Mechanisms underlying Economic Choice

Thesis by

Gideon Nave

In Partial Fulfillment of the Requirements for the
degree of
PhD

CALIFORNIA INSTITUTE OF TECHNOLOGY
Pasadena, California

2016
Defended 5/6/2016

Gideon Nave
ORCID: 0000-0001-6251-5630

In memory of Charlotte Prevost, 1985-2016.

ACKNOWLEDGEMENTS

Six years ago, no one (myself included) would have guessed that my future career would be in the Social Sciences. While working on my masters in Electrical Engineering at the Technion, I was suffering from a broken heart and a lack of meaning in my life. I have pledged to find a career path that excited me but had no idea where to start. Sitting in a digital communication lecture, I have learnt how optimal digital receivers use Bayesian principles when "deciding" how to convert noisy analog signals into binary digits. An interesting idea jumped into my mind: could it be that our brains use similar principles when deciding what to do in the chaotic world we live in? I was somewhat disappointed when I found out that others had thought of this idea before. But despite the lack of novelty, this was a defining moment. I was blessed to stumble upon a topic that excited me, and realized that I had just touched the tip of an iceberg. Those were the early days of Neuroeconomics, an entire scientific discipline dedicated to understanding how the human brain makes decisions in uncertain environments.

I will never forget my first meeting with Professor Colin Camerer, at the CNS interview weekend at Caltech. I had never had a chance to chat with such a prominent scientific figure, and was surprised when Colin, wearing sneakers and a t-shirt, shook my hand and introduced himself as if it wasn't obvious that I knew who he was. Working with Colin was a life changing experience. I was lucky to have a mentor who has endless curiosity and openness, that allowed me to pursue some of the crazy ideas I had. Thanks to Colin's trust and generosity, I had the opportunity to run pharmacological experiments in unprecedentedly large samples, study how the presence of a provocatively dressed female experimenter affects the experience of wine tasting in Parisian men, and try to investigate how Xenopus frogs make food choices. Colin taught me that being a world-class scientist is not at odds with kindness, modesty and a great sense of humor.

Colin is not an outlier among the professors I have interacted with at Caltech. I very much enjoyed the interaction with the other members of my dissertation committee, Ralph Adolphs, John O'doherty and Shin Shimojo. I always felt that their doors were open, and was lucky to get their advice when needed and present my work to their research groups. I had similar experiences with the other 'Neuroeconomics' CNS faculty, Antonio Rangel and Peter Bossaerts, and with the directors of the program, Pietro Perona and Thonas Siapas. The pleasant interactions with Caltech

staff, especially Barbara Estrada, Tanya Owen, Laurel M. Auchampaugh and Tiffany Kim gave me the essential peace of mind to make my research happen.

There is a seemingly endless list of colleagues and friends who greatly contributed (and continue to contribute) to my personal and professional development. I am thankful for many hours of conversation with Juri Minxha, who has been my roommate during my first three years at Caltech. I was fortunate to have a friend like Juri, who has the rare combination of sharp intelligence, frank sensitivity and lightness. I was blessed with many new friends who have also turned into superb research collaborators during my time at Caltech. Three of them remarkably contributed to this dissertation. Chapter 2 is the fruit of a my work with Cary Frydman, whose last year of grad school overlapped with my first year one. I was fortunate to share an office with Alec Smith. I have learned a great deal of Econometrics and Game Theory by merely sitting with Alec in the same room, and chapter 3 is the fruit of our close collaboration with Colin. The friendships of Cary and Alec greatly helped me to navigate through the loneliness and difficulties of moving from the busy streets of Tel-Aviv to the lively but unwalkable City of Angels.

Chapter 2 of my dissertation is the first product of a joint research program with Amos Nadler. Although Amos and I grew up just a few blocks from each other in Ahuza neighborhood in Haifa, we met for the first time only a few years ago. Amos was especially fun to work with, and he taught me a great deal about conducting hormonal studies.

Last but not least, my parents' faith and unconditioned love, especially in times of uncertainty and struggle, gave me the confidence to unapologetically pursue happiness in my personal and professional lives. Their support and generosity, along with the encouragement and sensitivity of my sisters Tal and Orly and their partners Shiran and Tom, and the joyfulness of my niece Eliya and my nephews Omri, Omer and Daniel, is the fortress from which I embarked on this journey, and the place where I will always come back to.

ABSTRACT

The current dissertation proposes three manners in which findings about the neuroscience of decision-making can inform traditional questions in economics that historically has been investigated using choice data alone, and without delineating the mechanism of choice.

The first chapter investigates the origins of a critical component of both economic and perceptual decision-making under uncertainty: the belief formation process. Most research has studied belief formation in economic and perceptual decision-making in isolation. One reason for this separate treatment may be the assumption that there are distinct psychological mechanisms that underlie belief formation in economic and perceptual decisions. An alternative theory is that there exists a common mechanism that governs belief formation in both domains. Here,we test this alternative theory by combining a novel computational modeling technique with two well-known experimental paradigms. I estimate a drift-diffusion model (DDM) and provide an analytical method to decode prior beliefs from DDM parameters. Subjects in our experiment exhibit strong extrapolative beliefs in both paradigms. In line with the common mechanism hypothesis, we find that a single computational model explains belief formation in both tasks, and that individual differences in belief formation are correlated across tasks. These results suggest that extrapolative beliefs in economic decision-making may stem from low-level automatic processes that also play a role in perceptual decision-making, and therefore might be difficult to suppress.

The second chapter investigates the role of the sex steroid hormone testosterone as a biological mediator that translates environmental changes into shifts in cognition, that influence decision-making. Correlational studies have linked testosterone with aggression and disorders associated with poor impulse control, but corresponding mechanisms are poorly understood and there is no evidence of causality. Building on a dual-process framework, I identify a mechanism for testosterone's behavioral effects in humans: reducing cognitive reflection. In the largest testosterone administration study to date, 243 men received either testosterone or placebo and took the Cognitive Reflection Test (CRT) that estimated their capacity to override incorrect intuitive judgments with deliberate correct responses. Testosterone administration reduced CRT scores. The effect was robust to controlling for age, mood, math skills, treatment expectancy, and 14 other hormones. The effects were enhanced in subjects

with high cortisol and estradiol levels. These findings suggest a unified mechanism underlying testosterone's varied behavioral effects in humans and provide novel, clear, and testable predictions.

In the third chapter, I study dynamic unstructured bargaining with deadlines and one-sided private information about the amount available to share (the "pie size"). Using mechanism design theory, I show that given the players' incentives, the equilibrium incidence of bargaining failures ("strikes") should increase with the pie size, and I derive a condition under which strikes are efficient. In our setting, no equilibrium satisfies both equality and efficiency in all pie sizes. I derive two equilibria that resolve the trade-off between equality and efficiency by either favoring equality or favoring efficiency. Using a novel experimental paradigm, I confirm that strike incidence is decreasing in the pie size. Subjects reach equal splits in small pie games (in which strikes are efficient), while most payoffs are close to either the efficient or the equal equilibrium prediction when the pie is large. I employ a machine learning approach to show that bargaining process features recorded early in the game improve out of sample prediction of disagreements at the deadline. The process feature predictions are as accurate as predictions from pie sizes only, and adding process and pie data together improve predictions even more. As process data can be much richer than the series of cursor locations that we have used (for example, by including skin conductance, pupil dilation or facial expressions), better inference of outcome variables is likely feasible. Thus, if a policy maker or a mediator can access an independent measure of private information, an arbitration mechanism may allow boosting efficiency by taking this measurement into account.

PUBLISHED CONTENT AND CONTRIBUTIONS

Frydman, Cary and Gideon Nave (2016). "Extrapolative Beliefs in Perceptual and Economic Decisions: Evidence of a Common Mechanism". In: *Management Science, Forthcoming*.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

Chapter 1

INTRODUCTION

A multi-disciplinary effort has led to impressive progress in the field of decision neuroscience (Neuroeconomics) over the past decade (Glimcher and Fehr, 2013). It remains to be seen whether new discoveries in the field will translate into major contribution to the disciplines from which it has emerged. More specifically, there's a need to evaluate how promising findings about the neuroscience of decision-making can inform traditional questions in economics that historically have been investigated using choice data alone and without delineating the mechanism of choice.

An optimistic view describing the potential contribution of "opening the black box" of the human mind to standard economics is described by C. Camerer, Loewenstein, and Prelec, 2005; C. F. Camerer, Loewenstein, and Prelec, 2004 and C. Camerer, 2008. Some economists greet this proposition with skepticism and argue that economists should, in principle, ignore non-choice measurements because economic theories make no testable predictions about such data (Gul and Pesendorfer, 2008). Other economists philosophically accept that non-choice data should not be ignored in principle, but take a practical "wait and see" approach (Marchionni and Vromen, 2010; Rubinstein, 2008; Rubenstein, 2013). Bernheim, 2008, for example, noted that neural models of decision-making are also black boxes: "We are not dealing with a single box, but rather with a Russian doll. Do we truly believe that a good economist requires mastery of string theory?"

This chapter discusses three manners in which Neuroeconomics can contribute standard economics. The remaining three dissertation chapters are proofs of concepts, demonstrating how investigating the computational and biological basis of human decision-making can enrich traditional economic theories.

1.1 Understanding the origins of decision biases

Research in behavioral economics over the past few decades has shown that people's decisions often deviate from those of "homo-economicus", the selfish rational agent who is the hero of most economic theory textbooks (Gintis, 2000; Henrich et al., 2001; R. H. Thaler, 2000). These deviations ("decision biases") often lead to sub-optimal outcomes in the individual and the societal levels (C. F. Camerer, 2004;

C. Camerer, Babcock, et al., 1997; DellaVigna and Malmendier, 2006) and have become the target of various policy interventions (Leonard, 2008). Examples of such interventions are the introduction of mandatory retirement saving plans (Statman, 2013; Bateman and Piggott, 1998), shifting people towards more desirable choices by setting them as "defaults" (Johnson and Goldstein, 2003; Choi et al., 2004) and mandatory information disclosure policies (Caswell and Mojduszka, 1996; Welker, 1995). Finding the balance between limiting people's freedom of choice and protecting them from the consequences of poor decision-making is an important challenge for policy makers (C. Camerer, Issacharoff, et al., 2003; Loewenstein, Brennan, and Volpp, 2007; Hausman and Welch, 2010). It is therefore crucial to understand whether and how people are capable of overcoming decision-biases without having their freedom of choice limited.

Behavioral economists have characterized many systematic decision biases that unlikely reflect arbitrary mistakes, taking the initial step towards understanding these anomalies (Kahneman, 2003). But what is causing them? Arguably, contemporary humans face decision problems that are quite different from those that our ancestors had encountered (Rubin and Capra, 2011). Deciding whether to go hunting or foraging for grains is different from choosing between 30 types of barbecue source on the supermarket shelf; forecasting tomorrow's rainfall based on today's weather is not the same as predicting tomorrow's stock prices based on today's trades. As our brains have evolved in environments that do not resemble modern markets, we might rely on assumptions that are no longer valid when making economic decisions (McDermott, Fowler, and Smirnov, 2008; Chen, Lakshminarayanan, and Santos, 2005; Li et al., 2012).

In contrast to economic decision-making, humans seem to make reliable judgments and decisions in the perceptual domain. Although sensory illusions are pervasive in carefully controlled experiments under unnatural settings (Gregory, 1968), people are remarkably good at making sense of perceptual information as they navigate the world outside the laboratory. A recent documentation of a visual illusion in the field, a photo of a blue dress that seemed white to the majority of the population (Lafer-Sousa, Hermann, and Conway, 2015), was regarded with so much astonishment, that it became a world-wide internet sensation overnight. As our brains have evolved in an environment governed by the same regularities that operate today (i.e., mechanical, optical, and acoustic physical laws), we still benefit from relying on the same computations that our ancestors' brains had used when making

decisions that translate sensory information into perceptual judgments and motor actions.

Many decision biases in the economic domain have parallel phenomena in the perceptual domain. Our sensitivity to light intensity and auditory loudness follows logarithmic laws (Reichl, Tuffin, and Schatz, 2013; Drago et al., 2003; Palmer, 1999) that resemble the manner in which we encode monetary rewards. We perceive the luminance and size of objects in relation to their surroundings (Palmer, 1999), in a manner that resembles framing effects in economic decision-making (Levin, Schneider, and Gaeth, 1998). Even the compromise and the attraction effects, two well-documented phenomena in consumer decision-making (Simonson, 1989), were recently documented in the perceptual domain (Trueblood et al., 2013). These findings suggest that decision biases might arise because our brains apply computational techniques that successfully solve perceptual problems when making economic decisions.

