatments varied in the order subjects made decisions. Subjects' first choice either offered no opportunity for plausible deniability (high external constraint) and this opportunity increased in subsequent decisions or their first choice offered the greatest opportunity for plausible deniability (low external constraint) and this opportunity decreased in subsequent decisions. I find evidence of persistence of behaviors, as I find that subjects' decisions are relatively stable over the series of decisions. I also find that by imposing stricter constraints on subjects' initial action, male subjects make more generous allocations and continue to behave similarly to women even in later decisions when these constraints are relaxed. Specifically, I find that when subjects' first decision has low external constraints, at every level of nature intervening, women are more likely to choose equal allocations than men. However, by simply changing the order of decisions so early decisions have higher external constraints, I mitigate gender differences, as men and women are equally likely to choose a 50-50 split of the pie in this treatment.

This paper provides two important contributions to the gender differences literature.

I first introduce a novel mechanism for understanding gender differences. This is the first model (to the best of my knowledge) that shows persistent gender differences without assuming differences in fundamentals. I also provide additional evidence of gender differences, even in contexts where the researchers were not looking for them. This suggests gender differences, and more largely adherence to social norms, may be more prevalent than we realize.

This paper additionally contributes to the literature on social norms and social prescriptions. These two are largely viewed as distinct, with the former relating to behaviors that are externally punished if not followed (e.g., Akerlof, 1976; Kandori, 1992; Cole et al., 1992) and the latter relating to behaviors that are self-enforced (e.g., Akerlof and Kranton, 2000; Huang and Wu, 1994). I connect these two literatures, as I propose a mechanism where one is generated by internalizing the other. This also suggests a powerful way in which social norms can perpetuate, as eventually external enforcement is no longer necessary for an individual to continue adhering to the norm.

Relatedly, this paper contributes to the literature on identity economics. There is a well-established literature (beginning with Akerlof and Kranton (2000)) on identity economics–the idea that individuals have an identity and derive disutility from taking an action inconsistent with that identity. While this mechanism makes good predictions for many behaviors, it does not address how these identities may form. It assumes that an individual is endowed with both a group membership and an identity with that group and does not want to deviate from the behavioral norms associated with that group. My model, in contrast, takes a step back. It does not assume that identities are endowed, but rather are endogenously formed. I assume that group membership is randomly assigned, but then individuals are incentivized to behave in ways consistent with the norms of their group membership. As agents continue taking actions consistent with their group, the association between themselves and the behaviors associated with their group strengthens through

habit formation. Then, over time, gender identities are solidified. After this point, agents will continue to act in accordance with the prescriptions of their group membership, even if actions are not observed (so there are no external incentives for adhering to the social norm).

The most closely related paper in the theoretical literature is Coate and Loury (1993). Coate and Loury determine that even if two identifiable groups are identical ex ante, an affirmative action policy can create a situation in which employers correctly perceive the groups to be unequally productive ex post. This relates to my model in that both Coate and Loury (1993) and I are able to generate differences in observable behaviors without assuming differences in fundamentals about the groups. The most important distinction between their model and my own is that in their model, differences only persist as long as observers (in their model, employers) are able to observe an individual's group membership. If employers were not able to observe a potential employee's group membership, groups would behave identically. In my model, because of the addition of habit formation, I show that group differences can persist even when observers cannot observe an individual's group membership.

With respect to the experimental literature, my empirical analysis is most closely related to Andreoni and Vesterlund (2001) and DellaVigna et al. (2013). Both of these papers re-analyze an existing dataset and test for gender differences. Andreoni and Vesterlund find that men are more sensitive to the price of giving, while women appear more egalitarian, even when giving is expensive. DellaVigna et al. find that men and women are equally generous in a door-to-door solicitation, but that women become less generous when it is easy to avoid the solicitor. While both papers report significant gender differences, each of these papers re-analyzes only one dataset. In this paper I analyze multiple datasets, which allows me to come to different conclusions than one of these papers. Notably, DellaVigna et al. conclude from their analysis that women are more likely to

be on the margin of giving, and are therefore more sensitive to experimental treatments. Using a larger number of datasets, I do not find support for this claim, as I find that men were more sensitive to experimental treatments in one of datasets I analyze as well as in my own experiments.

The paper proceeds as follows: In Section 2.2 I construct and analyze the model and develop a set of testable predictions. Section 2.3 presents the empirical analysis. Section 2.4 presents the experimental design, and the experimental results are presented in Section 2.5. Section 2.6 concludes. All proofs appear in the Appendix.

2.2 Model

I develop a model to analyze an individual's decision to make altruistic choices. An individual may behave altruistically either because they care about fairness or because they desire others to perceive them as fair. In the model, individuals make a choice–to act either selfishly or fairly–and this choice is observed. Observers, after seeing the individual's choice make an inference about their character, which is unobservable. Individuals may then act fairly because they inherently care about fairness to varying degrees and because they care about the inferences others make about their character. Individuals' gender is visible and observers may form different inferences based on the individual's gender. These different inferences provide different constraints for men and women, causing them to behave differently. Throughout an individual's lifetime, they continue to face these same types of choices. As they continue to do so, they begin to internalize these different constraints. Eventually, individuals begin to self-enforce these mechanisms as these constraints become internalized.

This model thus shows how gender differences can be perpetuated, as I show that even when members of the two groups are identical ex ante, if there exists a stereotype that influences observers' beliefs about the groups, group members may behave in ways

consistent with this stereotype in equilibrium. Then, due to habit formation, these group difference will persist, even after choices are no longer observed.

2.2.1 Setup

Two players–a decision-maker (D) and a receiver (R)–split a prize normalized to have unit value. D transfers $x \in [0,1]$ to R and consumes $c = 1-x$. Decision-makers belong to one of two groups and have label $L \in \{M,W\}$ that discloses group membership. L is visible, making D's group membership public information. Decision-makers are differentiated by a parameter, t, that indicates the importance D places on fairness; t is D's private information. The distribution of t has full support over the interval $[0,\bar{t}]$. K denotes the CDF, and I define K_T as the CDF obtained from T, conditioning on $T \leq t$. Groups and types (t) are uncorrelated, so groups are identical ex ante.

D cares about his own prize (c) as well as his social image (s), as perceived by an Audience (A), which includes R. $F(c,s)$ is a utility function of c and s. It is unbounded in both arguments, twice continuously differentiable, strictly increasing, and strictly concave in c. The decision-maker also cares about fairness, which is determined by the extent to which the outcome departs from the fair alternative, x^F.[2] D's total payoff is:[3]

$$U(x,s,t) = F(1-x,s) + tG(x-x^F)$$

G is twice continuously differentiable, strictly concave, and reaches a maximum at zero. D's social image, s, depends on A's perception of D's fairness. I normalize s so that if A is certain D's type is $\hat{t}$, then D's social image is $\hat{t}$. Φ denotes the CDF that represents A's belief about D's type and $S(\Phi)$ is the associated social image. A forms an inference Φ

[2] x^F is most commonly $\frac{1}{2}$, but I allow it to be a free parameter for generality.

[3] This utility function was originally introduced by Andreoni and Bernheim (2009). They were the first to propose that individuals may act generously because they care about being perceived as fair.

about t after observing x and L. S is continuous and satisfies $S(\Phi') > S(\Phi'')$ if Φ' first-order stochastically dominates (FOSD) Φ''. One possible functional form that the social image may take is $\mathbb{E}_D[\mathbb{E}_A(t)]$, so D's social image is her expectation of A's expectation of her type.

I allow audience members to be heterogeneous and for audience members and the decision-maker to hold non-common priors. The decision-maker does not observe the inference directly, but she knows that A will judge her based on x, so she accounts for this effect of her choice on A's inference. I restrict attention to pure strategy equilibria.

2.2.2 One Period Model

I first analyze the model where the game lasts only one period. For simplification, I restrict the decision-maker's choice to $x \in \{0, x^F\}$. Since the decision-maker's choice is binary but there is a continuum of types, this precludes perfect separation. The following lemma shows that there is a threshold type, t_L^*, and all types above this threshold will choose to transfer the fair allocation while all types below the threshold will choose to transfer zero.

Lemma 4. *There exists t^* such that $\forall t \geq t^*$, D chooses $x = x^F$ and $\forall t < t^*$, D chooses $x = 0$.*

I first examine the case where all audience members hold the correct belief that groups are identical and that the decision-maker knows that A holds these beliefs. In this case, the threshold type will be the same across groups, as the next result shows.

Proposition 1. *Let t_W^* denote the threshold type for group W and t_M^* denote the threshold type for group M. If all audience members believe $K(t;M) = K(t;W) = K$ and $\Phi(t;W,x) = \Phi(t;M,x) = \Phi$ and these beliefs are common knowledge, then $t_W^* = t_M^* = t^*$.*

When all audience members know the true distribution of types and the decision-maker knows that they hold correct beliefs, then there will be no differences in group behavior. However, if even one audience member holds an incorrect belief about the groups, this result may break down.

I define a type of belief where audience members, upon observing a decision-maker choose the selfish allocation, draw harsher inferences about the decision-maker if she comes from group W.

Definition 2. *Belief B1: Belief such that* $\Phi(t; M, x = 0)$ *FOSD* $\Phi(t; W, x = 0)$.

There are multiple conditions that would lead audience members to draw inferences consistent with *B1*. Sufficient conditions for *B1* include:

1. Distorted prior beliefs: An audience member believes that t is drawn from two different distributions and while the distribution of t still has full support over $[0, \bar{t}]$ for each group, he believes $K(t; W) > K(t; M)$ for $0 < t < t^*$ and $K(t; W) < K(t; M)$ for $t^* < t < \bar{t}$ ($K(t; W) = K(t; M)$ for $t \in \{0, t^*, \bar{t}\}$). These distorted beliefs imply that the audience member believes that members of group W are concentrated at the tails of the distribution, so members of group W are more likely to either be very low types or very high types.

2. Biased inferences: An audience member is biased (implicitly or explicitly) against group W, and so upon observing $x = 0$ and that D is a member of group W, he over-updates (i.e., puts more weight on the signal, and as a result his posterior beliefs about a decision-maker from group W are less favorable than his posterior about a decision-maker from group M). This means that after observing the same signal, the audience member arrives at different inferences about the decision-maker based on group membership.

Holding belief *B1* means that that the audience member holds incorrect beliefs about the decision-maker's type. I also allow for members of the audience to hold incorrect beliefs about the decision-maker's preferences. In this case, audience members misspecify the decision-maker's utility function. Specifically, audience members believe decision-makers from the two groups care about social image to different degrees and place different weights (α) on the social image in their utility.

Definition 3. *Belief B2: Belief that* $U_L = F(1-x, \alpha_L s) + tG(x - x^F)$, $\alpha_L > 0$ *and* $\alpha_W > \alpha_M$

If an audience member holds this belief, then he believes that members of group W care more about social image, and thus have a stronger preference for being perceived as fair, than members of group M.[4] If any audience members hold incorrect beliefs about either the decision-maker's type or the decision-maker's preferences, then group differences will arise, as the next result shows.

Proposition 2. *If there exists at least one member of A who holds belief B1 or belief B2 and D knows this, then* $t^*_W < t^*_M$.

The above result illustrates that it is sufficient for just one member of the audience to hold incorrect beliefs to result in group differences. The next result shows that even if all audience members hold correct beliefs, this is not sufficient to guarantee no differences between groups. As the next result shows, even if all audience members hold correct beliefs about both decision-makers' types and preferences, but the decision-maker believes that at least one audience member holds misspecified beliefs, then group differences will arise.

Proposition 3. *If all members of A believe* $\Phi(t; W, x) = \Phi(t; M, x) = \Phi$ *and* $\alpha_W = \alpha_M = \alpha$, *but D believes there exists at least one audience member who holds belief B1 or belief B2, then* $t^*_W < t^*_M$.

[4] Assume that if $x = 0$, then $s < 0$ and if $x = x^F$, $s > 0$. Then, under belief *B1*, a member of group W gets greater disutility from a low social image and greater utility from a high social image.

This section examined an equilibrium where actions are perfectly observed. We can easily imagine scenarios where this is not the case. The next section examines the case where the observation of the decision-maker's choice is noisy.

2.2.3 One Period Model with Noisy Signal

I now consider what happens if the signal, x, is noisy. Suppose now that nature intervenes with probability $p \in (0,1)$. If nature intervenes, $x = 0$ is transferred regardless of the decision-maker's choice. p is common knowledge, but R and A cannot observe if nature intervened.

The following result demonstrates that introducing some plausible deniability decreases the threshold type. This implies that a lower fraction of decision-makers will choose $x = x^F$, resulting in greater pooling at the bottom.

Lemma 5. *t^* is increasing in p.*

Although the threshold type falls when decision-makers can "hide" behind nature, unless the decision-maker and all audience members hold correct beliefs, at each level of p, the threshold will differ between groups, as demonstrated by the next result.

Proposition 4. *Let $t^*_{p,L}$ denote the threshold t for group L when the probability of intervention is p. If there exists at least one audience member who holds belief B1 or B2 or D believes there exists at least one audience member who holds belief B1 or B2, then $t^*_{p,W} < t^*_{p,M}$ for any $p \in (0,1)$.*

The above result shows that even with a noisy signal, group differences will still exist and that members of group W will behave more generously even when there is an opportunity for plausible deniability. Although the fraction of both groups voluntarily giving $x = 0$ grows, at every level of p this fraction will be smaller for group W than for group M.

2.2.4 Multi-period Model with Habit Formation

I show above that when there are stereotype-based ideas about groups and these ideas influence beliefs about the groups, then individuals will behave consistently with this stereotype in equilibrium. That is, even when the two groups are ex ante identical, expectations can result in group differences. Now I want to determine if these differences can persist in the long-run even in contexts where social image is not a concern (for example, because the decision-maker's choice is not observed in some period).

D participates in a sequence of dictator games, getting rematched with a different receiver and audience in each game. Each game is denoted by $g \in [1, \bar{g}]$. The sequence consists of two phases: in the first phase ($g \in [1, \hat{g}]$) actions are observed, and in the second phase ($g \in [\hat{g}+1, \bar{g}]$) actions are not observed. I assume that the decision-maker has habit formation, so the more times he has taken an action in the past, the more likely he is to take that action in the current period. Let x_g denote D's transfer in game g and s_g denote D's social image in game g (there is no transfer of social image between games because the audience is different in each game). $r \in [0, \bar{r}]$ is the weight D places on habit formation. The decision-maker places more weight on more recent actions, so past actions are time-discounted by a factor $\delta \in (0,1)$. D's utility function for each game can be written as the following:

Phase 1:

$$U = F(1-x_g, s_g) + tG(x_g - x^F) + rH(\sum_{j=1}^{g-1} \delta^j \mathbb{1}\{x_{g-j} = x_g\})$$

Phase 2:

$$U = F(1-x_g) + tG(x_g - x^F) + rH(\sum_{j=1}^{g-1} \delta^j \mathbb{1}\{x_{g-j} = x_g\})$$

I assume H is twice continuously differentiable, strictly increasing, non-negative, and $H(0) = 0$.

Note that the above utility functions differ in that s does not enter the utility function in Phase 2. Since actions are not observed in this phase, the audience cannot draw inferences about D's type and thus social image is not a concern.

The following result demonstrates that although initial group differences are due to contexts where social image is relevant, habit formation can eventually make these differences permanent, so members of the two groups behave differently even when choices are anonymous.

Proposition 5. *For $r > 0$, if there exists at least one audience member who holds the same belief B1 or belief B2 or D believes there exists at least one audience member who holds the same belief B1 or belief B2 $\forall g \in [1, \bar{g}]$, then $\exists \hat{g}^*$ such that $\forall \hat{g} > \hat{g}^*$, D chooses $x_g = x_{g-j}$ with probability 1.*

This result illustrates that my model gives rise to behavioral differences between groups that persist in the long-run, even in contexts where choices are anonymous, despite the assumption that groups were identical ex ante. Group differences are initially driven by the difference in inferences, but these initial differences will eventually become permanent due to habit formation.

2.2.5 Discussion

This model proposes a mechanism by which individuals internalize external constraints. While the external constraints were initially necessary for group differences to arise, eventually these external constraints become internalized. Individuals then self-enforce social norms and consequently adhere to the norm even when no one is watching.

This model also allows for gender identities to form endogenously. Previous papers

on identity assume either that identities are exogenously endowed ex ante (Akerlof and Kranton, 2000) or that another exogenous event, for example puberty, causes individuals to form gender identities (Bharadwaj and Cullen, 2017). My model does not require either of these, as in the model identities are formed entirely through habit formation. Thus, simply behaving in a way that is consistent with the norms of a particular group over time causes individuals to identify with that group.

2.2.6 Testable Predictions

The model generates two key testable predictions:

1. Women will be more generous when their choice is observable, even when they are offered opportunities for plausible deniability.

2. If women have sufficient life experience, women will be more generous even when no one is watching.

2.3 Empirical Analysis

I conduct an empirical analysis to test for evidence of the model's predictions. I use existing data on dictator games where gender data were collected but never analyzed. Compared to using published results from the gender differences literature, this is a cleaner test of the model's findings, as I was not aware of if there were gender differences in the data before conducting the analysis.

The empirical analysis uses data from two previous experiments that involve dictator games. I do not rely only on the results of these papers, but I use their raw data to perform new analysis. These datasets are Andreoni and Bernheim (2009) "Social Image and the 50-50 Norm: A theoretical and experimental analysis of audience effects" and Andreoni

and Rao (2011) "The Power of Asking: How communication affects selfishness, empathy, and altruism". Going forward, these studies will be referred to as AB and AR. Dataset AB allows me to test the model's first prediction that women will be more generous than men when offered plausible deniability and dataset AR allows me to test the model's second prediction that women will be more generous even in anonymous settings.

For each of these datasets, I first discuss the key features of the experimental design as well as the original paper's main result for comparison to my new analysis. Then, I present my new analysis using gender data.

2.3.1 AB

AB examines preferences for fairness versus preferences for being perceived as fair. The experimental design allowed individuals to "hide" their selfish actions by giving them plausible deniability. At the beginning of the experiment, subjects were divided into pairs, and partners were seated opposite one another, so all subjects knew with whom they were paired. Allocators needed to decide how to split \$20 between themselves and their partner. For 9 separate dictator games, there was a probability that nature intervened, which varied between 0, 0.25, 0.5, and 0.75. If nature intervened, the allocator could not choose the allocation, and instead a predetermined amount (x_0 or $20 - x_0$) was transferred. There were two treatments, one where $x_0 = 0$ and one where $x_0 = 1$.[5] At the end of the experiment, one of the decisions was randomly selected and the outcome for each pair was made public.[6] The experiment involved 120 subjects (60 men and 60 women), all undergraduates at the University of Wisconsin–Madison.[7]

[5]Each subject participated in only one of the treatments.

[6]The experimenter wrote the final allocation on the board at the front of the room. This decision sheet was also used to determine payments.

[7]One pair in condition $x_0 = 1$ did not complete the experiment, so only 118 subjects are included in analysis.

Original Results

Figures 2.1 and 2.2 show the distributions of dictators' voluntary choices in the two conditions ($x_0 = 0$ and $x_0 = 1$, respectively). Values of x are grouped into five categories: $x = 0$, $x = 1$, $2 \leq x \leq 9$, $x = 10$, and $x > 10$. Looking at Figure 2.1, when $p = 0$, 57 percent of allocators transfer half the prize. As p increases, this fraction steadily declines, and when $p = 75$, only 28 percent of subjects split the prize equally. As p increases, the fraction of subjects transferring nothing grows, starting at 30 percent when $p = 0$ and ending at 70 percent when $p = 75$.

Looking at Figure 2.2, a large fraction of subjects choose to split the prize evenly when $p = 0$ (69 percent) and, like in the previous condition, this fraction declines at p increases, shrinking to 34 percent when $p = 75$. Conversely, the fraction of subjects transferring 1 to their partner grows substantially as p increases, beginning at only 3 percent when $p = 0$ and growing to 48 percent when $p = 75$.

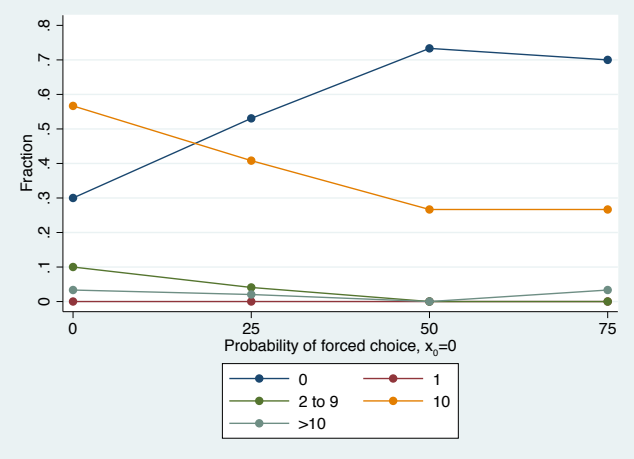


Figure 2.1: Distribution of amounts allocated to partners, condition $x_0 = 0$

Table 2.1 reports the results of two linear probability models. Looking at the first column of Table 2.1, the probability of choosing $x = x_0$ increases by approximately 27 percentage points when p increases from 0 to 0.25, and increases by approximately 15

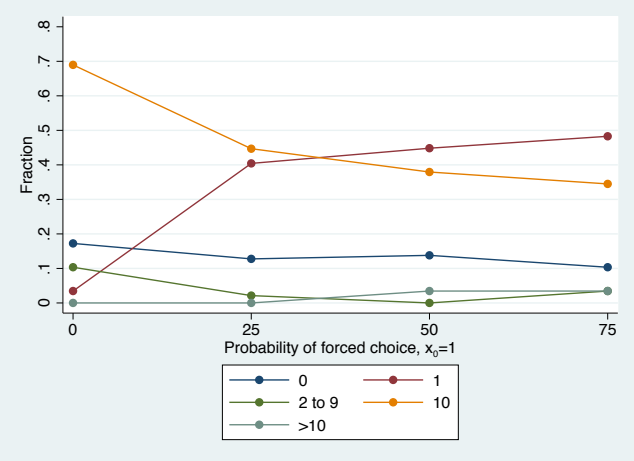


Figure 2.2: Distribution of amounts allocated to partners, condition $x_0 = 1$

percentage points when p increases from 0.25 to 0.5. This suggests that there is a significant increase in pooling at x_0 at these increases in p but not when p rises from 0.5 to 0.75. Looking at the second column, the coefficients imply that there is a significant decrease in pooling at $x = 10$ when p increases from 0 to 0.25, as the probability of choosing $x = 10$ decreases by nearly 24 percentage points, but there is no significant decline when p increases from 0.25 to 0.5 or 0.5 to 0.75. Similar results hold when I separate by condition (estimates reported in Table 2.2).

Table 2.1: Linear Probability Models

	Probability of Choosing $x = x_0$	Probability of Choosing $x = 10$
$p \geq 25$	0.271***	-0.237***
	(0.0786)	(0.0761)
$p \geq 50$	0.153***	-0.0678
	(0.0549)	(0.0476)
$p = 75$	0.000	-0.0169
	(0.0487)	(0.0525)
Constant	0.169***	0.627***
	(0.0549)	(0.0514)
Observations	236	236

Standard errors (clustered at the individual level) in parentheses. All specifications include individual fixed effects. *** p<0.01, ** p<0.05, * p<0.1

Table 2.2: Linear Probability Models by Condition

	Probability of choosing $x = x_0$		Probability of choosing $x = 10$	
	$x_0 = 0$	$x_0 = 1$	$x_0 = 0$	$x_0 = 1$
$p \geq 25$	0.233**	0.310**	-0.233*	-0.241**
	(0.108)	(0.118)	(0.121)	(0.0946)
$p \geq 50$	0.200**	0.103	-0.0667	-0.0690
	(0.0869)	(0.0673)	(0.0780)	(0.0560)
$p = 75$	-0.0333	0.0345	0.000	-0.0345
	(0.0683)	(0.0707)	(0.0793)	(0.0707)
Constant	0.300***	0.0345	0.567***	0.690***
	(0.0714)	(0.0860)	(0.0769)	(0.0701)
Observations	120	116	120	116

Standard errors (clustered at the individual level) in parentheses. All specifications include individual fixed effects. *** p<0.01, ** p<0.05, * p<0.1

Gender Analysis

This dataset allows me to test the model's first prediction that women will be more generous than men even when offered an opportunity to hide a selfish action behind a noisy signal.

Figures 2.3 and 2.4 show the distributions of dictators' voluntary choices in the two conditions ($x_0 = 0$ and $x_0 = 1$, respectively) separately for men and women. The differences in these distributions is particularly striking in Figure 2.3. Nearly 40 percent of men transfer nothing when $p = 0$ and this increases to over 80 percent when $p = 75$. By contrast, only half as many women (approximately 20 precent) choose $x = 0$ when $p = 0$ and this fraction increases to 57 percent when $p = 75$. Looking at even-splits, 56 percent of men transfer half the prize when $p = 0$ and this shrinks to 12 percent when $p = 75$. This decrease is less substantial for women, as 57 percent transfer half the prize when $p = 0$ and this only shrinks to 43 percent when $p = 75$. This means in the decision with the highest level of plausible deniability, compared to men, over 3.5 times as many women are still opting to share the pie equally.

It is also interesting to note where these increases/decreases come from. When p

increases from 0 to 0.25, the same fraction of women give $x = 0$ and the fraction of women giving intermediate amounts ($x \in [1, 9]$) decreases to 0 when $p \geq 25$. For men, however, the fraction giving intermediate amounts stays relatively constant when p increases from 0 to 0.25. This illustrates an interesting pattern in "switching" behavior. The increase in pooling at $x = 0$ for men when p increases from 0 to 0.25 is driven by men switching from giving equal divisions to giving zero when there is an opportunity for plausible deniability. The increase in pooling at $x = 0$ for women is driven by women who were giving intermediate amounts when choices were perfectly observable.

These results also suggest that women who switch from making equal divisions need a greater degree of plausible deniability before they are willing to change their behavior. While men changed their behavior from giving half to giving nothing at any positive level of plausible deniability, women needed this probability to be 0.5 in order for a majority fraction to choose $x = 0$. The willingness to take advantage of plausible deniability is clearly blunted for women compared to men.

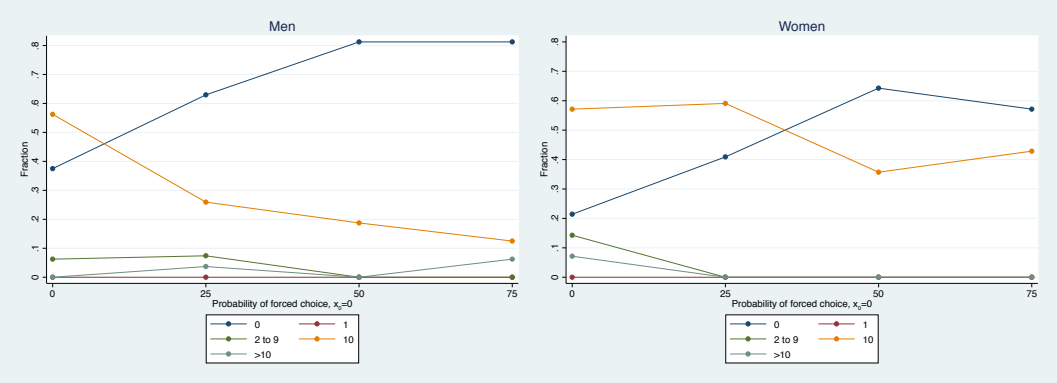


Figure 2.3: Distribution of amounts allocated to partners by gender, condition $x_0 = 0$

Table 2.3 reports the results of linear probability models. Columns 1 and 3 report results for the $x_0 = 0$ condition and columns 2 and 4 report results for the $x_0 = 1$ condition. Looking at the first column, there is a statistically significant increase in pooling at $x = 0$ when p increases from 0 to 0.25 for women, but not for men. Conversely, there is a

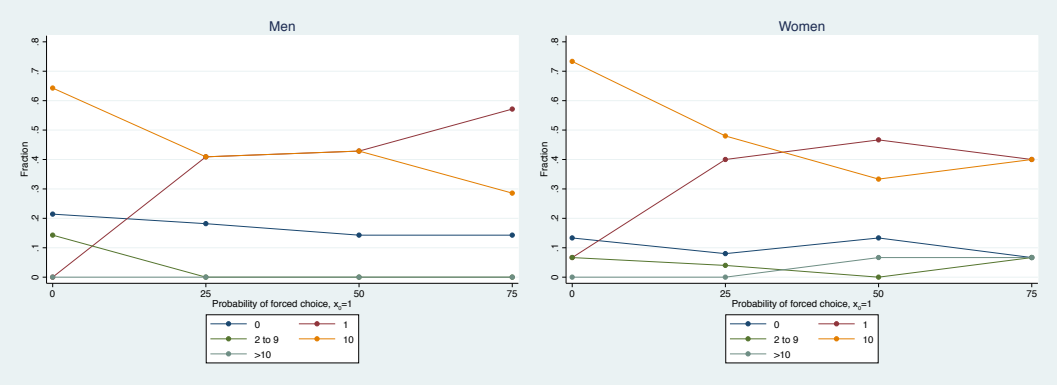


Figure 2.4: Distribution of amounts allocated to partners by gender, condition $x_0 = 1$

Table 2.3: Linear Probability Models by Condition

	Probability of choosing $x = x_0$		Probability of choosing $x = 10$	
	$x_0 = 0$	$x_0 = 1$	$x_0 = 0$	$x_0 = 1$
$p \geq 25$	0.187	0.357**	-0.375**	-0.214
	(0.159)	(0.155)	(0.147)	(0.133)
$p \geq 50$	0.250*	0.0714	-0.000	0.000
	(0.131)	(0.0835)	(0.107)	(2.63e-09)
$p = 75$	-0.000	0.143	-0.0625	-0.143
		(0.113)	(0.0733)	(0.113)
$p \geq 25 \times$ Female	0.0982	-0.0905	0.304	-0.0524
	(0.216)	(0.237)	(0.241)	(0.192)
$p \geq 50 \times$ Female	-0.107	0.0619	-0.143	-0.133
	(0.173)	(0.135)	(0.156)	(0.106)
$p = 75 \times$ Female	-0.0714	-0.210	0.134	0.210
	(0.148)	(0.138)	(0.165)	(0.138)
Constant	0.300***	0.0345	0.567***	0.690***
	(0.0727)	(0.0870)	(0.0754)	(0.0712)
$p \geq 25$ if Female	0.286*	0.267	-0.0714	-0.267*
$p \geq 50$ if Female	0.143	0.133	-0.143	-0.133
$p = 75$ if Female	-0.0714	-0.0667	0.0714	0.0667
Observations	120	116	120	116

Standard errors (clustered at the individual level) in parentheses. All specifications include individual fixed effects. *** p<0.01, ** p<0.05, * p<0.1

statistically significant increase when p increases from 0.25 to 0.5 for men but not for women. The coefficient for both is insignificant for $p = 75$. Looking at the third column, none of the coefficients are statistically significant for women. However, when p increases

from 0 to 0.25, the probability that a man divides the prize equally decreases by over 37 percentage points. This coefficient is statistically significant and over five times the magnitude of the coefficient for women.

Looking at the second and fourth columns ($x_0 = 1$ condition), gender differences are not as stark. The main notable difference is that there is a significant increase for men giving $x = 1$ when p increases from 0 to 0.25. Similarly, there is a significant decrease in pooling at $x = 10$ as p increases from 0 to 0.25 for women, but not for men.

Comparing condition $x_0 = 0$ to $x_0 = 1$, women behave relatively similarly between the two conditions. Even at the highest level of plausible deniability $p = 75$ in condition $x_0 = 1$, 40 percent of women chose an even split while another 40 percent chose to transfer x_0. This is similar to what happened in condition $x_0 = 0$, where these percentages were 57 and 43, respectively. On the other hand, 29 percent of men chose even splits and 48 percent chose $x = x_0$ when $p = 0.75$ while these percentages were 12 and 81, respectively in the $x_0 = 0$ condition.

Summary

There are clear differences in men's and women's behavior in the $x_0 = 0$ condition. A larger fraction of men, compared to women, chose to transfer nothing to their partner when choices were perfectly observable. While the pooling at $x = 0$ increased for both genders as subjects were able to "hide" their selfishness, at every level of plausible deniability, the fraction of men choosing to transfer zero was larger than it was for women. Conversely, while the fraction of men choosing to split the prize evenly sharply decreased as the probability of nature intervening increased, this decline didn't begin until after p was greater than 25 and the degree of decline was blunted compared to men.

Turning to condition $x_0 = 1$, the results were relatively similar between men and women. Women behaved relatively similarly between the two conditions. Thus, the lack of

a real difference between the groups stems from men acting more similarly to women under this condition rather than women acting more similarly to men.

These results are in line with the model's prediction that women will be more generous than men when provided opportunities to hide their selfishness behind noisy signal.

2.3.2 AR

AR examines the role of communication in giving decisions. The experiment involved an anonymous dictator game where they systematically varied who in the pair could speak. Pairs and roles were randomly assigned, and allocators decided how to split $10 between themselves and their partners.[8] Pairs communicated via written messages that contained both a pass allocation (numerical request) and a free response message.[9] There were five experimental treatments: Baseline (no communication), Ask (only the recipient sent a message), Explain (only the allocator sent a message), Ask-Explain (both sent a message, but the recipient sent his first), and Explain-Ask (both sent a message, but the allocator sent his first). Subjects made two allocations (with different partners) and participated in only one experimental treatment. The experiment involved 258 subjects (117 men and 141 women), all undergraduates at the University of California, San Diego.