Chapter 2 of this dissertation is concerned with a specific decision bias in economic decision-making, extrapolative beliefs, also known as the belief in the "hot hand" (Bloomfield and Hales, 2002; Rabin, 2000). People often rely on past observations when forecasting the future, even when they contain no credible information. This tendency is thought to underlie market-level phenomena such as over-reaction to news (Bondt and R. Thaler, 1985). Extrapolative belief formation is also an empirical regularity in lab experiments of perceptual decision-making (Cho et al., 2002; Huettel, Mack, and McCarthy, 2002): people respond faster and more accurately to sensory stimuli that continue an apparent pattern, even when explicitly told that the sequence is completely randomized. We investigate whether people use a common computational mechanism of belief formation when making perceptual and economic decisions in a within-subject design, where each participants took part in decision-making tasks from both domains.

People have no conscious representation of their perceptual beliefs, as perceptual judgments occur fast and automatically (Nissen and Bullemer, 1987; Curran and Keele, 1993). Researchers typically use response-times and accuracy rates as proxies of perceptual beliefs, but contrasting these measures with economic beliefs is comparing apples and oranges, as they are not calibrated to a probabilistic scale. Further, response times and accuracy rates are sensitive to many factors other than beliefs, such as the quality of sensory input, the time it takes to make a motor movement, and the speed-accuracy trade-off.

Chapter 2 provides the computational framework for decoding perceptual beliefs from response times and accuracy rates. By applying the drift diffusion model, a widely used computational technique from the literature on perceptual decision-making (Ratcliff and McKoon, 2008), we show theoretically that beliefs are encoded in the ratio between two parameters of the model, the initial point and boundary. This allows us to walk the extra mile and compare the belief formation process across the perceptual and economic domains in a within-subject design. We find a reliable correlation between the degree of extrapolative beliefs across the perceptual and economic decision task. Furthermore, calibrating a single parameter computational model based on the perceptual decision-making task further allows out of sample predicting the belief formation process in the economic task. These results show that extrapolative beliefs in economic decision-making may stem from low-level automatic processes that also play a role in perceptual decision-making, and therefore might be difficult to suppress. Therefore, non-paternalistic policies, such as mandating information disclosure (Brav and Heaton, 2002), might not be as effective as one might hope for reducing welfare losses associate with this decision bias.

1.2 Biological state influences decision-making

Traditional economic theories assume that humans make decisions based on complete preferences that are stable across time. This presumption makes theories parsimonious and tractable, but sits uneasily with much empirical evidence. For example, experimentally induced incidental emotions were shown to carry over to unrelated tasks and influence subsequent consumer decisions systematically (Lerner, Small, and Loewenstein, 2004; Loewenstein and Lerner, 2003). Acute stress induced by dipping hands in icy water increased framing effects in an unrelated financial decision-making task that followed it (Porcelli and Delgado, 2009). Few would disagree with the premise that economically relevant constructs such as productivity, sociability, or confidence are influenced by intake (or avoidance) of substances like caffeine, alcohol, antidepressants, and other therapeutic or recreational drugs.

Hormones are chemical signals that are released into the blood stream and in the brain in response to internal states and environmental cues. The degree of hormonal responses to an environmental change might be moderated by various factors: genetic, demographics, personality, and more (see Figure 1.1). Many brain regions involved in social behavior and decision-making contain hormonal receptors. Hormones affect information processing in these brain regions in long lasting way, from

seconds to hours, making them immediate candidate biological mediators for translating environmental changes into shifts in cognition, motivation, and emotion that influence decision-making.

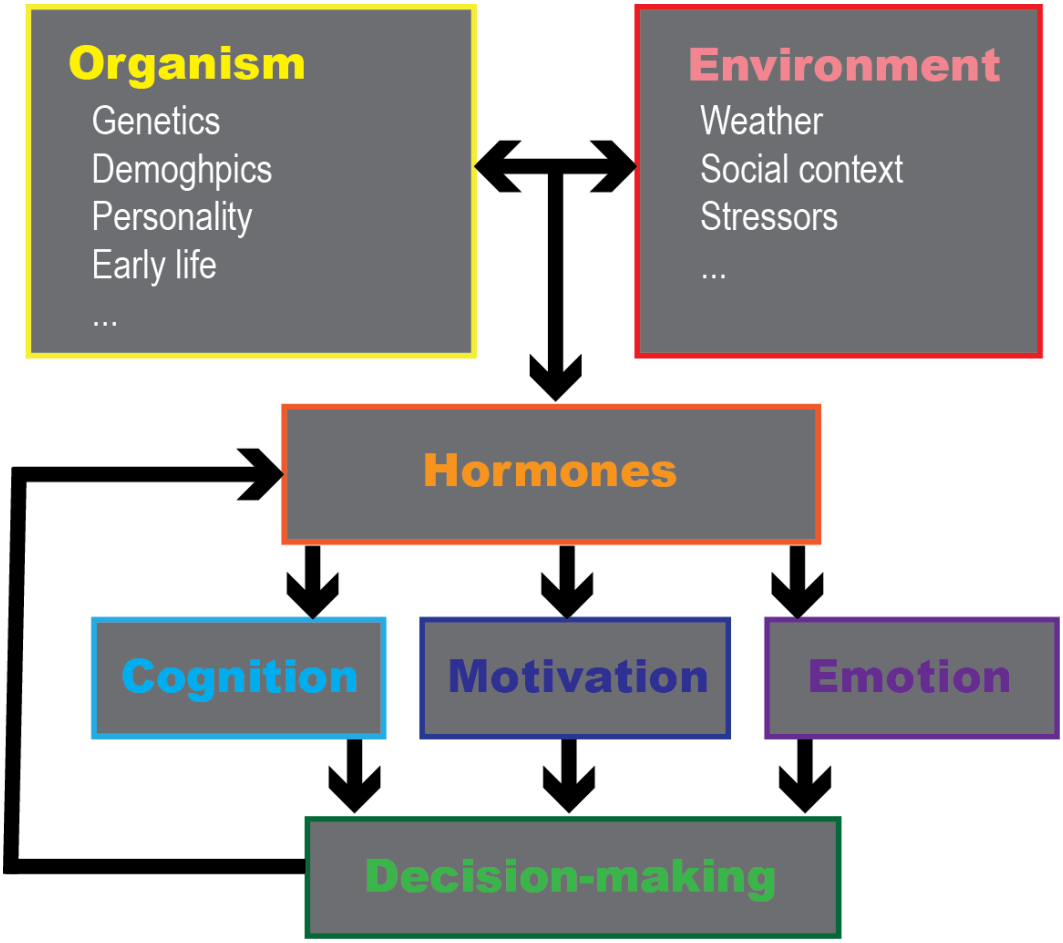


Figure 1.1: A framework for studying hormonal influences on decision-making. The interaction of environmental factors and the organism's inherent characteristics (e.g., genes) leads to hormonal variations that cause shifts in cognition, motivation and emotions and influence decision-making. Decisions might generate behaviors that influence the organism's hormonal levels (either directly or through environmental changes), as implied by the feedback arrow.

Traditionally, hormones had been "unobservables" to the economist, but this reality is changing. The development of relatively cheap and easy to perform hormonal assay methods and techniques for delivering pharmacological treatments (C. Wang et al., 2004; Bos et al., 2012) has made it possible to measure and manipulate hormonal levels in the lab and observe how they are linked to behaviors. In the near future, portable hormonal assay kits that connect to mobile devices will allow conducting fast and cheap hormonal measurements in the field as well (Ehrenkranz,

2013).

Chapter 3 uses a dual process framework (Evans, 2003) to study how exogenous administration of testosterone, the male sex steroid hormone influences the decision-making process in men. We show that testosterone impairs males' performance in the cognitive reflection tes , abbreviated CRT (Frederick, 2005). The CRT estimates one's capacity to override incorrect intuitive judgments that effortlessly "jump" into one's mind, in favor of accurate answers that require deliberate, yet easy to perform arithmetic calculations. Despite being only 3-items long, the CRT is a predictor of various behavioral outcomes, from the preference of an immediate gratification over a larger future reward, to reliance on sub-optimal on judgmental heuristics (e.g., the conjunction fallacy) and formation of asset market bubbles (Toplak, West, and Stanovich, 2011; Bosch-Rosa, Meissner, and Bosch i Domènech, 2015).

Various environmental factors influence testosterone levels in men. The presence of an attractive female (Ronay and Hippel, 2010), a win of one's favorite soccer team (Bernhardt et al., 1998), graduation (Allen Mazur and Lamb, 1980), and divorce (Booth and Dabbs, 1993; Allan Mazur and Michalek, 1998; Gettler et al., 2011) increase testosterone. Becoming a father (Gettler et al., 2011), losing a competition (Booth, Shelley, et al., 1989; Allan Mazur, Booth, and Dabbs Jr, 1992), and a decrease in female to male ratio (Miller, Maner, and McNulty, 2012) decrease testosterone. Testosterone could be a hidden variable that translates transitions in these environmental factors and many others into behavioral changes.

Understanding how testosterone causally influences decision-making may allow generating behavioral predictions under novel environmental conditions. For example, the one child policy in China - a nation of billion people - caused an increase of the sex ratio between male and female births (Cameron et al., 2013). Thirty million more men than women will live in China by 2020, potentially leading to social instability and courtship-motivated emigration (Ding and Hesketh, 2006; Zhu, Lu, and Hesketh, 2009). Studying the effects of testosteroneon decision-making could bring about non-obvious predictions on how unprecedented circumstances would influence the local and global economy.

1.3 Imputation of private information

Asymmetrical access to information is a feature of many economic environments and a major cause of inefficiencies. A famous example is the "market for lemons" problem (Akerlof, 1995). Consider a second hand vehicles market, where there are

two types of goods: damaged ("lemons") and well maintained cars ("cherries"). Sellers know how reliable their cars are, but cannot prove it to the buyers. Buyers are willing to pay more for cherries. In such settings, sellers have strong financial incentives to convince the buyers that they are selling "cherries", regardless of the true state of the car. Theoretical and experimental investigations show that buyers might be especially reluctant to pay the price premium for purchasing a cherry in such markets. The result is an inefficient equilibrium, where only lemons are sold for low prices (Lynch et al., 1986).

Non-choice measures can allow meaningful inferences of private information (Davatzikos et al., 2005; Farwell and Donchin, 1991; Meijer et al., 2007). Neuroeconomic methods are often criticized for being unnatural, costly, and inconvenient. This criticism might hold for methods such as functional MRI that require having subjects motionlessly lie inside a thunderous magnet. But fMRI is only one of many possible sources of non-choice data. Involuntary (autonomic) biological responses such as response times, heart rate, pupil dilation, changes in electro dermal activity, and facial expressions can be measured rapidly and with low marginal costs. These measures are linked with mental states (e.g., anxiety, cognitive difficulty or arousal) that may correlate with private information (J. T.-y. Wang, Spezio, and C. F. Camerer, 2010).

In the final chapter, I investigate unstructured dynamic bargaining with private information. Experimental economists have neglected this important topic for many years, perhaps because a lack of adequate behavioral paradigms and analytical techniques for exploiting the richness of unstructured bargaining data. In the game, only one party knows the surplus available to both players ("pie size"). Participants bargain on the payoff of the uninformed party and communicate offers using mouse clicks whenever they please. The game abstracts wage negotiations, where employees do not know the exact monetary value of their work to the firm. Under informational asymmetry, willingness to endure a strike might be the only credible means of the employers to convey their incapacity to pay a high wage (Kennan and Wilson, 1993). As a result, disagreements arise even when both parties act rationally given to their information and beliefs.

Using a game-theoretic framework, I show theoretically and empirically that disagreements arise in such settings, and the realization of the pie size is a strong predictor of their occurrences. Further, many deals are made in the last seconds of bargaining, a phenomenon previously termed as the "deadline effect". These results

highlight a great challenge to econometricians in the real world: the occurrence of a strike in any given negotiation is difficult to predict, as it depends on unobservable private information.