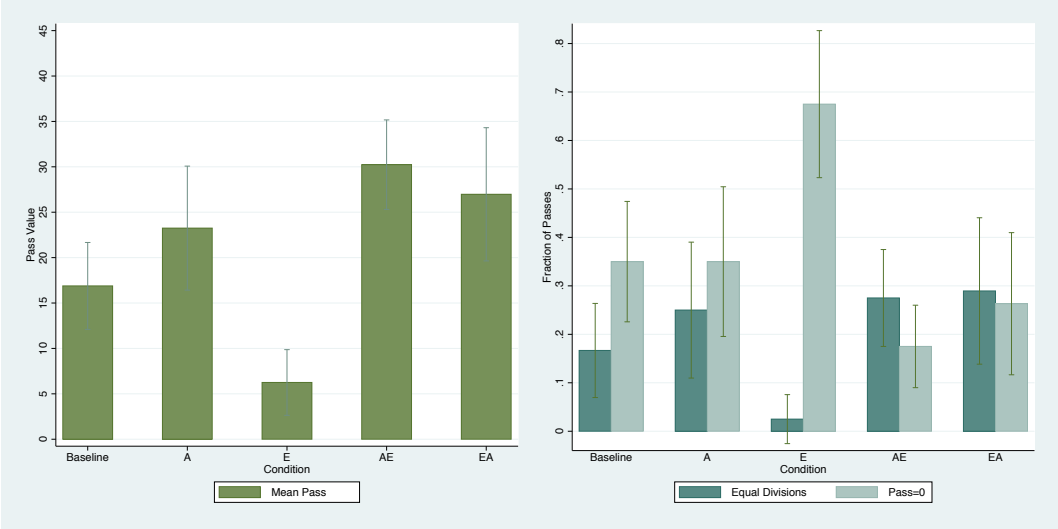
Original Results

Andreoni and Rao find that anytime the recipient spoke, giving increased. In the baseline (no-communication) condition, subjects passed 15.3 MU on average. Giving was higher in the Ask condition, with subjects passing 23.25 MU on average, and this difference becomes statistically significant when only requests for an even division or less

[8]Subjects divided 100 monetary units (MU) at an exchange rate of 1 MU = $0.10.

[9]The only restriction on messages was that they could not contain identitying information or promises outside the lab.

are considered (Wilcoxon rank-sum $z = 1.965, p < 0.049$). Giving was highest in the two-way communication conditions, and this difference is significantly different from Baseline (AE: $z = 3.29, p < 0.001$, EA: $z = 2.04, p < 0.041$). Figure 2.5 (left panel) presents mean pass values.[10]



(Left panel) Means of pass value by condition: Baseline (no communication), A (ask by recipients), E (explain by allocators), AE (ask then explain), EA (explain then ask). (Right panel) Fraction of equal divisions and pass=0 by condition. The allocator determined the final allocation of 100 MU between him/herself and an anonymous receiver. Bars give +/- 2 s.e.

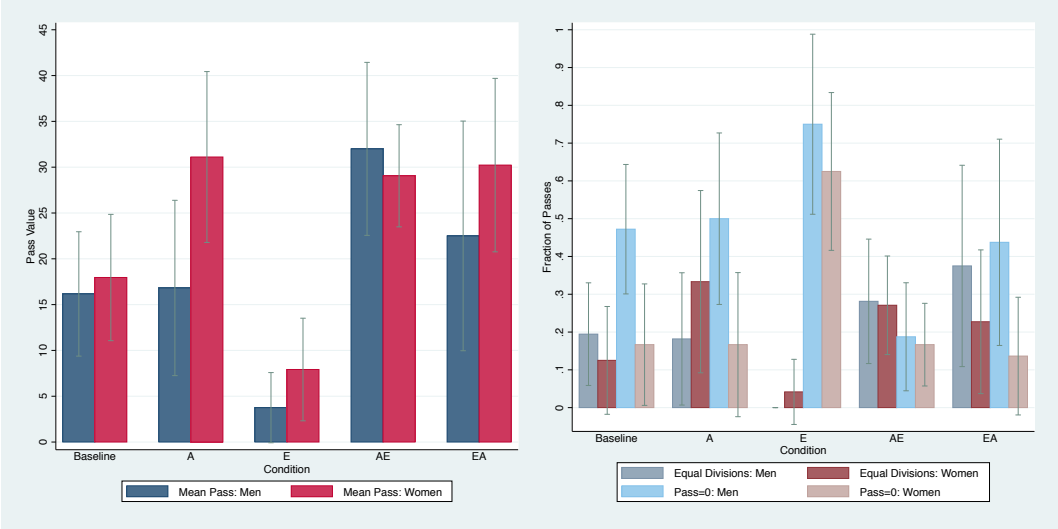
Figure 2.5: Means of pass values and fraction of equal divisions and passes of zero by condition

Gender Analysis

This dataset allows me to test the second prediction that women will be more generous even when choices are anonymous. Figure 2.6 (left panel) presents mean pass values and the fraction of subjects who chose equal divisions and to pass zero (right panel) separately for men and women. When subjects did not communicate with one another (Baseline condition), men and women were equally generous on average (16.2 MU vs. 17.96 MU, respectively). However, compared to women, nearly three times as many men chose

[10] This is a recreation of Figure 2 from Andreoni and Rao (2011).

to allocate nothing to their partners (47.2 percent of men vs. 16.67 percent of women; Fisher's Exact Test: $p = 0.026$).



(Left panel) Means of pass value by condition: Baseline (no communication), A (ask by recipients), E (explain by allocators), AE (ask then explain), EA (explain then ask). (Right panel) Fraction of equal divisions and pass=0 by condition. The allocator determined the final allocation of 100 MU between him/herself and an anonymous receiver. Bars give +/- 2 s.e.

Figure 2.6: Means of pass values and fraction of equal divisions and passes of zero by condition and gender

Differences between men and women become stronger when receivers are allowed to speak. When only recipients send a message, women are approximately twice as generous as men on average, as men give 16.8 MU on average while women give up 31.1 MU on average–nearly one-third of the total pie (t-test: $t = -2.21, p = 0.033$). And again in this condition, women are substantially less likely to give nothing to their partners (50.0 percent of men vs. 16.67 percent of women; Fisher's Exact test: $p = 0.046$). Comparing the distributions of allocations between men and women in this condition is even more striking. Figure 2.7 presents smoothed kernel densities of pass values for the Baseline and Ask conditions. The distributions for men and women in the Ask condition are both visibly and statistically significantly different (Wilcoxon rank-sum $z = -1.99, p = 0.046$; Kolmogorov-Smirnov $D = 0.42, p = 0.031$). Women were also more generous than men in

two-way communication when allocators spoke first (Explain-Ask condition), as they were again significantly less likely to make zero allocations (43.8 percent of men vs. 13.6 percent of women; Fisher's Exact test $p = 0.062$). Men and women were equally generous in the Ask-Explain condition, but this was due to men being more generous in this condition compared to the others. Namely, a much smaller fraction of men gave zero in this condition compared to all the others (18.8 percent in Ask-Explain compared to a minimum of 44 percent across the remaining conditions).

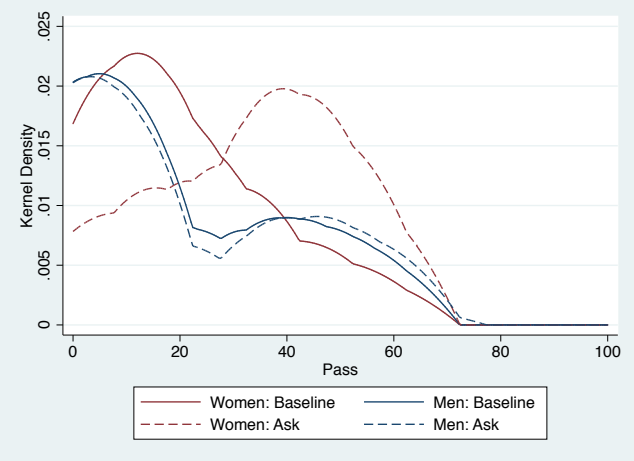


Figure 2.7: Smoothed kernel densities of pass values–Baseline and Ask conditions by gender

Men and women also respond differently to numerical pass requests. Looking at the difference between the recipient's numerical request and the allocator's pass value, again reveals large and significant gender differences. Women gave, on average, amounts closer to the request. In the Ask condition, the mean difference between the request men receive and what they give is more than twice that for women (35 MU vs. 14.5 MU), and this difference is statistically significant($z = 2.20, p = 0.028; D = 0.38, p = 0.07$). This size of this difference is heavily driven by a large number of men receiving requests of 50 MU (the modal request) and responding by giving nothing. This result is not due to men receiving higher pass requests (or conversely women receiving more "reasonable" requests), as the

requests allocators' received did not differ by gender in any condition.[11]

This difference stays relatively stable for women between one- and two-way communication (Ask-Explain: 18.6 MU, Explain-Ask: 14.0 MU), but decreases for men (Ask-Explain: 20.4, Explain-Ask: 25.6). However, the decrease between Ask and Ask-Explain is marginally insignificant ($D = 0.31, p = 0.102$), while the decrease between Ask and Explain-Ask is not statistically significant.

Summary

These results are consistent with the model's prediction that if given sufficient life experience, women will be more generous even in anonymous settings. Since the subjects in this experiment are college students, it is reasonable to believe that the women in the study have had enough experience to generate generous habits. Women were, in general, less likely to make perfectly selfish allocations. And when receivers were permitted to "speak," women were substantially more generous than men. Women were responsive to this social norm even when their identity was unknown to all those involved in the study, including the beneficiary of their generosity.

Additionally, considering gender leads to a very different conclusion of the original results drawn from this dataset. Andreoni and Rao found that whenever the recipient spoke, giving increased. However, this conclusion is only true for women. For male subjects to give more generous allocations, there needed to be two-way communication *and* the recipient needed to speak first. This challenges their finding that giving was highest under two-way communication.

[11] Since partners were anonymous to one another, and receivers therefore didn't know the gender of their partners, this is not surprising.

2.3.3 Summary of Results

In a public setting, women were less likely to exploit an opportunity to hide their selfishness when they were offered some degree of plausible deniability. Women were less likely to make perfectly selfish allocations and were substantially more generous in response to the presence of requests, even though choices were anonymous. These results provide empirical support consistent with the model's predictions.

Another interesting finding from this analysis is that in both of the datasets in the empirical analysis, one gender was responsible for driving some or all of the published results. The measured average treatment effect was not representative of the sample. Instead, it was an average of two extremes–one group that was strongly affected by the treatment and another group that was either not affected at all or was affected to a significantly lesser degree. This analysis provides strong evidence that heterogenous treatment effects due to behavioral differences between men and women may be responsible for many experimental results. This suggests that even if an experimental treatment was not designed with the intention of examining gender differences and even if it is not clear that the environment being studied should have differential effects on men and women, additional analysis to examine heterogenous treatment effects by gender should be performed.

2.4 Experimental Design

The empirical analysis provides evidence in support of the model's predictions, but it does not test the model's mechanism. In order to provide a more direct test of a key component of the model, I design and implement new experiments. While the model is designed to capture a complex process that takes place over an individual's lifetime, I distill this down into a key feature that can be tested in a laboratory setting: early decisions can have persistent effects even when the constraints of those decisions change.

In the experiment, subjects make a series of dictator game allocations. The games vary in the chance that nature intervenes and forces subjects to either keep everything or give everything to their partner with equal probability. If there is a chance that nature intervenes, this gives subjects an opportunity for plausible deniability if they choose to keep everything (because if others in the experiment observe an allocation where the subject keeps everything, they will be unable to determine if the subject made that choice or if nature forced that allocation). Experimental treatments vary in the order that subjects make decisions, so subjects' first choice either (i) offers no opportunity for plausible deniability and this opportunity increases in subsequent decisions or (ii) offers the highest level of plausible deniability and this opportunity decreases in subsequent decisions. In the experiment, I examine if exposing subjects to high external constraints in initial decisions mitigates gender differences even when subjects can take advantage of plausible deniability in later decisions.

2.4.1 Procedures

All sessions were conducted at UCSD's EconLab using undergraduate students recruited via email. Instructions were read aloud to subjects and they submitted all responses via experimental software. Subjects were divided into pairs, with partners and roles assigned randomly. Within each pair, one subject was designated as the decision-maker and the other pair the receiver. The decision-maker determined how the pair divided $30.

Each session proceeded as follows: Subjects were randomly divided into pairs, and partners were seated opposite one another. One-at-a-time, pairs stood up and greeted one another in order to identify themselves to their partner. Decision-makers made 27 decisions for how to split $30 between themselves and their partner. Decisions differed in the probability that they were forced to make a particular allocation. If a decision was "forced," the decision-maker kept all $30 and transferred nothing to his partner or kept

nothing and transferred all $30 to his partner with equal probability.[12] The probability that a decision was forced varied between 10 values (0, 0.01, 0.02, 0.03, 0.05, 0.10, 0.25, 0.50, 0.75, and 0.90). For each decision, decision-makers knew whether they were "forced" or free to make an allocation. This was to highlight for decision-makers that they knew whether their choice was forced but no one else did. After all subjects had submitted their decisions, one decision was selected at random to determine payments. At the end of the session, the outcome for all groups of this selected decision was written on the board at the front of the room. There were two treatment groups: one treatment where subjects made decisions in increasing order of being forced (starting with a zero probability of being forced and ending with 0.90) and another treatment where subjects made decisions in decreasing order of being forced (starting with 0.90 and ending with 0). I will refer to these as the Increasing treatment and the Decreasing treatment, respectively. This is a between subjects design (all subjects within a single session were in the same treatment and each subject participated in only one treatment).

At the end of the session, subjects were paid in cash. Sessions lasted approximately one hour, and subjects earned an average of $20, including a $5 show-up fee for their participation. 9 sessions (5 sessions of the Increasing treatment and 4 sessions of the Decreasing treatment) of 16-20 subjects per session were conducted, resulting in a total of 166 subjects (41 men and 42 women decision-makers).

2.5 Experimental Results

I seek to answer two questions: First, do individuals exhibit persistence in their choices–that is, is what individuals choose in each decision relatively stable even though the

[12]This was done to make the ex ante outcome of being forced equal for both the decision-maker and the partner. This was to ensure that individuals did not try to maximize ex ante fairness by being more generous in decisions where they were able to make an allocation in order to make up for forced decisions in which they were forced to make a selfish allocation.

opportunity for plausible deniability varies? Second, does the order of the decisions matter? That is, if individuals are initially exposed to a low probability of nature intervening, are they more generous initially and does this generosity extend to later decisions where the opportunity for plausible deniability is high?

I formalize these questions into three hypotheses:

Hypothesis 1: Choices will be relatively stable even though the opportunity for plausible deniability varies. This means that as subjects move to the next decision in the series, they will not be significantly more likely to change their allocation.

Hypothesis 2: Women in the Decreasing treatment will be more generous than men in the Decreasing treatment.

$$Pr(Pass = 15|W, D) > Pr(Pass = 15|M, D)$$

Hypothesis 3: Men and women in the Increasing treatment will be equally generous.

$$Pr(Pass = 15|W, I) = Pr(Pass = 15|M, I)$$

When subjects are initially exposed to a high probability of intervention, I predict men will be more likely to take advantage of this plausible deniability. These differences will persist through the series of decisions, so even when there are low or no opportunities for plausible deniability, men will still be less likely than women to choose equal allocations. However, when subjects' initial decisions have no probability of intervention, I predict that men will give equal allocations at approximately the same rate as women, and these initial generous actions will persist in later actions, even subjects are given the opportunity to hide a selfish action behind nature. These hypotheses mean that I predict that there will be gender differences in the Decreasing condition but these differences will be mitigated in the Increasing condition.

Looking first at Hypothesis 1, subjects' behavior appears to exhibit persistence to a high degree. When regressing the probability of choosing to pass 15 (an even split of

the pie) or pass zero on the probability that the choice was forced using linear probability models, only one coefficient is statistically significant. Looking at the first column of Table 2.4 (the outcome variable is the probability that the decision-maker passed 15), only one of the coefficients is statistically significant. This is the interaction term on the probability of forced being greater than 0.50 $\times$ Female. Although, choices seem to return back to their previous level, as the coefficient on $p \geq 75$ $\times$ Female is almost equal in magnitude but opposite in sign (it is not quite statistically significant). Moreover, the point estimates are very close to zero, with only two being greater than 0.10. Looking at the second column of this table (the outcome variable is the probability that the decision-maker passed zero), none of the coefficients are statistically significant. Similarly, the point estimates are very small in magnitude, with approximately one-third of them being approximately 0.03 or less in magnitude. Given that the opportunity for plausible deniability across choices varies greatly, the degree of stability of subject's choices is surprising.

Turning to Hypothesis 2, large gender differences are apparent when comparing men and women in the Decreasing treatment. Figure 2.8 depicts the fraction of subjects who chose equal allocations (pass 15) in this treatment. These results are also available in Table 2.5. Note that in the figure, the order of decisions goes from right to left (starting with 0.90 and ending with 0). As evidenced in the figure, the faction of women who chose to split the pie equally is greater than the fraction of men who chose this allocation at every level of intervention. That is, women are always more likely than men to choose equal allocations, and these differences are significant. Looking at subjects' first choice ($p = 0.90$), 56 percent of women chose to allocate 15 while only 21 percent of men did (two-sided Fisher's Exact test: $p = 0.045$). Even in subjects' last choice, where there no opportunity for plausible deniability, women were nearly twice as likely as men to choose to pass 15 (71 percent vs. 37 percent, two-sided Fisher's Exact test: $p = 0.054$).

While the differences between men and women's choices in the Decreasing condition

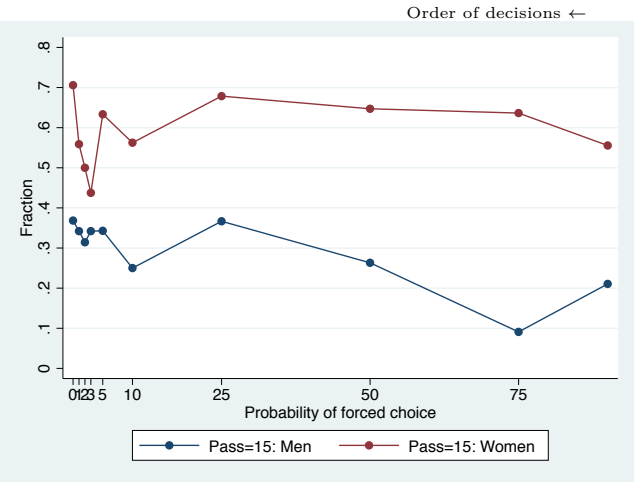


Figure 2.8: Fraction of 50-50 allocations to partners, Treatment D by gender

are large, when subjects made decisions in the opposite order, gender differences were mitigated. Looking at Hypothesis 3, men and women's behavior looks much more similar in the Increasing condition. Figure 2.9 depicts the fraction of subjects who chose 50-50 splits (pass 15) in this treatment. Note that in this figure the order of decisions goes from left to right (beginning with 0 and ending with 0.90). The fraction of men and women who chose equal divisions is not statistically different. In subjects' first choice, although a larger fraction of women choose to pass 15 to their partner–68 percent of women compared to 57 percent of men, this difference is not statistically significant (two-sided Fisher's Exact test: $p = 0.545$). Even when subjects are offered a large opportunity for plausible deniability in their last decision ($p = 0.90$), men are still as likely as women to give equal allocations (38 percent of men vs. 41 percent of women; two-sided Fisher's Exact test: $p = 1.00$)

The experimental results are in line with the hypotheses and illustrate that the the order of subjects' decisions has a large influence on their behavior. The difference in the parameters of the initial decision not only changed subjects' choices for that decision, but also their subsequent decisions. By initially exposing individuals to a high degree of plausible deniability, I relaxed the external constraints if subjects chose to act selfishly. This caused men to be less likely to give equal allocations in that decision, but this behavior

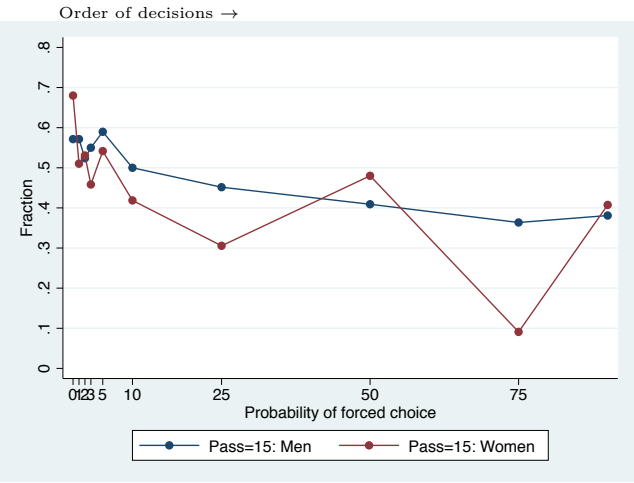


Figure 2.9: Fraction of 50-50 allocations to partners, Treatment I by gender

persisted over the series of decisions, even when there was no opportunity for plausible deniability. However, by exposing subjects to stricter external constraints on their first action, I mitigated gender differences, as men continued to behave generously even when they had ample opportunity to take advantage of plausible deniability. By simply changing the order in which subjects made decisions, I mitigated gender differences in subjects' behaviors. The results of this experiment thus present evidence in support of the model's mechanism.

2.6 Conclusion

I have proposed a theory of behavior that captures how the external constraints individuals face can eventually become internalized. This shows how social norms that are initially externally enforced later become self-enforced by the individual. This mechanism provides insight into gender differences in observed behaviors and provides an alternative explanation for behavioral differences between men and women.

This mechanism is important in the study of gender differences for two primary reasons. First, this mechanism provides a different interpretation for data on gender

Table 2.4: Linear Probability Models

	Probability of choosing Pass = 15	Probability of choosing Pass = 0
$p \geq 1$	-0.0750	0.0250
	(0.0700)	(0.0806)
$p \geq 2$	0.0000	0.0250
	(0.0538)	(0.0600)
$p \geq 3$	0.0250	0.0000
	(0.0464)	(0.0538)
$p \geq 5$	0.0500	-0.100
	(0.0531)	(0.0742)
$p \geq 10$	-0.100	0.100
	(0.0742)	(0.0834)
$p \geq 25$	0.0500	-0.0250
	(0.0531)	(0.0711)
$p \geq 50$	-0.0763	0.110
	(0.0692)	(0.0910)
$p \geq 75$	-0.0319	0.102
	(0.0858)	(0.0969)
$p = 90$	-0.0168	-0.113
	(0.0899)	(0.0949)
$p \geq 1 \times$ Female	-0.0917	0.142
	(0.0935)	(0.108)
$p \geq 2 \times$ Female	-0.0238	-0.0726
	(0.0785)	(0.0938)
$p \geq 3 \times$ Female	-0.0488	0.0714
	(0.0736)	(0.0777)
$p \geq 5 \times$ Female	0.0452	0.0286
	(0.0807)	(0.0930)
$p \geq 10 \times$ Female	-0.0190	0.0667
	(0.0985)	(0.110)
$p \geq 25 \times$ Female	-0.0500	-0.0226
	(0.0738)	(0.0945)
$p \geq 50 \times$ Female	0.172*	-0.158
	(0.0921)	(0.110)
$p \geq 75 \times$ Female	-0.139	0.0725
	(0.105)	(0.116)
$p = 90 \times$ Female	0.0959	0.0376
	(0.112)	(0.112)
Constant	0.583***	0.283***
	(0.0385)	(0.0431)
Observations	786	786

Standard errors (clustered at the individual level) in parentheses. All specifications include individual fixed effects. *** p<0.01, ** p<0.05, * p<0.1

Table 2.5: Fraction of passes by treatment and gender

Prob. of forced choice	Increasing						Decreasing					
	Number of Observations		Pass = 15		Pass = 0		Number of Observations		Pass = 15		Pass = 0	
	Men	Women	Men	Women	Men	Women	Men	Women	Men	Women	Men	Women
0	21	25	57.1	68.0	28.6	24.0	19	17	36.8	70.6	42.1	17.6
1	42	49	57.1	51.0	26.2	34.7	38	34	34.2	55.9	44.7	35.2
2	42	49	52.4	53.1	28.6	30.6	35	34	31.4	50.0	45.7	41.1
3	40	48	55.0	45.8	30.0	41.7	38	32	34.2	43.8	36.8	46.9
5	39	48	59.0	54.2	23.1	37.5	35	30	34.3	63.3	37.1	26.7
10	40	43	50.0	41.8	32.5	51.2	36	32	25.0	56.3	47.2	40.6
25	31	36	45.2	30.6	32.2	55.6	30	28	36.7	67.9	50.0	28.6
50	21	25	40.9	48.0	45.5	44.0	19	17	26.3	64.7	52.6	35.2
75	11	11	36.4	9.1	54.5	81.8	11	11	9.1	63.6	81.8	36.4
90	21	25	38.1	40.7	47.6	55.6	19	17	21.1	55.6	47.4	38.9

differences. Instead of differences in observables being due to differences in fundamentals (preference functions or types), differences in men and women's choices could be indicative of men and women facing different constraints and these constraints have become internalized over time. This analysis also provides evidence for the power and prevalence of social norms. A collection of experiments that did not set out to study gender differences actually captured very strong gender differences, so much so that the significance of their pooled results relied on the treatment effect to only one gender. In these data, even when choices were anonymous, the power of an internalized social norm was present.

Second, the results of this paper suggest that there is room for policy intervention. The danger of attributing gender differences to differences in fundamental characteristics between men and women is that is suggests that policy will be ineffectual. There is no need to construct policy if different choices are because men and women "are" different. My mechanism suggests that policy can be effective, specifically policies that either target established beliefs about men and women in order to relax the constraints put on women's behavior and policies that are targeted at habit breaking for women who have already learned to internalize social norms. There already exist a few policies that may be effective in achieving these ends. In July 2017, Britain's advertising regulator, the Committee on Advertising Practice, announced that new rules would be developed to ban advertising that promotes gender stereotypes or mocks those who do not conform to them. For example, one of the types of ads the UK policy is targeting is advertisements involving cleaning products, which typically feature women using them, and thus subtly enforce the association between women and domestic labor. Another potential for policy would be habit-breaking for women who have already formed habits for particular behaviors. Within economics, a group of female economists formed the "I just can't say no club" in order to address the frequent difficulty of women being able to say "no" to work requests that are often non-promotable in nature. Founding members include Linda Babcock and Lise Vesterlund, and the group

has since spread to three national clubs. Educating women on how to effectively decline requests is a promising potential policy.

While the idea that external constraints become internalized has clear applications to gender differences research, it is also a mechanism that could apply to other social norms. From a general policy perspective, potential research could examine how we might encourage socially desirable or welfare-improving behaviors, and eventually these behaviors will become self-perpetuating through habit formation and self-enforcement. Further examining this mechanism and its application to the way economists think about how individuals make decisions is a promising area of future theoretical, experimental, and applied research.

Appendix

Proof of Lemma 1. *Let s_x denote D's social image upon choosing x. Denote $U(x^F, s_{x^F}, t)$ as U_F and $U(0, s_0, t)$ as U_0. If we assume $G(-x^F) < 0$ and $G(0) > 0$, then $\frac{\partial U_F}{\partial t} > 0$ and $\frac{\partial U_0}{\partial t} < 0$. If we allow the domain of $U(x, s, t)$ to include $t \in (-\infty, \infty)$, then $U_F = U_0$ for some value of t. Call this value of t t^*.*[13]

Take any $\hat{t} > t^*$. Since $U_F = U_0$ at t^* and $\frac{\partial U_F}{\partial t} > 0$, then $U_F > U_0$ for $\hat{t}$. This means that any type $\hat{t}$ will choose $x = x^F$. A parallel argument holds for $t < t^*$ and choosing $x = 0$. Since I only consider pure strategy equilibria, assume that if the decision-maker is indifferent, he breaks ties by choosing $x = x^F$. □

Proof of Proposition 1. *As in Lemma 1, denote t^* as the t that satisfies $U(0, s_0, t^*) = U(x^F, s_{x^F}, t^*)$. If all members of A believe $K(t; W) = K(t; M) = K$ and $\Phi(t; W, x) = \Phi(t; M, x) = \Phi$, then $s_{0,W} = s_{0,M}$ and $s_{x^F,W} = s_{x^F,M}$. Then, for*

[13] This shows that t^* exists, but with only these assumptions, it could fall outside of the interval $[0, t^*]$. If it is the case that $t^* \geq \bar{t}$ or $t^* \leq 0$, then the equilibrium is a pooling equilibrium. If we want to examine the cases where there is partial separation, $t^* \in (0, \bar{t})$, then the assumptions that $U(0, s_0, 0) > U(x^F, s_{x^F}, 0)$ and $U(x^F, s_{x^F}, \bar{t}) > U(0, s_0, \bar{t})$ are needed.

any t, $U(0, s_{0,W}, t) = U(0, s_{0,M}, t)$ *and* $U(x^F, s_{x^F,W}, t) = U(x^F, s_{x^F,M}, t)$. *Therefore,* $U(0, s_{0,W}, t^*) = U(0, s_{0,M}, t^*) = U(x^F, s_{x^F,W}, t^*) = U(x^F, s_{x^F,M}, t^*)$ *and* $t^*_W = t^*_M = t^*$. □

Proof of Proposition 2. *First examine the case of one audience member holding belief B1. If an audience member holds belief B1, then* $\Phi(t; M, x = 0)$ *FOSD* $\Phi(t; W, x = 0)$. *Then,* $S(\Phi(t; W, x = 0)) < S(\Phi(t; M, x = 0)) \implies s_{0,W} < s_{0,M}$. *Suppose that* $t^*_W = t^*_M$. *This would imply that* $U(0, s_{0,W}, t^*_W) = U(x^F, s_{x^F,W}, t^*_W) = U(0, s_{0,M}, t^*_M) = U(x^F, s_{x^F,M}, t^*_M)$. *But this cannot be true because* $s_{0,W} \neq s_{0,M}$. *Next, suppose* $t^*_W > t^*_M$. *This implies that* $U(0, s_{0,W}, t^*_W) > U(0, s_{0,M}, t^*_W)$. *But this can't be true because* $s_{0,W} < s_{0,M}$. *Then it must be that* $t^*_W < t^*_M$.

Next examine the case of one audience member holding belief B2. Consider any social image s_0 *such that* $s_0 = s_{0,W} = s_{0,M}$. *If groups had utility functions* U_W *and* U_M, *then for any fixed* t, $U_W(0, s_0, t) < U_M(0, s_0, t)$, *since* $F(1, \alpha_W s_0) + tG(-x^F) < F(1, \alpha_M s_0) + tG(-x^F)$. *Then,* $t^*_W < t^*_M$. *When an audience member holds belief B2, he believes this is the case, and thus upon observing* $x = 0$, $\Phi(t; M, x = 0)$ *FOSD* $\Phi(t; W, x = 0) \implies s_{0,W} < s_{0,M}$. *Then, even when the utility function for both groups is* U, $s_{0,W} < s_{0,M}$, *then* $t^*_W < t^*_M$. □

Proof of Proposition 3. *If A holds correct beliefs and makes correct inferences, then* $\Phi(t; W, x) = \Phi(t; M, x) = \Phi$. *But, if D believes that at least one audience member holds belief B1 or B2, then* $S_W(\Phi(t; W, x = 0)) < S_M(\Phi(t; M, x = 0)) and s_{0,W} < s_{0,M}$. *Therefore,* $t^*_W < t^*_M$. □

Proof of Lemma 2. *If an audience member observes* $x = 0$, *then he knows there is some probability that nature, and not the decision-maker, made this allocation. I assume that upon observing* $x = 0$ *the audience member takes* p *into account and updates such that he believes that the probability* D *chose* $x = 0$ *conditional on observing* $x = 0$ *and* p *is decreasing in* p. *Take any* $p_1, p_2 \in (0, 1)$ *with* $p_1 > p_2$. *Then,* $\Phi(t, L, 0, p_1)$ *FOSD* $\Phi(t, L, 0, p_2)$ *and* $S(\Phi(t, L, 0, p_1)) > S(\Phi(t, L, 0, p_2))$. *Denote the social image for a given* x *and* p *as* $s_{x,p}$.