But what if we could observe correlates of the private information? As a proof of concept, I used a machine-learning algorithm to test whether the temporal dynamics of players' bargianing positions can predict disagreements before the deadline has arrived. Indeed, a rich set of behavioral features ("process data") is highly informative about bargaining outcomes. By using process data, one can predict disagreements just as accurately as when having access to the pie size realization. Combining the pie size information with process features improves predictions even further. As process data can be much richer than the series of cursor locations that we have used (e.g., by including skin conductance, pupil dilation, or facial expressions), better inference of outcome variables is likely feasible.

From a practical perspective, the premise of using process data to predict disagreements and reveal private information has the potential to reduce inefficiencies. A proof of concept was demonstrated by Krajbich et al., 2009, in a study that addressed the mechanism design problem of "free riders" while allocating costs of public goods among group members. When the true values of each member for the public good is private information, economic theory predicts that it is not feasible to achieve an allocation of costs in which each individual's benefit is greater than his costs. Using machine-learning techniques, the researchers showed theoretically and experimentally that a mechanism for allocating costs as a function of both announcements and neural proxies for private information could overcome the "free riders" barrier. Thus, if a policy maker or a mediator can access an independent measure of private information, an arbitration mechanism can boost efficiency by taking this measurement into account.

References

Akerlof, George (1995). *The market for "lemons": Quality uncertainty and the market mechanism*. Springer.

Bateman, Hazel, John Piggott, et al. (1998). "Mandatory retirement saving in Australia". In: *Annals of Public and Cooperative Economics* 69.4, pp. 547–569.

Bernhardt, Paul C et al. (1998). "Testosterone changes during vicarious experiences of winning and losing among fans at sporting events". In: *Physiology & Behavior* 65.1, pp. 59–62.

Bernheim, B Douglas (2008). *On the potential of neuroeconomics: A critical (but hopeful) appraisal*. Tech. rep. National Bureau of Economic Research.

Bloomfield, Robert and Jeffrey Hales (2002). "Predicting the next step of a random walk: experimental evidence of regime-shifting beliefs". In: *Journal of financial Economics* 65.3, pp. 397–414.

Bondt, Werner FM and Richard Thaler (1985). "Does the stock market overreact?" In: *The Journal of finance* 40.3, pp. 793–805.

Booth, Alan and James M Dabbs (1993). "Testosterone and men's marriages". In: *Social Forces* 72.2, pp. 463–477.

Booth, Alan, Greg Shelley, et al. (1989). "Testosterone, and winning and losing in human competition". In: *Hormones and behavior* 23.4, pp. 556–571.

Bos, Peter A et al. (2012). "Acute effects of steroid hormones and neuropeptides on human social–emotional behavior: a review of single administration studies". In: *Frontiers in neuroendocrinology* 33.1, pp. 17–35.

Bosch-Rosa, Ciril, Thomas Meissner, and Antoni Bosch i Domènech (2015). "Cognitive bubbles". In: *Available at SSRN 2553230*.

Brav, Alon and John B Heaton (2002). "Competing theories of financial anomalies". In: *Review of Financial Studies* 15.2, pp. 575–606.

Camerer, Colin (2008). "The case for mindful economics". In: *Foundations of positive and normative economics*, pp. 43–69.

Camerer, Colin F (2004). "Prospect theory in the wild: Evidence from the field". In: *Advances in behavioral economics*, pp. 148–161.

Camerer, Colin F, George Loewenstein, and Drazen Prelec (2004). "Neuroeconomics: Why economics needs brains". In: *The Scandinavian Journal of Economics* 106.3, pp. 555–579.

Camerer, Colin, Linda Babcock, et al. (1997). "Labor supply of New York City cabdrivers: One day at a time". In: *The Quarterly Journal of Economics*, pp. 407–441.

Camerer, Colin, Samuel Issacharoff, et al. (2003). "Regulation for Conservatives: Behavioral Economics and the Case for" Asymmetric Paternalism"". In: *University of Pennsylvania law review* 151.3, pp. 1211–1254.

Camerer, Colin, George Loewenstein, and Drazen Prelec (2005). "Neuroeconomics: How neuroscience can inform economics". In: *Journal of economic Literature*, pp. 9–64.

Cameron, Lisa et al. (2013). "Little emperors: behavioral impacts of China's One-Child Policy". In: *Science* 339.6122, pp. 953–957.

Caswell, Julie A and Eliza M Mojduszka (1996). "Using informational labeling to influence the market for quality in food products". In: *American Journal of Agricultural Economics* 78.5, pp. 1248–1253.

Chen, M Keith, Venkat Lakshminarayanan, and Laurie Santos (2005). "The evolution of our preferences: Evidence from capuchin monkey trading behavior". In:

Cho, Raymond Y et al. (2002). "Mechanisms underlying dependencies of performance on stimulus history in a two-alternative forced-choice task". In: *Cognitive, Affective, & Behavioral Neuroscience* 2.4, pp. 283–299.

Choi, James J et al. (2004). "For better or for worse: Default effects and 401 (k) savings behavior". In: *Perspectives on the Economics of Aging*. University of Chicago Press, pp. 81–126.

Curran, Tim and Steven W Keele (1993). "Attentional and nonattentional forms of sequence learning." In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 19.1, p. 189.

Davatzikos, Christos et al. (2005). "Classifying spatial patterns of brain activity with machine learning methods: application to lie detection". In: *Neuroimage* 28.3, pp. 663–668.

DellaVigna, Stefano and Ulrike Malmendier (2006). "Paying not to go to the gym". In: *The American Economic Review*, pp. 694–719.

Ding, Qu Jian and Therese Hesketh (2006). "Family size, fertility preferences, and sex ratio in China in the era of the one child family policy: results from national family planning and reproductive health survey". In: *BMJ* 333.7564, pp. 371–373.

Drago, Frederic et al. (2003). "Adaptive logarithmic mapping for displaying high contrast scenes". In: *Computer Graphics Forum*. Vol. 22. 3. Wiley Online Library, pp. 419–426.

Ehrenkranz, Joel RL (2013). *Device for performing a blood, cell, and/or pathogen count and methods for use thereof*. US Patent App. 13/862,188.

Evans, Jonathan St BT (2003). "In two minds: dual-process accounts of reasoning". In: *Trends in cognitive sciences* 7.10, pp. 454–459.

Farwell, Lawrence A and Emanuel Donchin (1991). "The Truth Will Out: Interrogative Polygraphy ("Lie Detection") With Event-Related Brain Potentials". In: *Psychophysiology* 28.5, pp. 531–547.

Frederick, Shane (2005). "Cognitive reflection and decision making". In: *The Journal of Economic Perspectives* 19.4, pp. 25–42.

Gettler, Lee T et al. (2011). "Longitudinal evidence that fatherhood decreases testosterone in human males". In: *Proceedings of the National Academy of Sciences* 108.39, pp. 16194–16199.

Gintis, Herbert (2000). "Beyond Homo economicus: evidence from experimental economics". In: *Ecological economics* 35.3, pp. 311–322.

Glimcher, Paul W and Ernst Fehr (2013). *Neuroeconomics: Decision making and the brain*. Academic Press.

Gregory, Richard L (1968). "Visual illusions." In: *Scientific American*.

Gul, Faruk, Wolfgang Pesendorfer, et al. (2008). "The case for mindless economics". In: *The foundations of positive and normative economics* 3, p. 39.

Hausman, Daniel M and Brynn Welch (2010). "Debate: To Nudge or Not to Nudge*". In: *Journal of Political Philosophy* 18.1, pp. 123–136.

Henrich, Joseph et al. (2001). "In search of homo economicus: behavioral experiments in 15 small-scale societies". In: *The American Economic Review* 91.2, pp. 73–78.

Huettel, Scott A, Peter B Mack, and Gregory McCarthy (2002). "Perceiving patterns in random series: dynamic processing of sequence in prefrontal cortex". In: *Nature neuroscience* 5.5, pp. 485–490.

Johnson, Eric J and Daniel G Goldstein (2003). "Do defaults save lives?" In: *Science* 302, pp. 1338–1339.

Kahneman, Daniel (2003). "Maps of bounded rationality: Psychology for behavioral economics". In: *The American economic review* 93.5, pp. 1449–1475.

Kennan, John and Robert B Wilson (1993). "Bargaining with private information". In: *Journal of Economic Literature* 31, pp. 45–45.

Krajbich, Ian et al. (2009). "Using neural measures of economic value to solve the public goods free-rider problem". In: *Science* 326.5952, pp. 596–599.

Lafer-Sousa, Rosa, Katherine L Hermann, and Bevil R Conway (2015). "Striking individual differences in color perception uncovered by 'the dress' photograph". In: *Current Biology* 25.13, R545–R546.

Leonard, Thomas C (2008). "Richard H. Thaler, Cass R. Sunstein, Nudge: Improving decisions about health, wealth, and happiness". In: *Constitutional Political Economy* 19.4, pp. 356–360.

Lerner, Jennifer S, Deborah A Small, and George Loewenstein (2004). "Heart strings and purse strings carryover effects of emotions on economic decisions". In: *Psychological Science* 15.5, pp. 337–341.

Levin, Irwin P, Sandra L Schneider, and Gary J Gaeth (1998). "All frames are not created equal: A typology and critical analysis of framing effects". In: *Organizational behavior and human decision processes* 76.2, pp. 149–188.

Li, Yexin Jessica et al. (2012). "Economic decision biases and fundamental motivations: how mating and self-protection alter loss aversion." In: *Journal of personality and social psychology* 102.3, p. 550.

Loewenstein, George, Troyen Brennan, and Kevin G Volpp (2007). "Asymmetric paternalism to improve health behaviors". In: *Jama* 298.20, pp. 2415–2417.

Loewenstein, George and Jennifer S Lerner (2003). "The role of affect in decision making". In: *Handbook of affective science* 619.642, p. 3.

Lynch, Michael et al. (1986). “Product quality, consumer information and" lemons" in experimental markets”. In:

Marchionni, Caterina and Jack Vromen (2010). “Neuroeconomics: hype or hope?” In:

Mazur, Allan, Alan Booth, and James M Dabbs Jr (1992). “Testosterone and chess competition”. In: *Social Psychology Quarterly*, pp. 70–77.

Mazur, Allan and Joel Michalek (1998). “Marriage, divorce, and male testosterone”. In: *Social Forces* 77.1, pp. 315–330.

Mazur, Allen and Theodore A Lamb (1980). “Testosterone, status, and mood in human males”. In: *Hormones and Behavior* 14.3, pp. 236–246.

McDermott, Rose, James H Fowler, and Oleg Smirnov (2008). “On the evolutionary origin of prospect theory preferences”. In: *The Journal of Politics* 70.02, pp. 335–350.

Meijer, Ewout H et al. (2007). “Combining skin conductance and forced choice in the detection of concealed information”. In: *Psychophysiology* 44.5, pp. 814–822.

Miller, Saul L, Jon K Maner, and James K McNulty (2012). “Adaptive attunement to the sex of individuals at a competition: the ratio of opposite-to same-sex individuals correlates with changes in competitors’ testosterone levels”. In: *Evolution and Human Behavior* 33.1, pp. 57–63.

Nissen, Mary Jo and Peter Bullemer (1987). “Attentional requirements of learning: Evidence from performance measures”. In: *Cognitive psychology* 19.1, pp. 1–32.

Palmer, Stephen E (1999). *Vision science: Photons to phenomenology*. Vol. 1. MIT press Cambridge, MA.

Porcelli, Anthony J and Mauricio R Delgado (2009). “Acute stress modulates risk taking in financial decision making”. In: *Psychological Science* 20.3, pp. 278–283.

Rabin, Matthew et al. (2000). *Inference by believers in the law of small numbers*. Institute of Business and Economic Research.

Ratcliff, Roger and Gail McKoon (2008). “The diffusion decision model: theory and data for two-choice decision tasks”. In: *Neural computation* 20.4, pp. 873–922.

Reichl, Peter, Bruno Tuffin, and Raimund Schatz (2013). “Logarithmic laws in service quality perception: where microeconomics meets psychophysics and quality of experience”. In: *Telecommunication Systems* 52.2, pp. 587–600.

Ronay, Richard and William von Hippel (2010). “The presence of an attractive woman elevates testosterone and physical risk taking in young men”. In: *Social Psychological and Personality Science* 1.1, pp. 57–64.