Define $t^*_{p_1}$ *to be the type such that* $U(0, s_{0,p_1}, t^*_{p_1}) = U(x^F, s_{x^F,p_1}, t^*_{p_1})$. *Since* $s_{0,p_1} > s_{0,p_2}$, $U(0, s_{0,p_1}, t^*_{p_1}) > U(0, s_{0,p_2}, t^*_{p_1})$. *Then, for* $U(0, s_{0,p_2}, t^*_{p_2}) = U(x^F, s_{x^F,p_2}, t^*_{p_2})$, $t^*_{p_1} < t^*_{p_2}$. □

Proof of Proposition 4. *Suppose* $t^*_{p,W} = t^*_{p,M}$. *This would imply that* $U(0, s_{0,p,W}, t^*_W) = U(x^F, s_{x^F,p,W}, t^*_W) = U(0, s_{0,p,M}, t^*_M) = U(x^F, s_{x^F,p,M}, t^*, M)$. *But this cannot be true because* $s_{0,p,W} \neq s_{0,p,M}$. *Next suppose* $t^*_{p,W} > t^*_{p,M}$. *This implies* $U(0, s_{0,p,W}, t^*_W) > U(0, s_{0,p,M}, t^*_W)$. *But this cannot be true because* $s_{0,p,W} < s_{0,p,M}$. *Then it must be that* $t^*_{p,W} < t^*_{p,M}$. □

Proof of Proposition 5. *Without loss of generality, I focus on the actions of group* W. *Define* $\tilde{t}$ *to be the type such that* $F(1-x^F) + \tilde{t}G(x^F - x^F) = F(1-0) + \tilde{t}G(0-x^F)$. *By Lemma 1,* $\forall t > \tilde{t}$, D *will choose* $x = x^F$. *Individuals of these types will give* x^F *in phase 2 even without habit formation. In phase 1, members of group* W *with type* $t < t^*_W$ *transfer 0, so they do not have any incentive to switch actions in phase 2. Then, restrict attention on decision-makers who are of are of types* $t \in [t^*_W, \tilde{t}]$. *These are types who would rather pick 0, but gave* x^F *in phase 1 because actions were observable.*

Looking at continuation payoffs,[14] D *will choose* $x = x^F$ *in all* g *iff the continuation payoff from giving* x^F *is greater than or equal to the continuation payoff from giving* $x = 0$. *If* D *transfers* $x = x^F$, D*'s utility is:*

$$U = \sum_{g=\hat{g}+1}^{\bar{g}} [F(1-x^F) + tG(x^F - x^F) + rH(\sum_{j=1}^{g-1} \delta^j)] \tag{2.1}$$

If D *transferes* $x = 0$, D*'s utility is:*

$$U = F(1-0) + tG(0-x^F) + rH(0) + \sum_{g=\hat{g}+2}^{\bar{g}} [F(1-0) + tG(0-x^F) + rH(\sum_{j=\hat{g}+1}^{g-1} \delta^j)] \tag{2.2}$$

[14] For this proof, I assume no future discounting, as this is a stronger result. The result will obviously still hold if the decision-maker discounts future periods.

D will choose x^F *in all periods iff* $(1) \geq (2)$.

Simplifying (1), we obtain

$$U = [\bar{g} - (\hat{g}+1)][F(1-x^F) + tG(x^F - x^F)] + \sum_{g=\hat{g}+1}^{\bar{g}} [rH(\sum_{j=1}^{g-1})]$$

Simplifying (2) yields

$$U = F(1-0) + tG(0-x^F) + [\bar{g} - (\hat{g}+2)][F(1-0) + tG(0-x^F)] + \sum_{g=\hat{g}+2}^{\bar{g}} [rH(\sum_{j=\hat{g}+1}^{g-1} \delta^j)]$$

$$= [\bar{g} - (\hat{g}+1)][F(1-0) + tG(0-x^F)] + \sum_{g=\hat{g}+2}^{\bar{g}} [rH(\sum_{j=\hat{g}+1}^{g-1} \delta^j)]$$

As $\hat{g}$ *increases, the incentive to switch from 0 to* x^F *decreases, because the habit formation term for staying with* x^F *increases and the number of periods to collect extra benefit of* $F(1-0) + tG(0-x^F)$ *decreases . So as* $\hat{g}$ *increases (approaches* $\bar{g}$*), (2) gets smaller and the second term of (1) gets larger. Then, if we make* $\bar{g}$ *arbitrarily large, there will be some* $\hat{g}^*$ *such that for* $\hat{g} > \hat{g}^*$*, (1)>(2). Then in games* $g > \hat{g}$*, D will choose* $x = x^F$*. Thus, for types* $t \in [0, t_W^*)$*, D chooses* $x = 0 \ \forall g \in [1, \bar{g}]$*, for types* $t \in [t_W^*, \bar{t}]$*, D will choose* $x = x^F \ \forall g \in [1, \bar{g}]$. □

Chapter 2, in full, is currently being prepared for submission for publication of the material. Giffin, Erin. The dissertation author is the sole author of this material.

Chapter 3

Recall of Repeated Games

3.1 Introduction

An implicit assumption of most game theory models is that players have perfect memory. In repeated games, players often condition their strategies on the entire history of the game, irrespective of its length or complexity. In recent years, theory models that relax this assumption have emerged. These models allow players to forget things, categorize events, ignore information and update infrequently (see Monte (2013) for a model of repeated games with bounded memory that captures all of these phenomena). While there has been increased attention to bounded memory in the theoretical literature, there has not been a similar uptick in empirical or experimental work by economists on this topic. This is an important area of study, as memory is dependent on the environment and different types of memory limitations would result in different behavioral implications. Thus, studying memory in these particular environments is critical to inform theoretical models.

Memory could be limited in three ways: (i) memory could be finite, (ii) memory could biased, and/or (iii) memory could include miscoded or false memories. We can think of memory being finite simply as memory having a limited capacity, meaning an individual cannot store every piece of information he observes. In this first case, the information storage process is stochastic, so every piece of information the individual observes has an

equal likelihood of being remembered or forgotten. In other words, memories are missing at random. However, in the second case of biased memory, there is a systematic bias in what information is stored. That is, what information is stored is not a random process and involves different probabilities of information being remembered or forgotten. These probabilities could be based on different qualities or characteristics of the information, including how recently the information was observed (i.e., primacy or recency effects), how important the information is thought to be (i.e., saliency effects), or how much the individual wants to believe the information (i.e., confirmation bias), to name a few possibilities. But memory could not always be accurate in what it includes; it could include information that was never observed. This is the third scenario: memory that contains miscoded or false memories. These three limitations can be summarized as follows: (i) memories can be missing at random, (ii) memories can be missing systematically, and (iii) memories can be incorrect.

These different memory limitations are analogous to the issues that arise with missing data in empirical studies. Due to difficulties with data collection or attrition, "missing data arise in almost all serious statistical analyses" (Gelman and Hill, 2006). Missing data can be missing according to different mechanisms (Gelman and Hill, 2006): missingness at random, missingness that depends on unobserved predictors, and missingness that depends on the missing value itself.

Missing data is not ideal, as the economist draws conclusions based on a subset of the existing set of all the information; however, that subset appropriately represents the full data set, and consequently leads to unbiased inferences.[1] But if the information is missing systematically (it is non-random), the subset of information that was able to be collected will not be representative. Thus, unless the missingness can be explicitly modeled, the inferences will be biased. Further, if there is information that is incorrect, this will

[1]Inferences will be unbiased if either the missingness is completely random or we control for the all the observables the missingness depends on.

further bias our estimates. If there is known measurement error, this can be corrected for during analysis. But if this error is either unknown, or if the error differentially affects the data (only some data are affected or different data are affected differently), there is no way to know which data are correct and which are errors. This will clearly bias any inferences made from these data.

While the problem of missing data is widely known, the mechanisms of memory in games, its limitations, and how it is related to behavior are not. We know the dangers associated with missing data in empirical studies and how it can affect inferences made from the data. Similar incorrect inferences can be made by players if they are operating with analogous memory limitations. What strategy a player selects, how effectively he can implement that strategy, and his ability to best respond to his opponent(s) will all be impeded if he is making these decisions based on missing or incorrect information about the history.

To examine these questions, I first build a theoretical framework to analyze how various memory types affect a decision maker's strategy selection and implementation. Using this framework, I show that memory limitations reduce the number of strategies a player can use and limit his ability to successfully implement that strategy. I find that memory limitations can result in inefficiencies, and these inefficiencies are greatest when the player has biased memory, especially when he has memory miscoding (false memories).

To gather empirical evidence on memory in a repeated game, I conducted a laboratory experiment and examine memory and behavior in the environment of a canonical repeated game: the Finitely Repeated Prisoner's Dilemma. I first establish general trends on how accurately players recall the history of the game and analyze different factors that are associated with recall. I find that memory is the history is often not perfect, with approximately half of subjects recalling their own action correctly for all rounds of a given match, and only approximately one-third of subjects recalling both their own action and

their partner's action correctly for all rounds of a given match. I then analyze factors that may be associated with recall accuracy, focusing on timing, outcomes, and the number of times players changed actions. I find some evidence of primacy effects, as earlier rounds are recalled accurately with higher probability. Subjects are also more likely to recall particular outcomes. Specifically, subjects are less likely to recall their own action when the outcome they cooperated on a defector and they are less likely to recall their partner's action when they defected on a cooperator. I finally find that the number of times a subject or his partner changed actions is negatively correlated with the subject's recall accuracy. Taken together, these results suggest that subjects' memory of the game's history both has a limited capacity and is biased.

I then examine how memory is associated with behavior, focusing on subjects' strategies and strategy implementation. To do this, I first develop a strategy estimation method to estimate individual subjects' strategies. For each match, I estimate two strategies per subject: one based on his observed behavior and one based on his recalled behavior. Using these estimated strategies, I find that subjects frequently recall using a different strategy than they actually did, and they believe they were more successful at implementing this strategy (as they recall making fewer deviations from their recalled strategy). I finally find that recall accuracy of the history is negatively correlated with the number of times a subject deviated from his estimated strategy.

This paper is the first, to my knowledge, to measure memory in an economic game. There is very limited research on memory in games, and those that do exist do not seek to measure memory in this environment. Huck and Müller (2002) seek to recreate the absent-minded driver in the lab. This study is essentially an existence proof, with the authors asking if there is an environment such that they can mimic the result of the classic absent-minded driver model. They have subjects play the game, but they are heavily distracted by having to complete another task simultaneously. So while this study examines

memory, it does not measure memory. In Ivanov et al. (2010), subjects play a second-price common-value auction. In one condition, subjects play against a computer who made the same choices they did previously; thus, subjects are asked to play against their past selves. While this task requires subjects to remember what they did previously, this study doesn't explicitly measure memory. Additionally, the focus of this study was not memory, but rather to determine beliefs and level-k thinking in players.

The paper that is most closely related to this one is a working paper by Fragiadakis et al. (2013). In their experiment, subjects play a sequence of two-player guessing games. Subjects are later asked to either replicate or best respond to their previous play. That is, subjects replay the same 20 games in the same role and are asked to replicate their choices from the first 20 games or subjects play against a computer that makes the same choices they made in the first 20 games[2]. While this paper is most similar in experimental procedure to mine, there are some important distinctions. The first is that an inability to successfully replicate past actions is not seen as a failure of memory but as evidence of idiosyncratic randomness of decisions. Further, the authors do not analyze different factors that are associated with recall, outside of subjects being classified as a behavioral type. There are many different factors that could be associated with memory accuracy, which is one of the focuses of this paper.

This paper also contributes to the literature on strategy estimation, as it is the first to estimate strategies at the individual-level. Dal Bó and Fréchette (2011) were the first to estimate strategies, developing the Strategy Frequency Estimation Method (SFEM). In their method, they find an estimation of the importance of strategies using maximum likelihood, meaning they find the likelihood that the data as a whole corresponds to a given strategy. My estimation procedure estimates something different: it estimates the strategy for an individual subject in a single supergame. I am also able to weaken the two

[2]This condition is very similar to that in Ivanov et al. (2010) (discussed in the previous paragraph).

key assumptions of SFEM: that all subjects have a given probability of using one of the six strategies they consider and that subjects do not change strategies from repeated game to repeated game.

The remainder of the paper proceeds as follows. In Section 2, I lay out a theoretical framework to examine how memory limitations can lead to inefficiencies. In Section 3, I describe the experimental design. Section 4 provides the results on general trends and patterns in recall accuracy and different factors that are associated with recall. Section 5 describes the strategy estimation method and provides results using this method. Section 6 concludes and provides directions for future research.

3.2 Theoretical Framework

Traditionally, in models of decision makers with bounded memory, memory is treated as a choice variable. That is, the decision maker has control over his memory and can choose to store, disregard, or purge information. In these models, the DM is restricted by memory constraints, but he has control of what information to keep in memory subject to these constraints.[3] Memory may not be so easily controlled by the individual, as some memory processes may be automatic or subconscious, making them exogenous from the perspective of the DM.

Introspection suggests that sometimes individuals would very much like to remember information but forget it. Sometimes individuals wish they could clear space in their memory by removing unnecessary information but are unable to do so.

I construct a theoretical framework that captures these different characteristics of memory. A DM may be able to manage his memory, actively deciding what information is retained or purged. Memory may be out of the control of the DM, and whether information

[3]Many of these models impose memory limitations and then the DM devises an optimal memory rule. See Wilson (2014), or other papers that use finite automata.

is remembered or forgotten is treated as exogenous. In reality, memory is a hybrid of these two, sometimes being automatic, other times within an individual's control.[4] In my model, I consider these two scenarios separately, first allowing the DM complete control over his memory and later treating memory as exogenous.

I analyze how different types of memory affect the DM and influence his optimal behavior in the Finitely Repeated Prisoner's Dilemma by determining how a DM's strategy selection and implementation are differentially affected by various memory types.

3.2.1 DM can control memory

The DM has memory m with capacity k (i.e., m has k memory "slots"). Upon observing a piece of information, i, he decides to store i, which uses up one of the slots, or forget i. Once all k slots are full, the DM can remove one of the is currently in m and replace it with a new i.

Unlimited memory capacity

I first consider $k = \infty$. Here the DM does not have any capacity constraints, so he chooses to store all information he observes. The result is that he remembers everything, can use any strategy, implements it successfully, and is able to learn his opponents' strategy, which allows for strategy adaptation (if necessary) in future supergames. In this case, efficiency is always achieved.

Finite Capacity

I now consider that the DM has finite memory, by having $k < \infty$. There are different implications based on the size of k. For example, if k is greater than sum of the total

[4]These types of memory are referred to as implicit (also called unconscious or automatic memory) and explicit (or declarative) memory.

number of rounds, then the DM will be able to remember everything over the course of the game. Since I want to analyze when this constraint has bite, I proceed assuming that k is less than the number of rounds in any single supergame.

As long as k is greater than or equal to the memory requirements of the DM's strategy, he will be able to implement it properly. Thus, he will select a strategy that has these memory requirements and will successfully implement it. The DM can also elect to use one of the k slots to reserve summarized past information about past games to be able to learn and adapt his strategy for future games, if necessary. Thus, depending on the precise size of k, the length of each supergame, and the number of supergames, the DM can still achieve an efficient outcome where he can successfully implement his strategy and adapt his strategy between supergames if he chooses.

However, we can end up at an inefficient outcome in certain cases. Take for example, $k = 1$. There are many memory-1 strategies, for example, Tit-for-Tat, Grimm, or any threshold strategy. Which of these strategies the DM should use is determined by what strategies the population of opponents is using. However, if the DM can only hold one piece of information, he will not be able to determine the strategy of his opponent, and thus will not know if his current strategy was a best response. As a result, he will not be able to learn which memory-1 strategy to select. For example, if the DM's opponents are using a threshold m strategy, the DM should adopt a threshold $m-1$ strategy, as this is a best response. However, since he cannot determine that his past opponents used a threshold m strategy, he will end up not best responding (unless through pure luck he happened to use a threshold $m-1$ strategy).

Further, take the extreme case of $k = 0$. The DM will select a memory-0 strategy (like Always Defect or Always Cooperate), and will be able to implement it, since it has no memory requirements. But since he has no memory, he cannot adapt his strategy between supergames based on the actions of his past opponent(s). He may decide to

change his strategy between supergames, but this would not be based on learning from past supergames. This means that inefficiencies can result. If the DM's opponents are playing other strategies, such as Tit-for-Tat or a threshold strategy, AD is not a best response. This means the DM is worse off than if he would be if he had a higher memory capacity. So although the DM selected a strategy that he can implement successfully, the outcome can still be inefficient.

Having limited memory capacity can additionally prevent the evolution of more complex strategies (Nowak and Sigmund, 1992, 1993; Axelrod, 1987; Lindgren, 1991; Hauert and Schuster, 1997). Since the DM is limited in his possible menu of strategies due to his limited memory, he is worse off than under the case of unlimited memory capacity. This is inefficient.

Memory miscoding

The case of memory miscoding is straightforward. It is never efficient for the DM to miscode information, so if he has control over what information is stored in memory, he will never choose to remember incorrect information.

3.2.2 Memory is exogenous

I now consider the case that memory or a memory rule may not be a choice variable. The decision maker (DM) has memory m. Upon observing a piece of information, i, the DM either remembers i with probability p_i or forgets it with probability $1-p_i$. Whether or not i is stored in memory is determined immediately. When analyzing the cases of finite memory, I assume that the DM is aware that his memory is finite, but is naive to the process that created his memories and always assumes that his memories are both accurate (i.e., the information that is in his memory is correct) and unbiased.