Rubenstein, Ariel (2013). “Response time and decision making: An experimental study”. In: *Judgment and Decision Making* 8.5, p. 540.

Rubin, Paul H and C Monica Capra (2011). "The evolutionary psychology of economics". In: *Applied Evolutionary Psychology*, pp. 7–15.

Rubinstein, Ariel (2008). "Comments on neuroeconomics". In: *Economics and Philosophy* 24.03, pp. 485–494.

Simonson, Itamar (1989). "Choice based on reasons: The case of attraction and compromise effects". In: *Journal of consumer research*, pp. 158–174.

Statman, Meir (2013). "Mandatory retirement savings". In: *Financial Analysts Journal* 69.3, pp. 14–18.

Thaler, Richard H (2000). "From homo economicus to homo sapiens". In: *The Journal of Economic Perspectives* 14.1, pp. 133–141.

Toplak, Maggie E, Richard F West, and Keith E Stanovich (2011). "The Cognitive Reflection Test as a predictor of performance on heuristics-and-biases tasks". In: *Memory & Cognition* 39.7, pp. 1275–1289.

Trueblood, Jennifer S et al. (2013). "Not just for consumers context effects are fundamental to decision making". In: *Psychological science* 24.6, pp. 901–908.

Wang, Christina et al. (2004). "Measurement of total serum testosterone in adult men: comparison of current laboratory methods versus liquid chromatography-tandem mass spectrometry". In: *The Journal of Clinical Endocrinology & Metabolism* 89.2, pp. 534–543.

Wang, Joseph Tao-yi, Michael Spezio, and Colin F Camerer (2010). "Pinocchio's pupil: using eyetracking and pupil dilation to understand truth telling and deception in sender-receiver games". In: *The American Economic Review* 100.3, pp. 984–1007.

Welker, Michael (1995). "Disclosure Policy, Information Asymmetry, and Liquidity in Equity Markets*". In: *Contemporary accounting research* 11.2, pp. 801–827.

Zhu, Wei Xing, Li Lu, and Therese Hesketh (2009). "China's excess males, sex selective abortion, and one child policy: analysis of data from 2005 national intercensus survey". In: *BMJ* 338, b1211.

Chapter 2

EXTRAPOLATIVE BELIEFS IN PERCEPTUAL AND ECONOMIC DECISIONS: EVIDENCE OF A COMMON MECHANISM

ABSTRACT

A critical component of both economic and perceptual decision-making under uncertainty is the belief formation process. However, most research has studied belief formation in economic and perceptual decision-making in isolation. One reason for this separate treatment may be the assumption that there are distinct psychological mechanisms that underlie belief formation in economic and perceptual decisions. An alternative theory is that there exists a common mechanism that governs belief formation in both domains. Here, we test this alternative theory by combining a novel computational modeling technique with two well-known experimental paradigms. We estimate a drift-diffusion model (DDM) and provide an analytical method to decode prior beliefs from DDM parameters. Subjects in our experiment exhibit strong extrapolative beliefs in both paradigms. In line with the common mechanism hypothesis, we find that a single computational model explains belief formation in both tasks, and that individual differences in belief formation are correlated across tasks.

2.1 Introduction

A common goal of economics and psychology is to understand the mechanisms that govern decision-making in uncertain environments (Platt and Huettel, 2008). This is an ambitious goal that seeks to explain the computational processes underlying both (i) perceptual decisions – those that are determined by judgments about objective states of the world, and (ii) value-based (economic) decisions – those that are determined by subjective beliefs and preferences. Over the past several decades a vast amount of research has been devoted to studying these two types of decision-making. Most of this research has proceeded along two parallel tracks, perhaps because of an implicit assumption that the psychological mechanisms governing perceptual and economic decisions are distinct.

However, a small but growing body of literature has begun to investigate the links between economic and perceptual decision-making (Summerfield and Tsetsos, 2012; Summerfield and Tsetsos, 2015). For example, the evidence accumulation process that is used to describe perceptual decision-making can also explain the dynamics in simple economic decisions (Krajbich and Rangel, 2011; Tsetsos, Chater, and Usher, 2012; Polania et al., 2014). Furthermore, robust decision biases from economic decision-making, such as context effects, have recently been uncovered in perceptual decision-making (Trueblood et al., 2013). There is also evidence of an interaction between the two domains, as visual saliency can systematically bias economic decisions (Mormann et al., 2012; Towal, Mormann, and Koch, 2013). However, despite this growing evidence, it is unknown whether there is a similar link between dynamic (across-trial) aspects of perceptual and economic decision-making. In other words, does a common belief formation mechanism across these two domains exist?

The answer to this question is important because it can shed light on the origins and mechanisms of belief-based biases in judgment and decision-making. Moreover, these biases are key ingredients in many behavioral models in finance (Barberis, Shleifer, and Vishny, 1998; Hong and Stein, 1999; Barberis, Greenwood, et al., 2015), and thus understanding the micro-foundation of such biases may be useful in building more psychologically realistic models of financial markets. For example, recent survey evidence shows that many investors hold extrapolative beliefs, where they expect stock prices to continue rising after they have previously risen, and to fall after they have previously fallen (Greenwood and Shleifer, 2014). If these extrapolative beliefs are driven by the same mechanism that governs belief formation in lower level perceptual processes, characterizing this common mechanism can be

useful in modeling higher-level economic expectation formation.

In this paper, we combine two well-known experimental paradigms with computational modeling to investigate whether belief formation is governed by the same psychological mechanism in economic and perceptual decision-making. Before investigating the relationship between belief formation in each domain, it is first necessary to precisely measure beliefs. Fortunately, economists have long since provided incentive-compatible experimental methods for measuring beliefs in economic decisions (Brier, 1950; Becker, DeGroot, and Marschak, 1964; Selten, 1998; Karni, 2009). Less work has been devoted to developing incentive compatible methods for eliciting beliefs in perceptual decision-making, but we do so here by building on previous work that uses computational modeling to decode beliefs from choice and response-time (RT) data. Specifically, our technique builds on the large literature of sequential sampling models (SSMs) in perceptual decision-making (Townsend and Ashby, 1985; Usher and McClelland, 2001; Ratcliff and Smith, 2004; Bogacz et al., 2006; A. R. Teodorescu and Usher, 2013) and, more recently, in value based decision-making (Fehr and Rangel, 2011; Webb, 2013; Woodford, 2014).

Most studies that use SSMs manipulate an attribute of the environment (e.g., stimulus motion coherence or subjective value) and test which computational parameters encode the change in environment (Klauer et al., 2007; Mulder et al., 2012; White and Poldrack, 2014). In contrast, the computational technique we employ to measure beliefs focuses on decoding perceived changes in the environment (i.e., changes in subjective beliefs) from an SSM. In particular, we show theoretically how the estimated initial point and boundary parameters of a drift diffusion model can be used to infer a subject's prior belief. Our study therefore provides a novel example of how neuro-computational models can be used to measure a subject's belief formation process.

To demonstrate this, we recruited subjects to participate in two separate tasks: an economic decision-making task (EDT) and a perceptual decision-making task (PDT). While each task has been used several times in its own literature separately (Bloomfield and Hales, 2002; Cho et al., 2002; Huettel, Mack, and McCarthy, 2002; Asparouhova, Hertzel, and Lemmon, 2009), in the current study we have subjects participate in both the EDT and PDT in the same experimental setting. This within-subject design allowed us to measure and compare the computations governing belief formation across different decision domains. We hypothesized that if a single psychological mechanism governs belief formation across economic

and perceptual decision domains, then (i) a common computational model should explain belief-updating (across trials) in both tasks and (ii) individual differences in the degree to which subjects rely on recent stimulus history to update beliefs should be correlated across tasks.

2.2 Methods

Subjects

Thirty-eight subjects (17 females) aged 17-29 (mean: 20.24 SD: 3.11) participated in the study. Subjects were students at Caltech or at a nearby community college and the sample size was chosen to match the exact sample size employed in previous work with the same task (Bloomfield and Hales, 2002). The California Institute of Technology and University of Southern California Institutional Review Boards approved this study, and the subjects gave informed consent.

perceptual decision task

The PDT consisted of four blocks of 300 trials each, preceded by 5 training trials. Each trial began with the appearance of a white fixation cross in the center of a black screen; after 800 milliseconds, either a white circle (diameter: 10.5 cm) or a blue square (width: 10.5 cm) appeared in the place of the cross. Subjects were instructed to respond by either pressing the "right arrow" key when a circle appeared or the "left arrow" key when a square appeared (Figure 2.1a). Subjects were told that they would receive one cent for each correct response, and their earnings would be reduced by 0.05 cents for every 100 milliseconds of delay in their response. If the response was slower than 2 seconds they would receive 0 cents. A new trial started immediately following a response, with the appearance of a new fixation cross on the screen. Subjects were told that there were only two possible stimuli (a circle or a square), that the probability of seeing either shape was 0.5, and that the stimuli of previous trials had no influence on future trials. The actual sequence of stimuli was an independent and identically distributed pseudo-random binary process, such that the stimuli sequence was identical for all subjects. Subjects had a break of 20 seconds between blocks, during which they received feedback about the number of correct responses they made, but did not receive feedback about their mean RT. The task instructions are available in the appendix.

Economic decision task

In the EDT, subjects were told that data from publicly traded firms was used to create a model that generated sequences of "performance surprises" – a time series of actual performance minus predicted performance (Bloomfield and Hales, 2002; Asparouhova, Hertzel, and Lemmon, 2009). Subjects were presented with a sequence of performance surprises from a typical firm, and were instructed to predict whether the next performance surprise would be positive or negative. In each period, subjects saw a history chart of performance surprises from the last 14 periods plotted in yellow on a black screen (see Figure 2.1b). Each period, subjects were endowed with 100 units of experimental cash and were asked to state the maximum price at which they would be willing to buy a share of stock in the company. To elicit an incentive-compatible measure of a subject's willingness to pay (WTP) for the stock, we used a Becker–DeGroot–Marshak (BDM) auction (Becker, DeGroot, and Marschak, 1964). Thus, after the subject stated his WTP, the actual price was drawn from a uniform distribution between 0 and 100, and the stock would be purchased at the price drawn if and only if its price was less than the WTP. If the subject purchased the stock, it would return 100 units of experimental cash in the case of a positive surprise and 0 units of experimental cash otherwise. The WTP measures the expected value of the stock, and therefore, for a risk neutral subject, it provides a measure of the subjective belief of a positive performance surprise. Subjects were explicitly told that in order to make the most money in the task, they should set the price equal to the probability of seeing a positive surprise. After stating their WTP, subjects received feedback about the price drawn, the stock performance, and their financial outcome. The task consisted of a single block of 400 trials preceded by 10 training trials, where each period started at the same point in the sequence where the previous trial had ended. The actual sequence was a pseudo-random independent and identically distributed binary process (the stimuli were identical for all subjects). After every 50 trials, subjects saw their accumulated payoff and were allowed to take a short break. At the end of the experiment, subjects were paid for all of their trials, using a conversion rate of 5,000 units of experimental cash = 1 USD. The task instructions are available in the appendix.

Experimental Procedures

The data was collected over three experimental sessions, all conducted at the Caltech Social Science Experimental Laboratory (SSEL). At the beginning of each session, participants were randomly allocated to cubicles in the lab, where they could not see

or interact with each other. Before each task, subjects received printed instructions that were also read out loud by the experimenter, and subjects subsequently had an opportunity to ask questions. Each session started with the PDT, followed by the EDT (this design choice is discussed in the appendix). Both tasks were programmed using Matlab Psychtoolbox (Brainard, 1997).

2.3 Results

Basic results from the perceptual decision-making task

We found that RTs and error rates systematically varied as a function of recent stimulus history, despite the fact that subjects were explicitly told the probability of displaying a circle was 0.5 on all trials, in accordance with previous studies (Cho et al., 2002; Huettel, Mack, and McCarthy, 2002). On trials where the stimulus continued the recent streak ("continuation" trials), RTs decreased with streak length ($p < 0.001$, see Table 2.3). In contrast, Figure 2.2a shows that on trials where the stimulus violated the recent streak (e.g. 3 circles followed by a square), RTs moderately increase with streak length ($p < 0.001$, see Table 2.3). Figure 2.2b displays error rates as a function of streak length, where an error is defined as misclassifying the stimulus. We found that as the streak length increases, error rates decrease for continuation trials, but they increase for violation trials (both $p < 0.001$, see Table 2.4).

The perceptual decision-making literature has highlighted two main types of sequential effects on response times and error rates. First, automatic facilitation (AF) effects occur when response times following a certain streak length are faster regardless of whether the current trial extends the streak or not. For example, after the sequence, {square, square, square}, an AF effect predicts that the response on the subsequent trial to either a circle or square will be faster than after the sequence {circle, square, square}. This type of effect could be driven by post-response residual activity (Roitman and Shadlen, 2002) depending on the decay rate of neuronal activity in the motor cortex. In contrast, strategic expectancy (SE) effects occur when response times following a certain streak length are faster only when the current trial extends the streak. This implies that after the sequence, {square, square, square}, an SE effect predicts that the response on a square trial will be faster than on a circle trial. The results in Figure 2.2a are inconsistent with AF effects because the mean RT for each streak length depends on the stimulus identity of the current trial. However, they are consistent with SE effects, which can be interpreted as effects driven by expectations about future stimuli. This finding is consistent with an extensive body

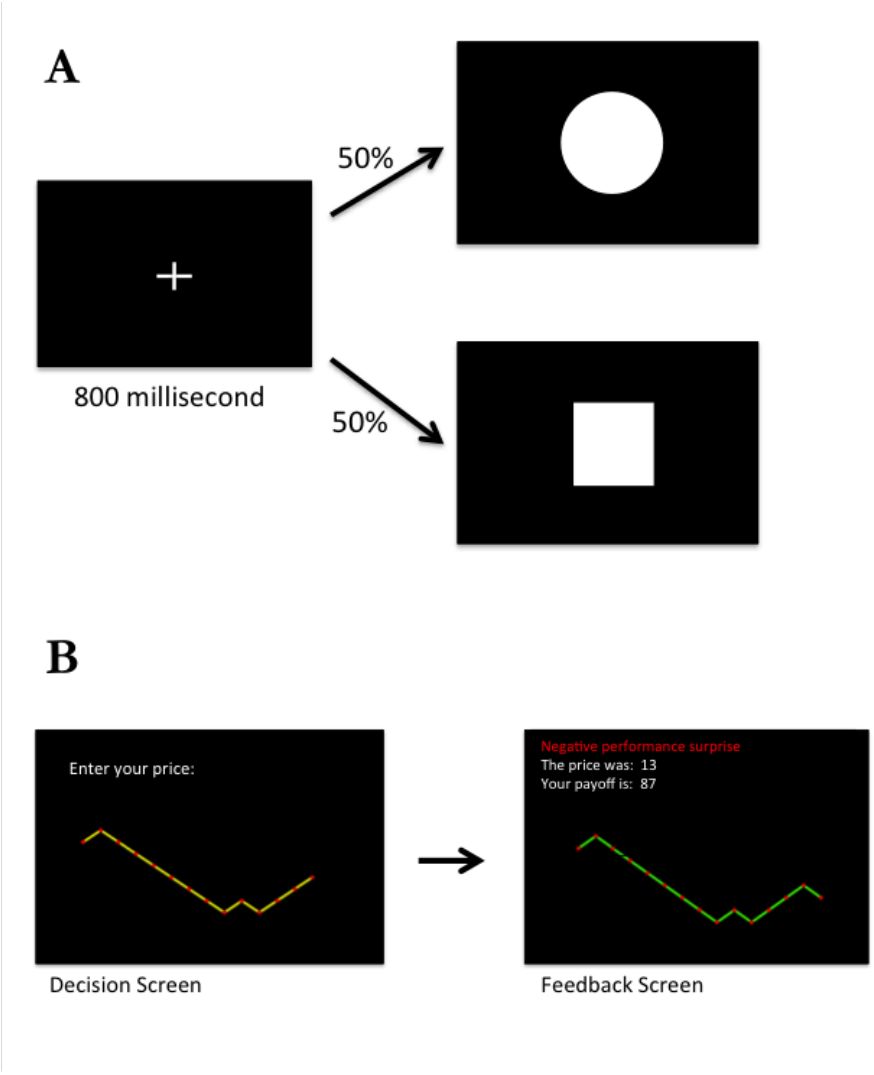


Figure 2.1: Experimental design of the task (EDT) and perceptual decision-makingtask (PDT). (A) PDT: Following a display of a fixation cross at the center of the screen (800 milliseconds), either a circle (p=.5) or a square was presented in random order over the course of 1200 trials. Subjects were incentivized to respond to each shape with a different key press as quickly and as accurately as possible. A new trial started immediately following the response, with the appearance of a new fixation cross. (B) EDT: On each of 400 trials, subjects entered the price, p, at which they would be willing to buy a stock. A price x was then randomly drawn and if x<p, the subject purchased the stock at a price of x on that trial. The stock then paid $100 if there was a positive performance surprise, and $0 otherwise.

of literature showing that AF effects do not occur when the response-stimulus interval (RSI) is greater than 250 milliseconds (e.g., Soetens, Boer, and Hueting, 1985). Indeed, as the objective of this study was to investigate expectation formation, we intentionally designed the experiment with an RSI of 800 milliseconds in order to

minimize AF sequential effects (Cho et al., 2002; Gao et al., 2009).

Basic results from the economic decision-making task

Consistent with previous research, we found substantial evidence of extrapolation based on the previous history of stimuli in the EDT (Bloomfield and Hales, 2002; Asparouhova, Hertzel, and Lemmon, 2009). Specifically, as shown in Figure 2.2c, we found that the longer the current streak of positive (negative) performance surprises, the higher the reported probability of a subsequent positive (negative) performance surprise ($p < 0.001$, see Table 2.6).

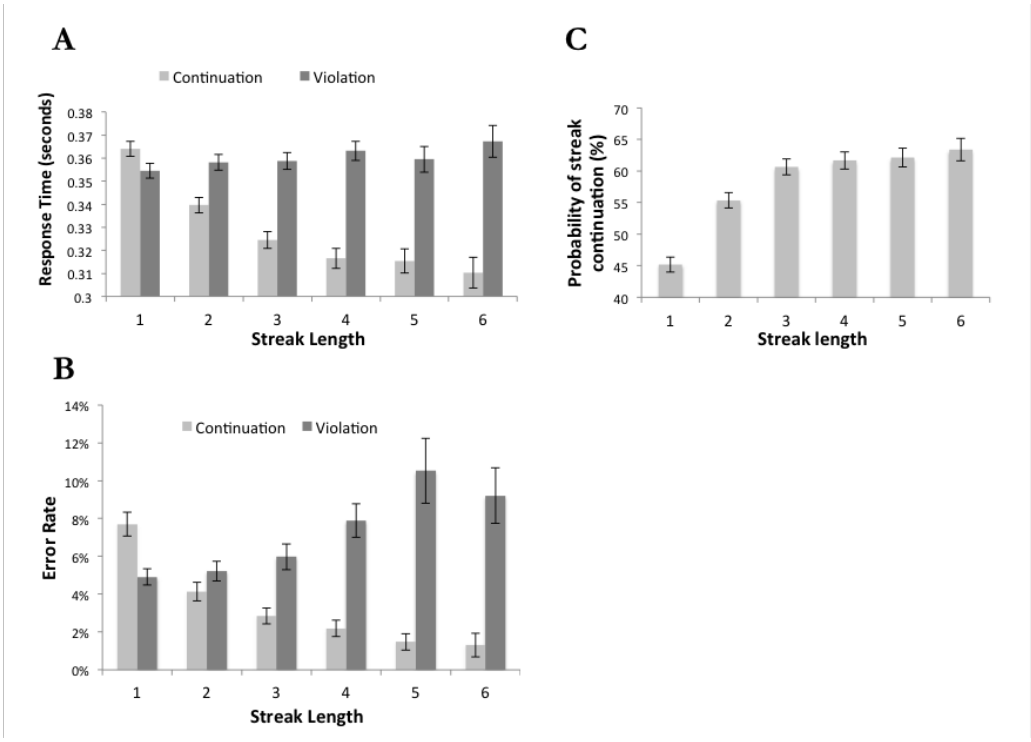


Figure 2.2: Basic experimental results (A) Average RT as a function of current streak length (PDT), where streak length is defined as the number of consecutive identical stimuli. Continuations are those trials where the streak continues; violations are those trials where the steak is violated. Data is shown only for correct trials (94% of the data) (B) Error rate as a function of current streak length (PDT). (C) Average reported beliefs (EDT) that the current streak would continue as a function of streak length. All error bars represent standard errors clustered at the subject level.

Structural model of decision-makingin the PDT

Drift Diffusion Model Overview

In order to investigate whether there is a common link between belief formation in the EDT and the PDT, we needed to first transform the RT and accuracy data from the

PDT into probabilities (prior beliefs), allowing us to compare them with the beliefs elicited in the EDT. This is a non-trivial exercise, but we were able to overcome this obstacle by estimating a structural model of the decision-making process in the PDT. While there is a variety of structural models of perceptual decision-making in the psychology and neuroscience literature, we chose to model our data using a drift-diffusion model (DDM) for two main reasons. First, in addition to providing an accurate description of perceptual decision-making data, recent work has shown that this model can also provide a good fit for behavior and response times in economic decision-making tasks (Fehr and Rangel, 2011; Krajbich and Rangel, 2011). As our goal is to investigate the link between belief formation in perceptual and economic decision-making, it is valuable to use a structural model that can explain the data well in both domains. Second, there are specific parameters of the DDM that have been shown to map on to changes in prior beliefs, which is the key object in our study (Mulder et al., 2012; A. R. Teodorescu and Usher, 2013; White and Poldrack, 2014).[1]

The DDM was originally developed to explain the response times and accuracy of perceptual decisions in binary choice tasks (Ratcliff, 1978; Ratcliff and McKoon, 2008). The basic assumptions of this model are that incoming sensory evidence about the identity of a stimulus (e.g., a circle or square) is noisy and decision time is costly. The DDM implements a choice algorithm that minimizes the decision time for a given level of accuracy (Bogacz et al., 2006).[2] A brief description of the DDM as applied in our experimental setting is useful to understand the analytical technique we use to decode prior beliefs from the PDT. The DDM assumes that the brain computes a relative decision value (RDV) that measures the accumulated relative "evidence" in favor of the correct option, and this RDV evolves over time until a choice is made (Figure 2.3).[3] The RDV follows a diffusion process:

$$dRDV(t) = Mdt + sdW, \tag{2.1}$$

with initial point $RDV(0) = c_i$, for condition i. A choice is made once the RDV reaches one of two thresholds, where we normalize the lower threshold to zero and

[1] In the next sub-section, we test which specific parameters from the DDM encode changes in prior beliefs.

[2] Conversely, it can be seen as solving the dual optimization problem of maximizing decision accuracy for a given amount of decision time.

[3] See A. R. Teodorescu, Moran, and Usher, 2015 for a recent discussion on the comparison between relative and absolute decision values.

the upper threshold to a constant, a. In this model, M is the drift rate and represents the strength of incoming sensory information that a subject uses to infer the identity of the current shape. When the discriminability between the two possible stimuli is high, the drift rate is large; if instead, the two shapes are difficult to discriminate, then the incoming sensory evidence in favor of one option versus the other is low and the drift rate will be small. The variable c_i represents the initial point in condition i and can parameterize the prior bias towards selecting the correct alternative (we use the convention that the upper boundary is associated with the correct alternative). Finally, s represents the standard deviation of mean-zero Gaussian distributed noise, which we set to $s = 0.1$ without loss of generality, and dW is a Weiner process.[4]

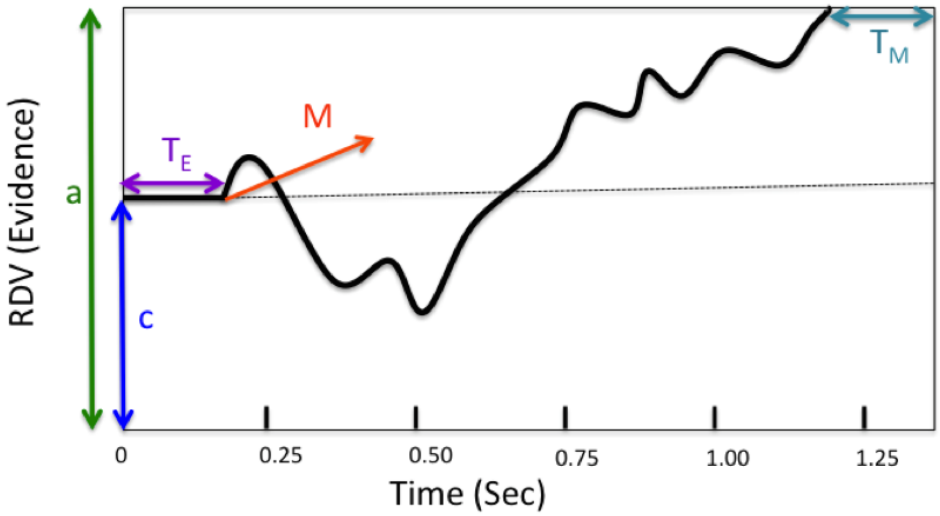


Figure 2.3: A graphical illustration of the drift diffusion process. The bold path indicates the evolution of the relative decision value (RDV) that tracks the relative evidence in favor of the alternative associated with the upper boundary. TE denotes the time required for encoding the stimulus, TM denotes the time required for making the motor response (such that the non-decision time, T, equals to the sum of TE and TM). c denotes the initial point that captures prior bias, a denotes the upper boundary that captures the speed-accuracy trade-off, and M denotes the drift rate that represents the quality of sensory input. When the RDV reaches a boundary, the process terminates and a decision is made. Without loss of generality, the lower boundary is set to zero.