Perfect Memory

In the case of perfect memory, $p_i = 1, \forall i$. That is, the DM remembers every piece of information he observes. When a player has perfect memory, he can select any strategy, implement it successfully, and learn his opponent's strategy so he can determine if he should change or adapt his strategy before the next supergame. This is the same result as when there was perfect memory and the DM had control over his memory. As is in that case, efficiency is always achieved. Even though the DM's memory is exogenous here, we still achieve the efficient outcome.

Fading memories

I now consider that the DM may have memories that fade (decay) over time. That is, p_i now specifically depends on t, with only recent events being remembered perfectly and earlier events being forgotten. So $p_{it-r} = 1$ for $r < \hat{r}$ and $p_{it-r} = 0$ for $r > \hat{r}$.[5]

In this case, the DM must choose a strategy with a memory requirement less than or equal to his memory fading (less than $\hat{r}$). As long as the DM selects this type of strategy, he can implement it perfectly. The DM is still limited; however, in that if can only remember the previous $\hat{r}$ rounds, he may not be able to determine his opponent's strategy. Thus, he may not be able to determine if he is using the optimal strategy and may not select the best strategy given the frequency of strategies in the population.

For example, if a player is only able to remember his opponent's action in the previous round, he will only be able to properly implement at most a memory-1 strategy. But there are many memory-0 and memory-1 strategies in the FRPD. Which one he should use is based on what is the best response to what other strategies players are using. But if he cannot determine what strategy his first opponent is using, then he cannot learn if

[5]This could be extended so that more recent events are remembered with higher probability (but not perfectly) and/or earlier events are remembered with lower probability (but not always forgotten). I choose to analyze the binary case for simplicity.

his strategy is a best response to the player population and may not be able to adapt his strategy for future games. Thus, this memory limitation allows a player to implement a strategy successfully (as long as its memory requirement is below his memory limitation), but he may not be able to learn. In this case, the outcome can be inefficient.

Finite and Unbiased Memory

In this case, $p_i < 1$, meaning memory is finite. But $p_i = p, \forall i$, so it is unbiased, as the probability of remembering i is the same for all i and independent of any qualities or characteristics of i.

Assume the DM is using a memory-1 strategy. If he stored the information of his opponent's last move from the previous round, he can implement his strategy correctly in this round. If the information from the last round was not stored, the DM can reach back in his memory and recall what happened in previous rounds with this opponent that he did store. Since this information is unbiased, it gives an accurate depiction of his opponent. The DM can then make a prediction of what his opponent did in the previous round. Although this does not guarantee that the DM takes the correct action (conditional on his strategy), he can still do fairly well.

If the DM wants to guarantee that he can always successfully implement his strategy, he must select a memory-0 strategy. However, if these memory-0 strategies are not a best response to the population of opponents, this is inefficient.

Finite and Biased Memory

In this case, $p_i < 1$, but I now allow p_i to depend on i. Different characteristics that may affect p_i, include timing, beliefs, or salience.

Biased memory can result in the DM not being able to best respond to his opponent. For example, if m is such that information where the opponent defected is remembered with

higher probability than when she cooperated, the DM would conclude that his opponent was less cooperative than she actually was. As a result, the DM will defect on her more frequently than he should (given all her actions). If the DM forgets the last round, but remembers previous rounds when the opponent defected, he will best respond to that information and not to the information from the last round (which is necessary information for most FRPD strategies).

Moreover, if this bias is based on beliefs, where the DM is more likely to remember information that is consistent with his prior (i.e., confirmation bias), the beliefs of the DM can become increasingly polarized. These polarized beliefs can result in the DM not best responding, as he selectively forgets information that is inconsistent with his prior beliefs.

Unlike the case of finite but unbiased memory, the DM is unable to use past rounds that are in his memory to draw an unbiased conclusion about his opponent. Since I assume that the DM is naive to his memory bias, even if he tries to impute the missing memories from the memories that were successfully stored, he will neglect to correct for the bias in the memory storage process, and his inferences will consequently be biased.

Memory miscoding

As before, the probability that i is remembered is p_i, and the probability that i is forgotten is $1-p_i$. Previously, if the information was forgotten, no new information entered memory. I now allow for misinformation. That is, conditional on i being forgotten, with probability q_i a "false memory" is stored, denoted as e.

If memories are miscoded, inefficiencies will result. Not only is the DM not responding to all the information, as he forgets some of it, but he is now responding to incorrect information. This creates even more inefficiencies than just biased forgetting, as the DM is not aware that there is incorrect information. In the previous case, he attempts to adjust his beliefs knowing that there is missing information. But now, information is in his

memory, but this information is wrong, which further biases the inferences drawn from his memories.

Unlike simply finite memory, in the cases of biased memory and memory miscoding, merely using a strategy with lower memory requirements will not mitigate the problem. Now the DM cannot properly implement even simple strategies. He also cannot properly update based on the actions of his opponent to learn if he should adapt his strategy. He may end up getting stuck in a "memory trap" where he stays in a sub-optimal outcome because he incorrectly remembers the history.

3.2.3 Summary

As demonstrated in this section, unless memory is "perfect," in that it has both an unlimited capacity and is unbiased, inefficiency can occur. These inefficiencies are guaranteed when memory is biased. While it is a nice idea that decision makers are always able to optimally allocate finite resources, this may not be true of memory. As a result, decision makers can end up making sub-optimal choices. I ran an experiment in order to gather evidence of memory accuracy of a game's history, which I describe in the next section.

3.3 Experimental Design

3.3.1 The Finitely Repeated Prisoner's Dilemma

To study memory in a repeated game, I chose a familiar environment: the Finitely Repeated Prisoner's Dilemma.

The Finitely Repeated Prisoner's Dilemma (FRPD) is a canonical game in the theoretical literature, and therefore there is extensive experimental work using this game. Because of the number of studies using the FRPD, we have a large amount of information

about individuals' behavior, strategies that subjects likely employ, how behavior changes as subjects become more familiar with the game, and factors that affect cooperation (see Andreoni and Miller, 1993; Cooper et al., 1996; Dal Bó, 2005; Bereby-Meyer and Roth, 2006; Friedman and Oprea, 2012; Embrey et al., 2017). Because so much is known about subject behavior in this environment, it yields an ideal environment to examine individual recall of strategic interactions. I exploit what is known about the FRPD both in my experimental design and analysis.

In addition to its prevalence in the literature, a benefit of the FRPD is its simplicity. It is a 2×2 stage game that is repeated a known and limited number of times. Because of this simplicity, my results on recall provide an upper bound on memory in repeated games. While the theoretical assumption of perfect memory extends to games of any length and complexity, using the FRPD provides a clean test of the upper bound of recall we can expect in these environments.

The Stage Game

For the FRPD in my experiment, I leverage the setup used by Embrey, Fréchette, and Yuksel (2017), which is a meta-analysis of the FRPD.[6] They examine different factors that affect cooperation in the FRPD, and I select the condition that is most sustainable to cooperation (so that subjects will employ strategies besides always defect, which occurs in some of their other treatments). The payoffs and length of the game encourage cooperation. The stage game is depicted in Figure 3.1.[7] One supergame (called a "match" in the experiment) lasts for 8 rounds, meaning that subjects play the supergame 8 times with the same partner. Subjects play a total of two supergames.

[6] They peform a meta-analysis of existing studies of the FRPD and then conduct experiments to test the hypotheses they derive from their previous analysis.

[7] All payoffs are in Experimental Points, which are converted to dollars at the end of the experiment. The exchange rate for all sessions was 1 Experimental Point = \$0.005

		Other's Choice	
		A	B
Your Choice	A	51, 51	22, 63
	B	63, 22	39, 39

Figure 3.1: Stage Game in the Experiment

3.3.2 Procedures

All sessions were conducted at UCSD's EconLab using undergraduate students recruited via email. Instructions were read aloud to students and they interacted with each other solely through computers. The procedure for each session was as follows: After the instruction period,[8] subjects were randomly matched into pairs for the length of a repeated game (supergame). In each round of the supergame, subjects played the stage game. The length of the supergame was finite (8 rounds) and provided in the instructions so it was known to all subjects. After each round, subjects were shown their choice, the choice of their partner, and their payoff for that round.[9] Pairs were randomly rematched for the second supergame. After completing two supergames, subjects were asked to recall both their own action and the other's action for each round of both supergames. Then subjects were shown a series of coin flips (2 coins are flipped 16 times, divided into two sets of 8) and afterward they were asked to recall the outcomes of all 32 flips.[10]

At the end of a session, subjects were paid according to the total number of Experimental Points earned during the course of the experiment in addition to a $5 show-up fee for their participation. Subjects earned between $9 and $12 and sessions lasted

[8]During the instructional period subjects were told that the session would consist of three different tasks, but were not told exactly what these three tasks are. They were told that they will be given more detailed instructions about each task as it arose (so that recall was unanticipated). They were then given verbal instructions on the game.

[9]The outcome for the round was displayed once and could not be accessed again once the subject chose to move on to the next round. Subjects were also not allowed to take notes at any time during the course of the experiment, so they could not record the history.

[10]The coins were flipped before any of the sessions began, so all subjects observed the same sequence of coin flips.

approximately 30 minutes. A total of 4 sessions were conducted and 69 subjects are included in analysis.[11]

3.4 Results: General Trends

I first seek to establish some empirical facts. I examine general trends and patterns in the data to give a first look at recall in this environment.

For this purpose, I aggregate the data from all subjects together. In this analysis, I temporarily ignore that subjects are using different strategies and observing different histories. While this introduces some noise into the analysis, I am still able to determine some patterns to recall of game histories.

I focus my attention on two categorizes of general trends: limited capacity and memory biases. As shown in the theoretical framework, different memory limitations will have different behavioral implications. Thus, I want to first determine if there is evidence that memory has a limited capacity (i.e., not all observed information is stored) and second if there are some systematic biases in what types of information are stored.

3.4.1 Limited Capacity

To examine if memory has a limited capacity, I look at memory accuracy over a match and if histories that involve a player changing actions are recalled less accurately.

Recall Accuracy

I first explore subjects' recall accuracy over an entire match. I begin by examining recall of actions separately; that is, how accurately subjects recall their own actions and the other's actions individually. Figures 3.2 and 3.3 are frequency diagrams that reflect

[11] 70 subjects participated, but one was unable to complete the experiment and was dropped during analysis.

how many of each action a subject recalled correctly for a given match. There are not large differences between subjects' recall of their own action or the other's action, and there is slightly better recall in match 2 compared to match 1. What is notable is that approximately 50 percent of subjects recalled all 8 actions correctly for a given action in a given match. While the median accuracy for each action in each match is 8, or perfect, this is still a notable difference from the assumption of perfect memory for game histories, as being able to recall a given action with perfect accuracy does not apply to nearly half of subjects.

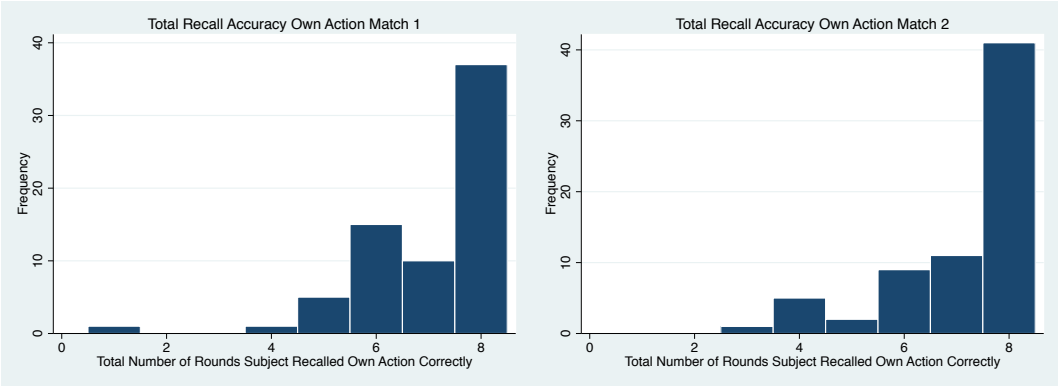


Figure 3.2: Total Recall Accuracy of Own Actions

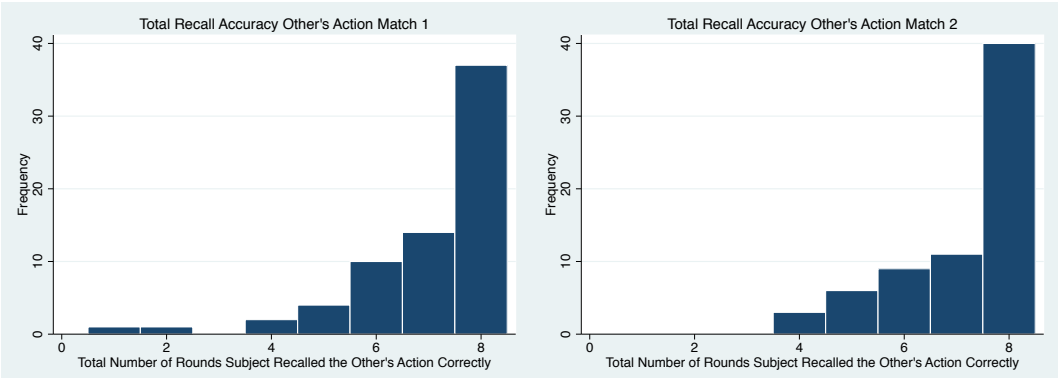


Figure 3.3: Total Recall Accuracy of Other's Actions

Recall of Outcomes

What the preceding analysis does not inform us of is whether the same individuals are getting all the actions correct in all rounds. To examine this question, I look first at the recall for outcomes. I define an individual as recalling the outcome of a round correctly if he recalls both his own action correctly and the other's action correctly for that round. Figure 3.4 is a frequency diagram that reflects how many outcomes a subject recalled correctly for a given match.

When looking at outcomes, approximately 1/3 of subjects recall all outcomes correctly for a given match. This is significantly lower than when I analyze actions separately, suggesting that a subject's recall of his own action and the other's action are independent to some degree. To analyze this further, I regress the probability that a subject recalls the other's action correctly on if he recalled his own action correctly, using a linear probability model. The results are shown in Table 3.1.[12]

While the coefficient is significant, it is 0.221 (looking at the model specification that includes round by match fixed effects). If subjects were recalling outcomes, meaning that recall of a subject's own action and the other's action are perfectly correlated, we would expect this coefficient to be (at least very close to) 1. An F-test to test if the coefficient is equal to 1 is strongly rejected (p=0.004), which suggests that recall of the actions is somewhat independent.

3.4.2 Changing Actions

There is large variation in the history the subjects observed. I use this variation to determine if histories where the subject (or the other) changed actions more times are recalled less accurately. I define changing actions or switching as if $a_r \neq a_{r-1}$, this is coded

[12]I also ran the same regression using a logit model for robustness. All results hold.

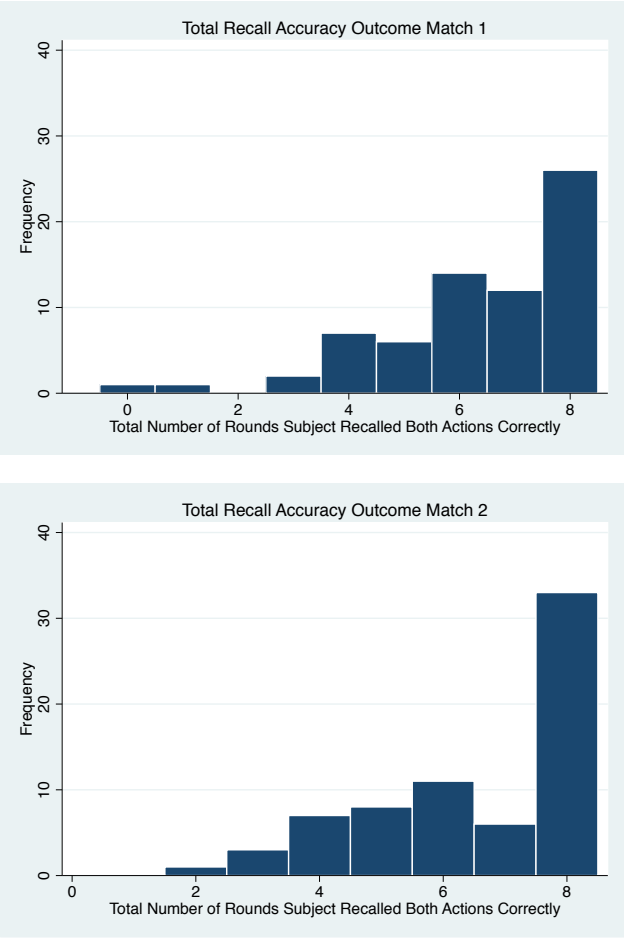


Figure 3.4: Total Recall Accuracy of Outcomes

Table 3.1: Linear Probability Model: Subject Recalled the Other's Action Correctly

	(1)	(2)	(3)
Subject Recalled Own Action Correctly	0.224**	0.224**	0.221**
	(0.0595)	(0.0589)	(0.0604)
Baseline Memory Task		-0.00458	-0.00430
		(0.0535)	(0.0539)
Constant	0.688***	0.691***	0.715***
	(0.0664)	(0.0817)	(0.0889)
Round × Match FEs	No	No	Yes
Observations	1,103	1,103	1,103

Clustered (at the session level) standard errors in parentheses. *** p<0.01, ** p<0.05, *p<0.1

as a switch (1) and 0 otherwise.