The non-decision time, denoted by T, represents the time required to encode the

[4]Because the noise parameter s, drift rate M, and boundary separation a, are only defined up to positive affine transformations, one could fix any of these three parameters and estimate the remaining two. We choose to follow the convention of the studies that fix the noise parameter and estimate the boundary and drift rate (Ratcliff and Smith, 2004). The boundary and drift rate are then interpreted in units of standard deviation of noise.

stimulus (TE) and implement the motor action (TM). In addition to the four parameters (M, a, c_i, and T) we include three additional across trial variability parameters that have been shown to improve the accuracy of describing observed RT distributions (Ratcliff and McKoon, 2008). First, a variance parameter M_{σ^2}, characterizes the distribution from which the drift rate is sampled in each trial, such that M is normally distributed with mean M_μ and variance M_{σ^2}. Second, a range parameter T_{σ^2} characterizes the distribution from which the non-decision time is sampled in each trial, such that $T \sim U[\frac{T_\mu - T_{\sigma^2}}{2}, \frac{T_\mu + T_{\sigma^2}}{2}]$. Finally, a range parameter c_{σ^2} characterizes the distribution of the initial point, such that in condition i $c_i \sim U[\frac{c_{i\mu} - c_{\sigma^2}}{2}, \frac{c_{i\mu} + c_{\sigma^2}}{2}]$. We estimated the DDM at the individual subject level using the DMA-Toolbox (Vandekerckhove and Tuerlinckx, 2008). This toolbox estimates the DDM parameters by minimizing the multinomial log-likelihood function:

$$LL(x_t, b_t|\theta_i) = -2 \sum_{c=i}^{C} \sum_{t=1}^{T_i} log(L(x_t, b_t|\theta_i)), \tag{2.2}$$

where $L(x_t, b_t|\theta_i)$ is the likelihood of observing, on trial t, a response (correct or incorrect) x_t and RT in bin b_t, conditional on parameters θ_i, where i denotes the experimental condition for trial t. Following the literature on sequential effects in perceptual decision-making we define an experimental condition by the 4-back history of repetitions (R) and alternations (A), thus setting $C = 16$ total conditions (Soetens, Boer, and Hueting, 1985; Cho et al., 2002; Yu and Cohen, 2009; Jones et al., 2013; M. H. Wilder et al., 2013).

Priors can be encoded by other DDM parameters

Our baseline model defines the parameter vector $\theta_i = [M, a, c_i, T, M_{\sigma^2}, T_{\sigma^2}, c_{\sigma^2}]$ which implies that only the initial point is allowed to vary across the C experimental conditions. This parameter restriction assumes that if prior probabilities vary across conditions, they must be encoded in the initial point. This baseline model is motivated by recent work showing that variation in prior beliefs is explained by variation in initial points, instead of by drift rates (Mulder et al., 2012; A. R. Teodorescu and Usher, 2013; White and Poldrack, 2014). However, the debate on whether priors are encoded in the initial point or the drift rate is still ongoing (Gao et al., 2009; Rorie et al., 2010; Ravenzwaaij et al., 2012) and recent work has developed experimental paradigms to examine this particular issue (Mulder et al., 2012; A. R. Teodorescu and Usher, 2013). Because this remains an open question in the literature, we estimated several additional versions of the DDM to investigate whether allowing drift rates to vary across experimental conditions yielded a better fit to our data. When

comparing the baseline model to a model where both drift rates and initial points are allowed to vary, $\theta_i = [M_i, a, c_i, T, M_{\sigma^2}, T_{\sigma^2}, c_{\sigma^2}]$, we find that the baseline model performs better in 37 of 38 subjects according to the Akaike Information Criterion (AIC). Furthermore, when comparing the baseline model to a model where only the drift rate is allowed to vary, we find that the baseline model performs better in 31 of 38 subjects. Another analysis that can help distinguish whether prior bias is encoded in the drift rate or initial point investigates the difference in average RTs on correct and incorrect trials. In particular, we followed Mulder et al., 2012 and defined (i) a valid prior trial as a trial where either a stimulus repetition had followed two or more repetitions, or an alternation had followed two or more alternations; (ii) an invalid prior trial as a trial where stimulus alternation had followed two or more repetitions, or repetition had followed two or more alternations; and (iii) a neutral prior trial as all other trials. The key prediction is that on valid trials where subjects respond incorrectly, RTs should be fast if prior bias is encoded in the drift rate and slow if prior bias is encoded in the initial point. In line with this hypothesis, we found that when making correct responses, subjects were fastest in trials with valid priors compared to neutral priors, and slowest in trials with invalid priors; critically, the opposite pattern was found for incorrect trials (see Figure 2.4). These results support the notion that for the PDT in our setting, changes in the prior are better captured by changes in the initial point of evidence accumulation rather than changes in the drift rate.

Finally, although previous research on sequential effects in binary perceptual decision-making tasks often sets the number of conditions to C=16 (Cho et al., 2002; Yu and Cohen, 2009), this itself is a parameter that is set by the researcher. To investigate this parameter choice, we re-estimated the baseline DDM allowing C to vary over the set $[2, 4, 8, 16]$ by collapsing the number of repetition and alternation histories accordingly. We find that the baseline DDM that uses 16 conditions performs better in 35 of 38 subjects (according to AIC) than the baseline DDM that use 2, 4, or 8 conditions. Similarly, when comparing a model where only the drift rate is allowed to vary across conditions, the model with 16 conditions performs best according to the AIC in 36 of 38 subjects. Taken together, the baseline model where only initial points are allowed to vary across 16 experimental conditions is the best fit, and we therefore use this model in all subsequent analyses.

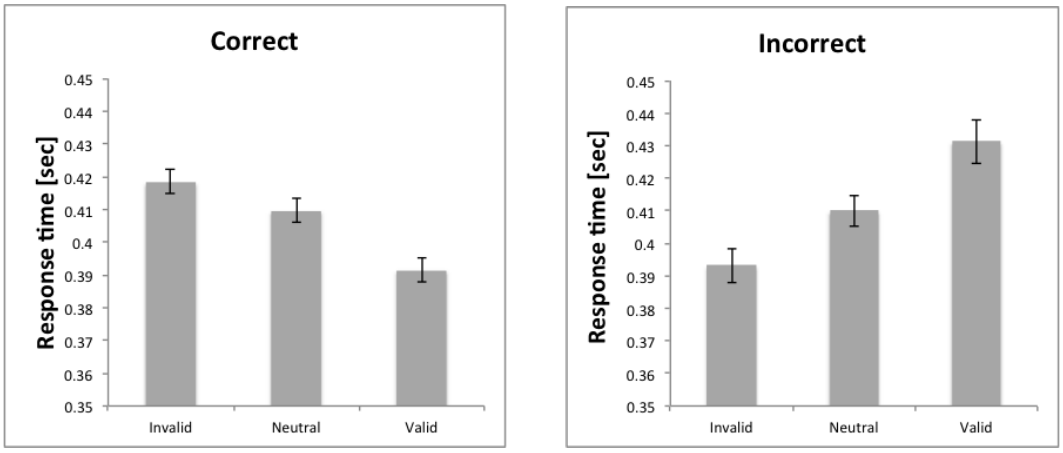


Figure 2.4: Average response times of correct and incorrect responses following "valid", "neutral" and "invalid" cues.
Valid cues are repetition following two or more repetitions or alternation following two or more alternations). Invalid cues (i.e., alternation following two or more repetitions or repetition following two or more alternations) and 'neutral' cues are all other trials.

Decoding prior beliefs using the DDM

Using the estimated DDM parameters, we decoded the prior probability for each subject and condition. To understand how the decoding process works, recall that the drift rate of the DDM, M, encodes the informativeness of the sensory signals in discriminating between the two shapes. At every instant within a trial, a new noisy signal is sampled where the noise is governed by the volatility of the process, s^2. All else equal, when M decreases, the signal to noise ratio decrease and a subject must rely more heavily on his prior belief. In the limit, when the drift rate goes to zero, the subject relies exclusively on his prior. In the appendix, we analytically solve for the probability of hitting the upper boundary when the drift rate goes to zero, and find that, in condition i, this prior equals $\frac{c_i}{a}$. Using this analytical result, we then collapsed the data from 16 to 8 conditions, because the identity of the current trial does not vary with the prior (e.g., trials in condition $AAAR$ and trials in condition $AAAA$ provide information only about $Pr(A|AAA) = 1 - Pr(R|AAA)$). Figure 2.5 plots the average priors for each of the eight conditions (all possible three element histories) for both the PDT and EDT. The Figure shows that prior probabilities are indeed a function of the recent stimulus history, and that these priors are highly correlated across the PDT and EDT ($r = 0.90, p < 0.005$).

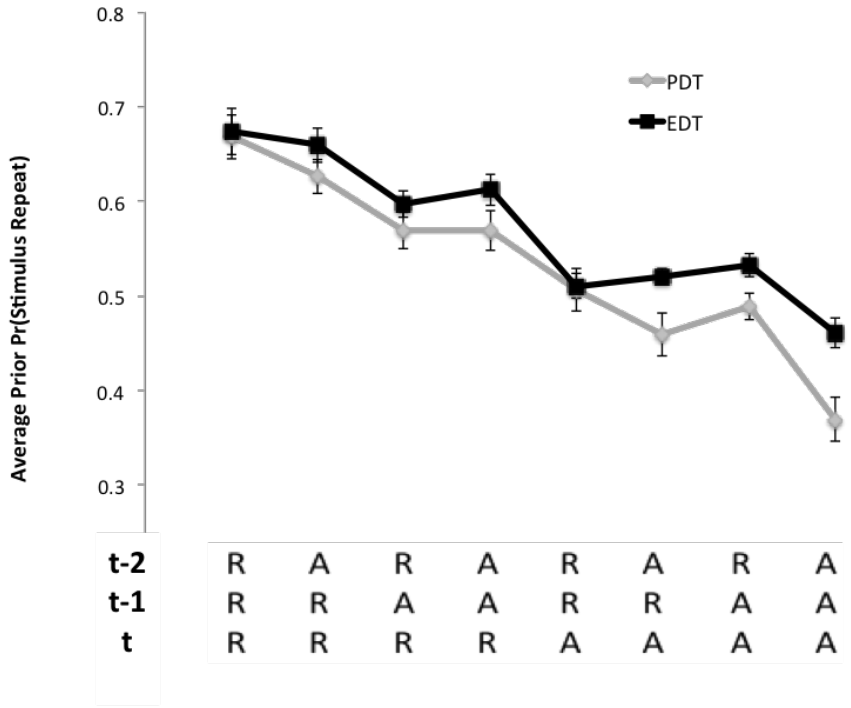


Figure 2.5: Average priors for the EDT and PDT as a function of the four most recent stimuli history. For each of the eight different conditions, the black line shows the average belief that a repetition will occur on the subsequent trial in the EDT, elicited using the BDM procedure. The gray line shows the average prior that a repetition will occur on the subsequent trial in the PDT, decoded from the initial point and boundary parameters of the DDM.