I run separate regressions for the total number of rounds a subject recalled his own action correctly on the number of switches and the total number of rounds a subject recalled the other's action correctly on the number of switches, using a censored tobit model. Note that 0 switches (meaning the action was the same over all 8 rounds) is omitted.[13] The results are summarized in Tables 3.2 and 3.4. Counts and group means associated with each number of switches are in Tables 3.3 and 3.5.

For both regressions, all coefficients are negative and significant. F-tests for equality of the coefficients are strongly rejected (F=93.1 for subject's recall of his own action and F=13.7 for subject's recall of the other's action). The resulting pairwise comparisons for equality, with one exception in each regression, are also rejected.[14] This is evidence that the number of times either player switched actions is strongly correlated with recall accuracy, with more switches decreasing recall accuracy.

[13]It may appear that 7 switches is the omitted variable, but that is because no subject switched 7 times, so this variable doesn't appear in the regression.

[14]I performed pairwise comparisons for 1 vs. 2, 2 vs. 3, 3 vs. 4, 4 vs. 5, and 5 vs. 6. In the first regression (recall of own action) all are rejected except 2 vs. 3. In the second regression (recall of other's action), all are rejected except 3 vs. 4.

Table 3.2: Total Number of Rounds Subject Recalled His Own Action Correctly

Number of times subject switched actions	Tobit
1	-2.016***
	(0.533)
2	-3.801***
	(0.764)
3	-4.584***
	(0.605)
4	-5.421***
	(0.762)
5	-7.254***
	(0.677)
6	-5.421***
	(1.346)
Constant	10.75***
	(0.677)
Observations	138

Clustered (at the session level) standard errors in parentheses. 0 switches is omitted.
*** p<0.01, ** p<0.05, * p<0.1

Taken together, these results suggest that memory has a limited capacity. Only 1/3 of subjects recall both actions correctly in every round of a given match. Additionally, histories where either the player or his partner changed actions more times are recalled less accurately (this suggests that histories with more different pieces of information are harder to remember). The next question is if memory is also biased.

3.4.3 Memory Biases

Timing Effects

I next examine how recall is related to the timing of the events. In contrast to the notion of memory decay (the idea that event further in the past are more likely to be

Table 3.3: Total Number of Rounds Subject Recalled His Own Action Correctly

Number of times subject switched actions	Count	Group Mean
0	36	7.97
1	47	7.65
2	22	6.64
3	16	6.13
4	12	5.33
5	2	3.5
6	3	5.33
7	0	–

forgotten) that is present in some bounded memory models, I don't find any evidence supporting that more recent events are remembered more accurately. In fact, when regressing the probability that an action in correctly on the round (using a linear probability model), the only significant coefficient are those associated with very early rounds. Specifically, only the coefficients on rounds 1 and 2 are significant for subjects recalling their own action or both actions, and round 1 is significant for subjects recalling the other's action (results in Table 3.6). None of the coefficients for match 2 are significant. Since round 8 is omitted, this is evidence that the first few rounds are *more* likely to be recalled correctly.

Recall of Particular Outcomes

I next analyze if subjects are more likely to recall particular outcomes. To examine this, I regress the probability that a subject recall's his own action correctly on all possible

Table 3.4: Total Number of Rounds Subject Recalled the Other's Action Correctly

Number of times the other switched actions	Tobit
1	-2.693***
	(0.461)
2	-3.760***
	(0.676)
3	-4.640***
	(0.929)
4	-5.353***
	(0.610)
5	-8.083***
	(0.784)
6	-6.416***
	(1.086)
Constant	11.08***
	(0.784)
Observations	138

Clustered (at the session level) standard errors in parentheses. 0 switches is omitted.
*** p<0.01, ** p<0.05, * p<0.1

outcomes: (C,C), (C,D), (D,C), and (D,D) using a linear probability model. Results are summarized in Table 3.7.[15] The only coefficient that is significant (and negative) is that associated with when the subject cooperated and the other defected. Note that both players defecting is the omitted variable in this specification, which means that relative to (D,D) a subject is less likely to recall the outcome (C,D). So fixing the fact that in both of these cases the subject plays defect, he is less likely to recall the outcome if he cooperated as opposed to also defecting.

I run this regression again, using the probability that the subject recalls the other's action correctly on all possible outcomes. The results are summarized in Table 3.8. While in the first specification the coefficient on both subjects having cooperated is significant in

[15] As before, a logit model was run for robustness. All results hold.

Table 3.5: Total Number of Rounds Subject Recalled the Other's Action Correctly

Number of times the other switched actions	Count	Group Mean
0	36	7.97
1	46	7.43
2	24	6.96
3	15	6.27
4	12	5.67
5	2	3
6	3	4.67
7	0	–

addition to the coefficient on subject defected and the other cooperated, the former loses significance after fixed effects are introduced. The coefficient on subject defected and other cooperated is negative, meaning that the subject is less likely to recall the other's action correctly when he defected on a cooperator.

This second result is in line with some very recent results on what has been deemed "unethical amnesia," which suggest that memories of unethical actions become obfuscated over time (Kouchaki and Gino, 2016). However, if unethical amnesia was driving the results, then subjects should also be more likely to forget their own action when they defected on a cooperator.

To explore this further, I look at the types of errors that are made. Of all of the (C,D) outcomes, 55 were recalled incorrectly. Of these 55 errors, 34 were misremembered as (D,D). Similarly, of all of the (D,C) outcomes, 54 were recalled incorrectly and 34 of

Table 3.6: Linear Probability Models

	Recalling Own Action	Recalling Other's Action	Recalling Both Actions
round 1 match 1	0.0784*	0.0664**	0.131***
	(0.0295)	(0.0138)	(0.0204)
round 2 match 1	0.0660**	0.0674	0.104**
	(0.0163)	(0.0326)	(0.0284)
round 3 match 1	0.0324	-0.0357	-0.0187
	(0.0198)	(0.0347)	(0.0601)
round 4 match 1	-0.0229	-0.0349	-0.0584
	(0.0370)	(0.0347)	(0.0282)
round 5 match 1	-0.0569	0.0197	-0.0515
	(0.0462)	(0.0481)	(0.0825)
round 6 match 1	-0.0238	-0.0654	-0.104
	(0.0231)	(0.0642)	(0.0625)
round 7 match 1	-0.0683	0.00736	-0.0322
	(0.0376)	(0.0362)	-0.0322
Outcome Controls	Yes	Yes	Yes
Individual FEs	Yes	Yes	Yes
Observations	1,104	1,104	1,104

Clustered (at the session level) standard errors in parentheses. Logits run for robustness; results hold. *** p<0.01, ** p<0.05, * p<0.1

these were misremembered as (D,D). This suggests that subjects are recalling many more (D,D) outcomes than there actually were. Thus, when the outcome is in reality (C,D), they recall their own action incorrectly, and when the outcome is in reality (D,C), they recall the other's action incorrectly.

Table 3.7: Linear Probability Model: Recalling Own Action Correctly

	(1)	(2)	(3)	(4)
Both Cooperated	0.0282	0.0265	0.0182	-0.00191
	(0.0251)	(0.0290)	(0.0324)	(0.0521)
Subject Cooperated, Other Defected	-0.210***	-0.213***	-0.233***	-0.219***
	(0.0198)	(0.0215)	(0.0198)	(0.0233)
Subject Defected, Other Cooperated	-0.0455	-0.0458	-0.0664	-0.0267
	(0.0676)	(0.0665)	(0.0744)	(0.0873)
Baseline Memory Task		0.103	0.105	-0.217***
		(0.165)	(0.166)	(0.0208)
Round × Match FEs	No	No	Yes	Yes
Individual FEs	No	No	No	Yes
Observations	1,103	1,103	1,103	1,103

Clustered (at the session level) standard errors in parentheses. Logits run for robustness; results hold. *** p<0.01, ** p<0.05, * p<0.1

Table 3.8: Linear Probability Model: Recalling Other's Action Correctly

	(1)	(2)	(3)	(4)
Both Cooperated	0.0103	0.0101	0.00622	0.00225
	(0.0192)	(0.0200)	(0.0222)	(0.0233)
Subject Cooperated, Other Defected	-0.0603**	-0.0607**	-0.0654	-0.0416
	(0.0153)	(0.0172)	(0.0283)	(0.0255)
Subject Defected, Other Cooperated	-0.218***	-0.218***	-0.224***	-0.211***
	(0.0239)	(0.0239)	(0.0347)	(0.0235)
Baseline Memory Task		0.0129	0.0135	0.635***
		(0.0701)	(0.0713)	(0.0190)
Round ×Match FEs	No	No	Yes	Yes
Individual FEs	No	No	No	Yes
Observations	1,103	1,103	1,103	1,103

Clustered (at the session level) standard errors in parentheses. Logits run for robustness; results hold. *** p<0.01, ** p<0.05, * p<0.1

3.4.4 Summary of Results so Far

The results thus far suggest that memory is both finite and biased. While recall of actions is good, it is far from perfect, and gets worse if we consider outcomes rather than individual actions. This phenomenon seems to be due to the recall of subject's own action and the other's action only being weakly correlated. Further, histories that contain more, different information (as measured by the number of times either player switched actions) are recalled less accurately. Memory also appears to be biased in that certain characteristics of the information are associated with its likelihood of being remembered correctly. Specifically, the timing of information (primacy effects) and the type of outcome are related to recall accuracy. Further, there is a systematic bias in the way that the forgotten information is subsequently recalled, as, conditional on the outcome being recalled incorrectly, subjects are significantly more likely to recall the outcome as (D,D).

Since recall in repeated games has not been studied empirically before by economists, it is important to explore general trends and patterns of recall in this environment. However, how memory is related to behavior is a key quesetion. In the following section, I examine how memory is related to strategies.

3.5 Results: Strategy

3.5.1 Strategy Estimation

Since my data includes only observed behaviors, I need to impute the strategies used by subjects. I develop a method to estimate what strategy a subject may be using. While I do not claim to know that subjects are definitely using this strategy (since I observe only actions and don't explicitly elicit strategies), I can determine which strategy best rationalizes the observed behavior of each subject.

Since there are an infinite number of possible strategies but the data are finite, it is impossible to identify the strategies used by subjects. To mitigate this problem, I restrict my attention to strategies that are relevant based on the theoretical and experimental literature: Always Defect (AD), Always Cooperate (AC), Grim (G), Tit for Tat (TFT), Win Stay Lose Shift (WSLS), Threshold 2-8 Strategies (TH2-TH8), and a trigger strategy with two periods of punishment (T2).[16]

I allow for individuals to make mistakes and take an action that is not prescribed by the strategy. I analyze each match separately in order to consider that subjects may switch strategies when beginning a new supergame.

The procedure is as follows: I generate data for each subject based on the aforementioned strategies and conditional on the actual history he observed. Thus, these data are what the subject should have done, conditional on the history, if he was using that strategy.

Formally, I denote the action taken by subject i in round r of a match m by a_{imr} and the action that the strategy k prescribes in that round when he is matched with subject j by $s^k_{imr}(a_{jm1},...,a_{jm(r-1)};a^k_{im1},...,a^k_{im(r-1)})$ if $r>1$. Then, for each match, I find the strategy, s^{k*}_{im}, that minimizes $\sum_{r=1}^{8}|a_{imr}-s^k_{imr}|$.[17]

I allow for the "errors"[18] for the selected strategy ($\sum_{r=1}^{8}|a_{imr}-s^{k*}_{imr}|$) to be at most 3. Since the total number of rounds in a match is 8, 4 errors is the same number as predicted by random chance, so I assume that the subject is not using any of the strategies under consideration. If his estimated strategy has 4 errors I code him as having "no strategy".[19]

[16]WSLS starts by cooperating and then conditions behavior on the outcome in the previous round. If both players cooperated or neither player cooperated, then WSLS cooperates, and it defects otherwise. T2 starts by cooperating and a defection by the other player triggers two periods of defection, after which it goes back to cooperating. Threshold strategies are strategies that conditionally cooperate until a predetermined period, m, at which point it always defects. It can thus be seen as a combination of Grim and AD.

[17]I allow for ties if there is more than one strategy that minimizes this expression. If there is a tie, I assume that the subject could be using any of the tied strategies.

[18]Errors are the number of times a subject deviated from the strategy estimated for him.

[19]It may not actually be the case that the subject is not using *any* strategy, because in principle there are an infinite number of strategies. I simply use this term to denote that the subject is not using one of

I perform this procedure again, but instead use the subject's recalled history. I use his recall of his own action for his choices and his recall of the other's action as the conditional history.

This strategy estimation method is a new application of a least absolute deviation regression, or, since the data are binary choices, non-linear least squares. In the least absolute deviation regression, we are looking for an $f(x)$ that best fits a set of observations. So the problem can be formulated as finding an $f(x)$ to minimize $S = \sum_{i=1}^{n} |y - f(x_i)|$. In my procedure, the strategies in my consideration set are the possible $f(x)$s I am testing.[20]

There exist other strategy estimation methods in the literature, namely the Strategy Frequency Estimation Method (SFEM) developed by Dal Bó (2005). In their method, they find an estimation of the importance of strategies using maximum likelihood, meaning they find the likelihood that the data as a whole corresponds to a given strategy. My estimation procedure estimates something different: it estimates the strategy for an individual subject in a single supergame. I am also able to weaken the two key assumptions of SFEM: that all subjects have a given probability of using one of the six strategies they consider and that subjects don't change strategies from repeated game to repeated game.

3.5.2 Stategy Estimation Results

I calculate the number of subjects for whom each strategy was the one that best rationalized their behavior. These results are summarized in Tables 3.9 and 3.10. The columns add up to more than 69, because of my allowing for ties. Most of the ties are concentrated in Grim and the threshold strategies, as if a subject defects following a defection by the other in the previous round early in the game, this is consistent with Grim as well as any threshold that comes after the round of first conditional defection. The

the strategies that I am considering in this analysis.

[20]Since my choice data are binary, the absolute deviations are equivalent to the squared deviations, so this is equivalent to non-linear least squares.

most frequently used strategies are strategies with very low memory requirements, as the two most frequent strategies are Always Defect and Always Cooperate, both of which are memory-0, and the third and fourth most common are Grim and Tit-for-Tat, which are both memory-1.[21]

Table 3.9: Stategy Frequency (Based on Actual History)

	Match 1		Match 2	
Strategy	Number of subjects estimated strategy	Percent of subjects estimated strategy	Number of subjects estimated strategy	Percent of subjects estimated strategy
AD	23	33.3	22	31.9
AC	14	20.3	16	23.2
G	19	27.5	27	39.1
TFT	14	20.3	31	44.9
WSLS	11	15.9	12	17.4
TH2	12	17.4	13	18.8
TH3	12	17.4	15	21.7
TH4	12	17.4	15	21.7
TH5	12	17.4	16	23.2
TH6	12	17.4	15	21.7
TH7	13	18.8	18	26.1
TH8	15	21.7	18	26.1
T2	12	17.4	13	18.8
NS	4	5.8	1	1.4

Note that these numbers add up to more than 69 because I allow for ties.

Table 3.10: Strategy Frequency (Based on Recalled History)

	Match 1		Match 2	
Strategy	Number of subjects estimated strategy	Percent of subjects estimated strategy	Number of subjects estimated strategy	Percent of subjects estimated strategy
AD	22	31.9	22	31.9
AC	13	18.8	15	21.7
G	21	30.4	30	43.5
TFT	21	30.4	31	44.9
WSLS	10	14.5	12	17.3
TH2	15	21.7	15	21.7
TH3	13	18.8	17	24.6
TH4	13	18.8	20	29.0
TH5	14	20.3	20	29.0
TH6	16	23.2	19	27.5
TH7	14	20.3	22	31.9
TH8	17	24.6	20	29.0
T2	21	30.4	15	21.7
NS	0	0	0	0

Note that these numbers add up to more than 69 because I allow for ties.

3.5.3 "Switching" Stategies

Since I run my estimation procedure both on the observed history and the recalled history, I can compare the strategies estimated under each condition. I compare which strategy is estimated for the subject using the actual history and see if this is the same strategy that is estimated under the recalled history. For match 1, 11 subjects (15.9%)

[21] Dal Bó (2005) find using their strategy estimataion procedure that together Always Defect and TFT can explain the vast majority of their data.

have a different strategy between these two estimations, and for match 2, 13 (18.8%) of subjects have a different estimated strategy if look at actual history and recalled history.