Individual differences in belief formation across tasks

Using our decoding strategy, we found significant individual differences in the extent to which priors in the PDT deviated from the rational prior of 0.5. To quantify this deviation, for each subject u, we computed the sum of squared deviations from 0.5 and define this as the irrationality index (II):

$$II_u = \frac{1}{8}\sum_{i=1}^{8}(p_{i,u} - 0.5)^2. \tag{2.3}$$

Because subjects were explicitly told that the probability of seeing either shape was 0.5, independent of the stimulus history, a fully rational subject would exhibit an irrationality index of 0 in the PDT. Instead, we found that the average II across subjects was 1.93 (SD: 0.391), which is significantly greater than the optimal level of 0 ($p < 0.001$, Tobit regression left-censored at 0). If the extrapolative beliefs from the EDT are driven by the same psychological mechanism that generated the irrationality in the PDT, then the II should explain a portion of the cross-subject

heterogeneity in the extrapolation of beliefs in the EDT (Appelt et al., 2011). To test this, we defined an Extrapolation Index (EI) for each subject u, as

$$EI_u = \frac{1}{400}\sum_{t=1}^{400}(b_{t,u} - 0.5)^2, \tag{2.4}$$

where $b_{t,u}$ is the belief reported by subject u in trial t of the EDT that a repetition would occur on trial $t + 1$. As illustrated in Figure 2.6a, we found a significant positive correlation ($r(38) = 0.57, p < 0.001$) between the II and the EI.

One potential alternative explanation for the correlation between the EI and the II is that it is driven by task engagement. In particular, subjects who exhibit a high level of engagement in each task might also exhibit a greater tendency to perceive local patterns in recent stimulus history, which can then generate both a high EI and a high II. To rule out this alternative explanation, we use the estimated within trial noise parameter from the full DDM, denoted by s in equation 2.1, as a measure of subject-specific task engagement.[5] We then regressed the EI on the II while including each subject's within trial noise parameter, s, as a control variable. As column (1) of Table 2.1 shows, the II was still a significant predictor of the EI ($p < 0.001$), while the within trial noise parameter was not a significant predictor ($p = 0.651$). As an additional robustness check, we estimated an OLS model that included the across trial variability in drift rate M_{σ^2} as a control. This approach was motivated by recent experimental work showing that the across trial variability parameter in drift rate, M_{σ^2}, provides a good proxy of arousal during perceptual decision-making (Murphy, Vandekerckhove, and Nieuwenhuis, 2014). As task engagement is typically a quadratic function of arousal, we include both M_{σ^2} and the quadratic term $M^2_{\sigma^2}$, in a multiple regression. Column (2) of Table 2.1 shows that the II was a significant predictor of the EI (p<0.001), but neither M_{σ^2} nor $M^2_{\sigma^2}$ predicted the EI ($p = 0.322$ and $p = 0.360$, respectively). Column (3) provides an analogous regression using the standard deviation of RTs from the PDT as a control for task engagement, and column (4) includes all controls in a single model. Again, we found that the II remained a significant predictor of the EI after including these additional controls.

[5] In our original estimation of the full DDM, we set $s = 0.1$ (because core model parameters are only defined up to ratios of one another). We therefore re-estimate the full DDM parameters when fixing the boundary parameter $a = 0.1$, and allowing s to be a free parameter.

Table 2.1: OLS regression of extrapolation index (EI) on the irrationality index (II) and controls. The within trial and across trial noise parameters – estimated at the individual level from the DDM – are used to control for levels of task engagement. Standard errors are clustered by subject.

	Dependent variable: Extrapolation index			
Irrationality index	0.635^{***}	0.649^{***}	0.613^{***}	0.611^{***}
	(0.155)	(0.157)	(0.150)	(0.152)
Within trial noise	0.005			0.005
	(0.012)			(0.013)
Across trial noise		0.308		−0.118
		(0.307)		(0.121)
Across trial noise^2		−0.749		
		(0.807)		
Standard deviation of RT			0.183	0.269
			(0.104)	(0.139)
Constant	-0.088^{**}	-0.094^{**}	-0.101^{**}	-0.107^{**}
	(0.038)	(0.039)	(0.037)	(0.038)
Observations	38	38	38	38
R^2	0.328	0.344	0.379	0.398

Note: $^{*}p < 0.1$ $^{**}p < 0.05$ $^{***}p < 0.01$

Finally, we investigated whether individual differences in the payoff from the EDT can be explained by the II from the PDT. After all, if subjects hold extrapolative beliefs in the EDT, this should lead to low overall payoffs in the EDT since there is no predictability in the actual process that generated the EDT performance surprises.[6] If these extrapolative beliefs are governed by the same mechanism that generates the II in the PDT, then subjects with a higher II should have lower payoffs in the EDT. Consistent with this hypothesis, we found a significant negative correlation between the II and the payoff in the EDT, demonstrating that irrational behavior in the perceptual domain predicts performance in the economic domain ($r = -0.47, p < 0.01$, Figure 2.6b).[7]

[6] In line with this hypothesis, there was a strong negative correlation between payoffs in the EDT and the extrapolation index ($r = -0.73$, $p < 0.001$).

[7] Although we find a significant correlation between the irrationality index and payoffs in the

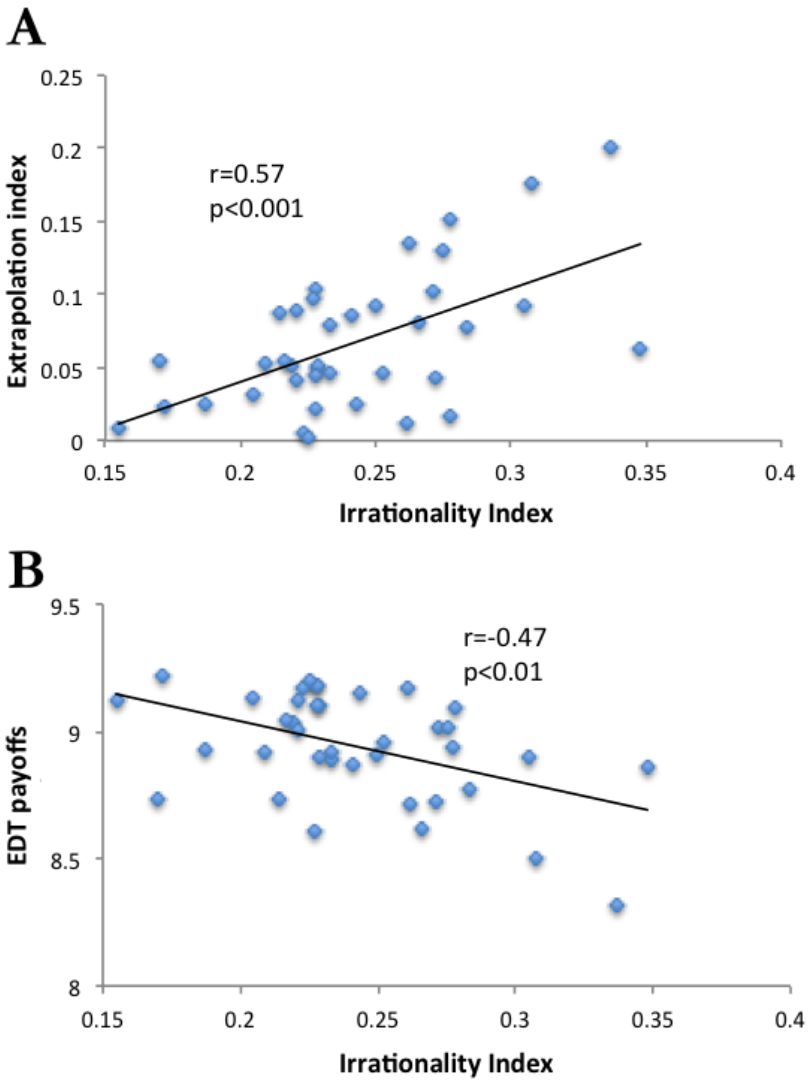


Figure 2.6: Individual differences. (A) Correlation across subjects between the extrapolation index and the irrationality index. (B) Correlation across subjects between the irrationality index and the EDT payoffs. Each point represents a single subject.

EDT, we do not find a significant correlation between the irrationality index and payoffs in the PDT. We reason that this may be because error rates were very low across all participants, and a significant component of the payoff structure for the PDT was based on error rates. Thus, the lack of a significant correlation between the II and PDT payoffs is most likely driven by the low variation across subjects in PDT payoffs.

A common dynamic belief model

So far we have shown that (i) conditional on recent stimulus history, the average beliefs in the EDT and PDT are correlated (Figure 2.5) and (ii) individual differences in the degree to which beliefs deviate from the 0.5 benchmark in the EDT and PDT are correlated across subjects (Figure 2.6A). These two pieces of evidence suggest that a common belief formation mechanism governs subjects' responses in both tasks. To further investigate this conjecture, we use a Bayesian model from the computational neuroscience literature, the Dynamic Belief Model (DBM) (Yu and Cohen, 2009; M. Wilder, Jones, and Mozer, 2009; Zhang, Huang, and Angela, 2014). The DBM relies on two key assumptions. First, the agent believes that the probability of observing a repetition on trial t (i.e., the stimulus on trial t matches the stimulus on trial $t-1$) is governed by a parameter γ_t. Second, the parameter γ_t is time-varying and changes on each trial with a constant probability α. These assumptions are motivated by empirical evidence in cognitive neuroscience, showing that the brain is well adapted to learning in non-stationary environments (Behrens et al., 2007; Nassar et al., 2012). Furthermore, the DBM can be well approximated by an exponential filter (Yu and Cohen, 2009), which has received empirical support both at the behavioral and neurophysiological levels (Sugrue, Corrado, and Newsome, 2004). To fix ideas, we describe the model in the context of the PDT. Let $X_t \in \{Square, Circle\}$, and suppose the agent believes that the state of the world is captured by γ_t, which represents the time-varying repetition rate, $Pr_t(X_t = X_{t-1})$. On each trial, with probability α, γ_t is resampled from a reset prior P_0 that is uniform over $[0, 1]$ (Fox and Rottenstreich, 2003). Formally, let $z_t = 1$ if $X_t = square$, and let $z_t = 0$ if $X_t = circle$, so that the probability that the next instance, X_t, is a square is as follows:

$$\begin{aligned} Pr(X_t = square | X_{t-1}, X_{t-2}, ..., X_1) = \\ Pr(X_t = square | \gamma_t, X_{t-1}) = \\ z_{t-1}\gamma_t + (1 - z_{t-1})(1 - \gamma_t). \end{aligned} \tag{2.5}$$

The model's prediction depends explicitly only on the most recent observation, X_{t-1} and on the current estimate of γ_t. The DBM algorithm operates iteratively by maintaining a prior distribution over γ_t, $Pr(\gamma_t, X_{t-1}, \gamma_{t-1})$. After observing a new stimulus, the posterior, $\hat{P}(\gamma_t | X_t, \gamma_{t-1})$ is computed using Bayesian updating:

$$\hat{P}(\gamma_t | X_t) \propto P(X_t | \gamma_t, X_{t-1}) P(\gamma_t | X_{t-1}). \tag{2.6}$$

The posterior of the current trial is then used to compute the prior for the next trial, as a sum of the posterior weighted by $(1 - \alpha)$ and the reset prior weighted by α:

$$P(\gamma_{t+1}|X_t) = (1 - \alpha)\hat{P}(\gamma_t|X_t) + \alpha P_0(\gamma_{t+1}). \tag{2.7}$$

The model generates predictions, $P(X_t|X_{t-1})$, by integrating Eq. 2.7 over the prior on γ_t. In our simulations, we maintain a discrete approximation to the continuous prior by dividing the interval [0, 1] into 100 equally spaced bins, where expectations are computed by summing over the discrete probability mass function. To test whether the DBM provides a common computational model of belief formation across both the EDT and PDT, we estimated α at the subject level for the PDT and found that the mean level was $\alpha^\star$. We then used $\alpha^\star$ to calibrate a DBM for the EDT, and computed the DBM predicted time series of beliefs for the EDT. Figure 2.7 shows the DBM theoretical predictions plotted against the actual average beliefs from the EDT. The two time series exhibit a strong positive correlation ($r(400) = .66$, $p < 0.001$).[8] Moreover, the mean level of α from the EDT is 0.46 - almost identical to the mean level of α from the PDT of 0.44. Figure 2.8 shows the sum of squared errors from the EDT estimation as a function of α, and indicates that the global minimum is indeed at 0.44.