The most interesting way that subjects switch strategies is if we focus on the "no strategy" category. 4 subjects are categorized as having no strategy in match 1 and 1 is categorized as having no strategy in match 2. But once I estimate strategies based on the recalled history, 0 subjects recall having no strategy. That is, subjects appear to recall having more pattern or order in their behavior than they actually had.

3.5.4 Allowing Strategy Flexibility

I now want to compare how many times a subject deviated from his estimated strategy. I first allow for strategy flexibility. This means that I allow subjects to have a different strategy estimated by the actual history and based on the recalled history. I then compare the number of strategy implementation "errors" made between these two estimated strategies. The results are summarized in Table 3.11.

In match 1, 14 subjects had negative differences between their strategy implementation errors for their recalled strategy and their observed strategy. That means that 14 subjects believed that they implemented a strategy better than they did in reality. This is twice the number of subjects who believed they made more mistakes than they actually did (7), and this difference is only marginally insignificant ($\chi^2 = 2.333$, $p = 0.1226$). There is a more even distribution if we look at match 2, with 7 subjects making fewer errors in their recalled strategy and 9 making more mistakes in their recalled strategy ($\chi^2 = 0.250$, $p = 0.6171$).

Since I allow for strategy flexibility, some of these subjects have different strategies under the two estimations (actual history vs. recalled history). Thus, subjects are sometimes recalling behavior that is better rationalized by a different strategy, but in either case, a large proportion of subjects recall implementing a strategy more successfully than they

actually did.

Table 3.11: Difference in Strategy Implementation "Errors"

	Match 1	Match 2
"Errors" in Recalled Strategy minus "Errors" in Observed Strategy	Count	Count
-2	1	2
-1	13	5
0	48	53
1	6	8
2	1	1

3.5.5 Fixing the Stategy

I repeat this comparison, but now I fix subjects' strategies. That is, I estimate the strategy under the observed history, and assume that a subject is using that strategy.[22] I fix their strategy as the one estimated from the actual history, so I prevent subjects from "switching," which I previously allowed. I then compare the difference in strategy implementation mistakes actually made and those the subject recalled making for this same strategy. The results are summarized in Table 3.12.

There are now many more subjects making more mistakes under recall (13 vs. 6 for match 1–$\chi^2 = 0.1083$–and 14 vs. 2 for match 2–$\chi^2 = 9.00$, $p = 0.0027$). This stark change from when I allow flexible strategies is stemming from the result that a sizable of subjects are recalling using a different strategy entirely.

[22]This is a slight abuse of terminology. By "use," I do not mean mean to imply that the subject is definitely using this strategy (with my data there is no way to definitively say that). What I mean is that I take that estimated strategy as given so I can compare the observed mistakes under that estimated strategy and the recalled mistakes under that same strategy.

Table 3.12: Difference Between Recalled Mistakes and Observed Mistakes in Strategy Implementation for a Fixed Strategy

	Match 1	Match 2
Recalled mistakes minus observed mistakes	Count	Count
-2	1	0
-1	5	2
0	46	51
1	10	11
2	3	3

3.5.6 Recall and Strategy Errors

I now examine how strategy implementation errors are associated with recall of the game history.

I run regressions on the number of strategy implementation errors made (based on the observed strategy) on the subject's recall of the game history. I run separate regressions for match 1 and match 2 and use a censored tobit model. Results are summarized in Tables 3.13 and 3.14.

The coefficient associated with the total number of rounds subject recalled own action correctly is significant and negative for both regressions, so the results are robust across the matches. What is more surprising about this regression result is that the coefficient on other's action is *not* significant. Since most of the strategies are memory-1, they rely only on recalling the other's action correctly and not the action of the player. Thus, it is surprising that only the coefficient on the subject's own action would be significant. An alternative explanation of this is that there is reverse causality. If a subject is better at implementing his own strategy, he has a clearer pattern to remember with fewer "unexplainable" actions to have to keep track of. Thus, it may be that those who implement their strategies better have a easier sequence to recall. The problem with this explanation, however, is if a player is playing a particular strategy and he implemented it properly, he should be able to back out the other's action. For example, if he is playing Tit-for-Tat and he implemented it

correctly, he should be able to determine what his partner did in any given round, because it is whatever he did in the next round.

Table 3.13: Number of Strategy Implementation Errors in Match 1 (based on observed strategy)

	Tobit
Total number of rounds subject recalled the other's action correctly in match 1	-0.00227
	(0.113)
Total number of rounds subject recalled own action correctly in match 1	-0.319***
	(0.0819)
Constant	2.917***
	(0.461)
Strategy Controls	Yes
Observations	65

Clustered (at the session level) standard errors in parentheses.
*** p<0.01, ** p<0.05, * p<0.1

3.6 Conclusion

The results of this paper suggest that memory is both finite and biased and is associated with a player's ability to successfully implement a strategy. Players appear to organize their behavior in their memory, creating patterns that better fit into strategies than their observed behavior would suggest. This is best evidenced by when estimating strategies, I find subjects who appeared to have no strategy under the consideration set when I look at their behavior in the game all have strategies as determined by my estimation procedure in their memory.

Players being able to recall an entire history of a repeated game is a common implicit assumption of many models. This assumption is made irrespective of how long

Table 3.14: Number of Strategy Implementation Errors in Match 2 (based on observed strategy)

	Tobit
Total number of rounds subject recalled the other's action correctly in match 2	0.0104
	(0.0598)
Total number of rounds subject recalled own action correctly in match 2	-0.422***
	(0.0775)
Constant	3.946***
	(0.666)
Strategy Controls	Yes
Observations	68

Clustered (at the session level) standard errors in parentheses. Includes strategy controls (the estimated strategy based on subject's actual choices for that match). *** p<0.01, ** p<0.05, * p<0.1

or complicated the history may be. Recent models have begun to relax this assumption, allowing players to have bounded memory for the history. While these new models have developed, rigorous empirical and experimental testing of how individuals recall histories in these environments has not been equally prevalent. This paper provides empirical evidence of recall inaccuracies in the simple, canonical environment of the Finitely Repeated Prisoner's Dilemma. Given my results, examining memory in more complex strategic environments and its implications on subjects' ability to learn, adapt, best respond, and payoffs is an important area of future research.

Chapter 3, in full, is currently being prepared for submission for publication of the material. Giffin, Erin. The dissertation author is the sole author of this material.

References

(1973). Lehnhausen v. Lake Shore Auto Parts Co.

Akerlof, G. A. (1976). The Economics of Caste and of the Rat Race and Other Woeful Tales. *The Quarterly Journal of Economics 90*(4), 599–617.

Akerlof, G. A. and R. E. Kranton (2000). Economics and Identity. *Quarterly Journal of Economics CXV*(3), 715–753.

Andreoni, J. and B. D. Bernheim (2009). Social Image and the 50-50 Norm: A Theoretical and Experimental Analysis of Audience Effects. *Econometrica 77*(5), 1607–1636.

Andreoni, J. and J. H. Miller (1993). Rational Cooperation in the Finitely Repeated Prisoner ' s Dilemma : Experimental Evidence. *The Economic Journal 103*(418), 570–585.

Andreoni, J. and J. M. Rao (2011). The power of asking: How communication affects selfishness, empathy, and altruism. *Journal of Public Economics 95*(7-8), 513–520.

Andreoni, J. and L. Vesterlund (2001). Which Is the Fair Sex? Gender Differences in Altruism. *Quarterly Journal of Economics 116*(February), 293–312.

Axelrod, R. (1987). Evolving New Strategies The Evolution of Strategies in the iterated Prisoner's Dilemma. *The dynamics of norms*, 1–16.

Babcock, L. and S. Laschever (2003). *Women Don't Ask: Negotiation and the Gender Divide*. Princeton University Press.

Babcock, L., M. P. Recalde, L. Vesterlund, and L. Weingart (2017). Gender differences in accepting and receiving requests for tasks with low promotability. *American Economic Review 107*(3), 714–747.

Bereby-Meyer, Y. and A. E. Roth (2006). The Speed of Learning in Noisy Games : Partial Reinforcement and the Sustainability of Cooperation. *The American Economic Review 96*(4), 1029–1042.

Bertrand, M., C. Goldin, and L. F. Katz (2010). Dynamics of the Gender Gap for Young Professionals in the Corporate and Financial Sectors. *American Economic Journal: Applied Economics 2*(3), 228–255.

Bertrand, M., E. Kamenica, and J. Pan (2015). Gender Identity and Relative Income Within Households. *The Quarterly Journal of Economics 130*(2), 571–614.

Bharadwaj, P. and J. B. Cullen (2017). Coming of Age: Timing of Adolescence and Gender Identity Formation.

Bolton, G. E. and E. Katok (1995). An experimental test for gender differences in beneficent behavior. *Economics Letters 48*, 287–292.

Bull, J. (2008). Costly evidence production and the limits of verifiability. *The BE Journal of Theoretical Economics 8*(1).

Bull, J. and J. Watson (2004). Evidence disclosure and verifiability. *Journal of Economic Theory 118*(1), 1–31.

Ceci, S. J., D. K. Ginther, S. Kahn, and W. M. Williams (2014). Women in Academic Science. *Psychological Science in the Public Interest 15*(3), 75–141.

Cho, I.-K. and D. M. Kreps (1987). Signaling games and stable equilibria. *The Quarterly Journal of Economics 102*(2), 179–221.

Coate, S. and G. C. Loury (1993). Will Affirmative-Action Policies Eliminate Negative Stereotypes? *The American Economic Review 83*(5), 1220–1240.

Cole, H. L., G. J. Mailath, and A. Postlewaite (1992). Social Norms , Savings Behavior , and Growth. *Journal of Political Economy 100*(6), 1092–1125.

Cooper, R., D. V. DeJong, R. Forsythe, and T. Ross (1996). Cooperation without Reputation : Experimental Evidence from Prisoner 's Dilemma Games. *Games and Economic Behavior 218*(12), 187–218.

Cooter, R. D. and D. L. Rubinfeld (1994). An economic model of legal discovery. *The Journal of Legal Studies*, 435–463.

Croson, R. and U. Gneezy (2009). Gender Differences in Preferences. *Journal of Economic Literature 47*(2), 448–474.

Dal Bó, P. (2005). Cooperation under the Shadow of the Future: Experimental Evidence from In nitely Repeated Games. *American Economic Review 95*(5), 1591–1604.

Dal Bó, P. and G. Fréchette (2011). The Evolution of Coorporation in Infinitely Repeated Games: Experimental Evidence. *The American Economic Review 100*(1), 411–429.

Daughety, A. F. and J. F. Reinganum (2000). On the economics of trials: adversarial process, evidence, and equilibrium bias. *Journal of Law, Economics, and Organization 16*(2), 365–394.

Davis, M. L. (1994). The value of truth and the optimal standard of proof in legal disputes. *Journal of Law, Economics, and Organization 10*, 343.

DellaVigna, S., J. A. List, U. Malmendier, and G. Rao (2013). The Importance of Being Marginal. *American Economic Review: Papers and Proceedings 103*(3), 586–590.

Demougin, D. and C. Fluet (2008). Rules of proof, courts, and incentives. *The RAND Journal of Economics 39*(1), 20–40.

Eckel, C. C. and P. J. Grossman (1998). Are Women Less Selfish Than Men?: Evidence from Dictator Experiments. *The Economic Journal 108*(448), 726–735.

Ehrlich, I. and R. A. Posner (1974). An economic analysis of legal rulemaking. *The Journal of Legal Studies 3*(1), 257–286.

Embrey, M., G. Fréchette, and S. Yuksel (2017). Cooperation in the Finitely Repeated Prisoner 's Dilemma.

Fragiadakis, D. E., D. T. Knoepfle, and M. Niederle (2013). Identifying Predictable Players : Relating Behavioral Types and Subjects with Deterministic Rules.

Friedman, D. and R. Oprea (2012). A Continuous Dilemma. *American Economic Review 102*(1), 337–363.

Friendly, H. J. (1975). Some kind of hearing. *University of Pennsylvania Law Review 123*(6), 1267–1317.

Froeb, L. M. and B. H. Kobayashi (1996). Na{\"\i}ve, biased, yet Bayesian: can juries interpret selectively produced evidence? *The Journal of Law, Economics, and Organization 12*(1), 257–276.

Gelman, A. and J. Hill (2006). Missing-data imputation. In *Data Analysis Using Regression and Multilevel/Hierarchical Models*, Chapter 25. Cambridge University Press.

Gould, J. P. (1973). The economics of legal conflicts. *The Journal of Legal Studies 2*(2), 279–300.

Hauert, C. and H. G. Schuster (1997). Effects of increasing the number of players and memory size in the iterated Prisoner's Dilemma: a numerical approach. *Proceedings of the Royal Society B: Biological Sciences 264*(1381), 513–519.

Hay, B. L. (1994). Civil discovery: Its effects and optimal scope. *The Journal of Legal Studies 23*(S1), 481–515.

Henry, E. (2009). Strategic disclosure of research results: The cost of proving your honesty. *The Economic Journal 119*(539), 1036–1064.

Huang, P. H. and H.-M. Wu (1994). More Order without More Law: A Theory of Social Norms and Organizational Cultures. *Journal of Law, Economics, & Organization 10*(2), 390–406.

Huck, S. and W. Müller (2002). Absent-minded Drivers in the Lab: Testing Gilboa's Model. *International Game Theory Review 4*(4), 435–448.

Ivanov, A., D. Levin, M. Niederle, S. Econometrica, and N. July (2010). Can Relaxation of Beliefs Rationalize the Winner's Curse?: An Experimental Study. *Econometrica 78*(4), 1435–1452.

Kandori, M. (1992). Social Norms and Community En orcement. *The Review of Economic Studies 59*(1), 63–80.

Kim, C. (2013). Adversarial and inquisitorial procedures with information acquisition. *The Journal of Law, Economics, & Organization 30*(4), 767–803.

Kouchaki, M. and F. Gino (2016). Memories of unethical actions become obfuscated over time. *Proceedings of the National Academy of Sciences* (5), 1–6.

Landes, W. M. (1971). An economic analysis of the courts. *The Journal of Law and Economics 14*(1), 61–107.

Lester, B., N. Persico, and L. Visschers (2009). Information Acquisition and the Exclusion of Evidence in Trials. *The Journal of Law, Economics, & Organization 28*(1), 163–182.

Lewis, T. and M. Poitevin (1997). Disclosure of information in regulatory proceedings. *The Journal of Law, Economics, and Organization 13*(1), 50–73.

LIMRA (2016). Men vs. Women: Who makes the financial decisions?

Lindgren, K. (1991). Evolutionary phenomena in simple dynamics.

Macey, J. R. (1994). Judicial preferences, public choice, and the rules of procedure. *The Journal of Legal Studies 23*(S1), 627–646.

Milgrom, P. and J. Roberts (1986). Relying on the information of interested parties. *The RAND Journal of Economics*, 18–32.

Monte, D. (2013). Bounded memory and permanent reputations. *Journal of Mathematical Economics 49*(5), 345–354.

National Center for Education Statistics (2016). Bachelor's, master's, and doctor's degrees conferred by postsecondary institutions, by sex of student and discipline division.

Nowak, M. A. and K. Sigmund (1992). Tit for tat in heterogeneous populations. *Nature 355*(6357), 250–253.

Nowak, M. A. and K. Sigmund (1993). A strategy of win-stay-lose-shift that outperforms tit-for-tat in the Prisoner's Dilemma game.

Persico, N. (2012). Evidence of Discrimination. *The Journal of Legal Studies 41*(2), 321–346.

Saccardo, S., A. Pietrasz, and U. Gneezy (2017). On the Size of the Gender Difference in Competitiveness. *Management Science*.

Sanchirico, C. W. (1997). The burden of proof in civil litigation: A simple model of mechanism design. *International Review of Law and Economics 17*(3), 431–447.

Sapienza, P., L. Zingales, and D. Maestripieri (2009). Gender differences in financial risk aversion and career choices are affected by testosterone. *Proceedings of the National Academy of Sciences 106*(36), 15268–15273.

Shavell, S. (1982). Suit, settlement, and trial: A theoretical analysis under alternative methods for the allocation of legal costs. *The Journal of Legal Studies 11*(1), 55–81.

Shin, H. S. (1994). The burden of proof in a game of persuasion. *Journal of Economic Theory 64*(1), 253–264.

Stephenson, M. C. (2008). Evidentiary Standards and Information Acquisition in Public Law. *American Law and Economics Review 10*(2), 351–387.

Thomas, D. (1990). Intra-Household Resource Allocation: An Inferential Approach. *The Journal of Human Resources 25*(4), 635–664.

Wilson, A. (2014). Bounded Memory and Biases in Information Processing. *Econometrica 82*(6), 2257–2294.

Wong, T.-N. and L. L. Yang (2015). When Monitoring Hurts: Endogenous Information Acquisition in a Game of Persuasion.