To provide a more formal statistical analysis of this result, we ran an OLS regression, where the dependent variable is subject i's response on trial t in the EDT. In column (1) of table 2.2, we specify a model with a single independent variable (and a constant), defined by the EDT predicted time series calibrated from the average α of the PDT (this calibrated time series is illustrated by the dark gray time series in Figure 2.7). Consistent with the result displayed in Figure 2.7, we find that the average predicted time series is a strong predictor of observed behavior in the EDT ($p < 0.001$). In column (2), we also included the individual EDT predicted time series, calibrated from the individually estimated subject level α's from the PDT. The individually calibrated EDT predictions explain additional variation in observed EDT behavior, even after controlling for the average EDT predicted time series ($p = 0.035$). Furthermore, the model that included the individual specific EDT predictions provided a better fit to the data than the model that only included

[8] We ran an additional analysis where we first computed the thirty-eight predicted time series for the EDT using the thirty-eight individually estimated alphas, and then averaged these time series. This average time series exhibits a 0.686 correlation with the actual average time series from the EDT, thus serving as a robustness check.

the average predictions, as indicated by the lower AIC in model (2) compared to that of model (1).

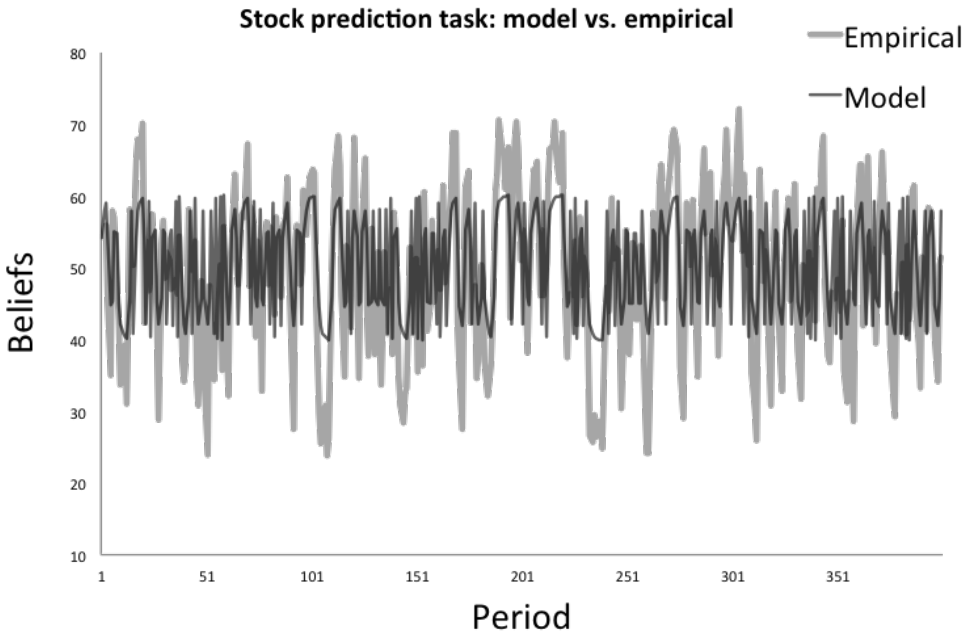


Figure 2.7: Out of sample DBM-based predictions of the average EDT beliefs. The predictions are calibrated from the PDT priors. A DBM model is estimated for each subject from the PDT. We then input the stimuli from the EDT into the DBM using the average α across subjects from the PDT, and generate a time series of theoretical predictions (red). The empirical average beliefs from the EDT are plotted in blue. The correlation between the two time series is .66($p < 0.001$).

2.4 Discussion

Our experimental results support the hypothesis that belief formation in perceptual and economic decision-making is governed, at least in part, by a common psychological mechanism that is rooted in Bayesian models of decision-making (Chater, Tenenbaum, and Yuille, 2006; Oaksford and Chater, 2009). Intriguingly, the use of the DBM as a computational strategy in the PDT is sub-optimal, because subjects are explicitly told the data generating process, and hence there is no reason to learn the underlying model parameters. When viewed separately, the results from the PDT and EDT are successful replications of several studies that have employed similar tasks, as RTs and error rates in the PDT and subjective beliefs in the EDT are heavily dependent on recent stimulus history (Bloomfield and Hales, 2002; Cho et al., 2002; Huettel, Mack, and McCarthy, 2002; Asparouhova, Hertzel, and Lem-

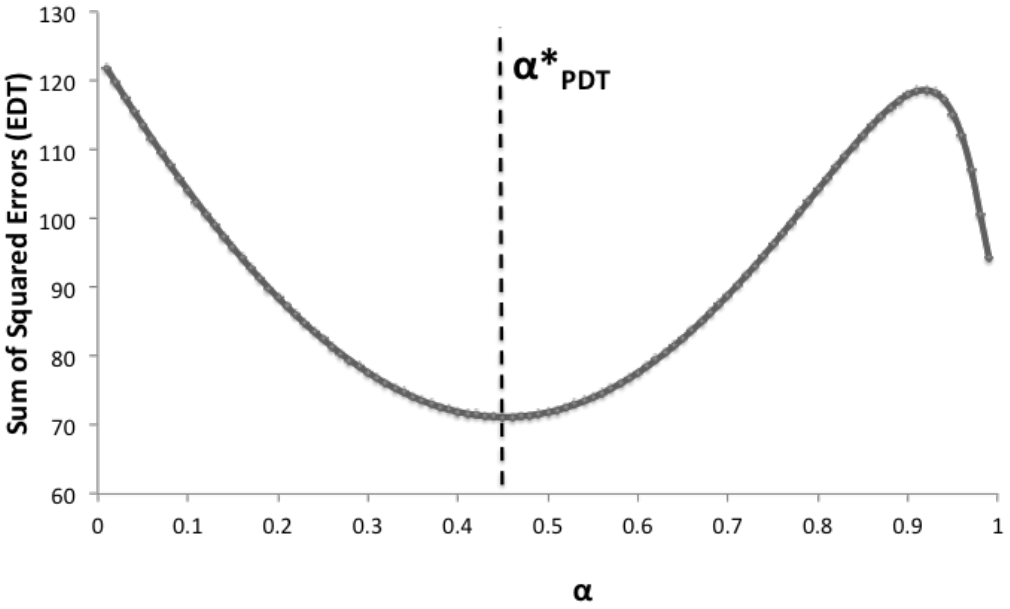


Figure 2.8: The sum of square errors of the DBM prediction for the average reported beliefs in the EDT, as a function of parameter α. The mean parameter found in the PDT, α^*, is marked by the dashed line.

Table 2.2: Individual differences in PDT predict behavior in EDT. The dependent variable is subject i's response on trial t of the EDT. "Average prediction" is the model prediction using the mean α from the PDT. "Individual prediction" is the model prediction using individual estimates of α from the PDT. Standard errors are in parentheses and are clustered by subject.

	Dependent variable: Beliefs (EDT)	
Average prediction	1.06^{***}	0.900^{***}
	(0.142)	(0.145)
Individual prediction		0.333^{*}
		(0.152)
Constant	−0.005	−0.130
	(0.073)	(0.089)
Observations	15, 200	15, 200
AIC	791	747
R^2	0.080	0.083

Note: $^{*}p < 0.1$ $^{**}p < 0.05$ $^{***}p < 0.01$

mon, 2009). However, by employing a within-subject design, we were also able to compare the computational processes underlying belief formation across the tasks.

Because subjects were explicitly told that the probability of seeing a square on each trial in the PDT was 0.5, an optimal Bayesian agent using the DBM would set α equal to zero (i.e., the autocorrelation parameter would change with 0 probability on each trial). Instead, we found the average level of α in our subject pool was 0.44. This suggests that subjects may have a strong prior on a non-stationary model of the world, and they have difficulty adjusting their belief formation mechanism in the face of explicit information that would increase earnings (Yu and Cohen, 2009). Moreover, the average value of α in the EDT was 0.46, nearly identical to the parameter's average value in the PDT (see Figure 2.8). This similarity in α rules out the possibility that working memory constraints are responsible for subjects' strong reliance on recent stimulus history; because we display the previous fourteen stimuli in the EDT and we display no previous stimuli in the PDT, we would expect differences in average α if behavior was driven by working memory constraints, but we do not find this. Moreover, the fact that individual differences in α from the PDT (in which no historical data is available at the time of decision-making) can explain behavior in the EDT also casts doubt on the working memory hypothesis.

It is important to highlight that the combination of the DDM with the DBM is just one of many models that can be used to explain sequential effects in perceptual decision-making at long RSIs. For example, another candidate model is proposed by Gao et al., 2009 and allows for a unified explanation of both within and across trial dynamics over a broad range of RSIs. While this model surely provides a generalized and detailed account of the sequential effects in our data, we choose to model our data with a joint DBM-DDM because the DBM is a portable model that can be applied directly to many other settings, including the EDT. As the core question in our study is concerned with the relationship between belief formation across domains, the ability to use the same single parameter model across separate tasks is especially useful. Moreover, the DDM is increasingly being used to model economic decisions, and we therefore believe it is valuable to use a model that has been shown to explain data well in both perceptual and tasks (Fehr and Rangel, 2011). Furthermore, the key feature of the DBM that enables it to flexibly explain sequential effects is the assumption that the autocorrelation parameter is perceived to be time varying, and that subjects use Bayesian inference to update their estimate of this parameter. While it may be difficult for subjects to implement these precise Bayesian computations, it has been shown theoretically that the DBM is well approximated by an exponential filter (Yu and Cohen, 2009), which has received empirical support both at the behavioral and neurophysiological levels (Sugrue, Corrado, and Newsome, 2004).

This property is especially relevant to the current study because recent theories of belief formation in economics and finance explicitly use this exponential decay property (Camerer and Hua Ho, 1999; Malmendier and Nagel, 2009; Malmendier and Nagel, 2015; Barberis, Greenwood, et al., 2015).

At a methodological level, one contribution of our study is to provide the analytical framework that enabled us to measure the trial-by-trial subjective beliefs in the PDT. In particular, we develop a new methodology that can be used in future work on perceptual decisions to elicit prior probabilities. While most previous DDM research has focused on estimating computational parameters as a function of experimental conditions, we instead use the estimated computational parameters to reverse-engineer the belief formation process. While response time and accuracy data have been used for decades to infer mental representations of the environment (Luce, 1986; Achtziger and Alós-Ferrer, 2013; Jones et al., 2013), our study provides the extra structure necessary to map RTs into exact probabilities, thus providing a common framework for studying beliefs in perceptual and economic decisions.

One potential concern about our experimental design is that the EDT and PDT are not substantially different from each other. Indeed, the stimuli in both the EDT and PDT are binary processes and both tasks require making judgments about the underlying state of the world. However, the tasks do differ on three fundamental dimensions. First, in the EDT subjects are asked to make judgments about the likelihood of a future event (a firm's performance) whereas in the PDT subjects are asked to make fast motor responses about the identity of the currently displayed stimulus. In other words, the subjects in the EDT rely on cognitive resources to predict the future whereas in the PDT they rely on perceptual resources to classify the present state of the world. Another major difference between the two tasks is the information that is given to subjects about the underlying data generating process. In the PDT subjects are explicitly informed that the probability of observing each stimuli is 0.5, whereas subjects are not told anything about the process governing the stimuli in the EDT. Finally, the tasks differ with respect to the information that is available to subjects about the stimulus history. In the EDT, subjects are given access on-screen to a history of the previous fourteen stimuli; in the PDT, only the current stimulus is displayed.

One could also argue that because of these different aspects across the two tasks, the individual differences in behavior that we document may be driven by two distinct mechanisms. Perhaps the strongest alternative hypothesis is that the sequential

effects we observe in the PDT are not driven by expectation formation, but are instead driven by post-response residual activity of the motor cortex. This type of effect has indeed been documented in the sequential effects literature, but it is most prevalent in tasks that features RSIs that are no longer than 250 milliseconds (Soetens, Boer, and Hueting, 1985). Because the goal of this study is to examine expectation formation explicitly, we designed the PDT with a long RSI of 800 milliseconds, for which expectation formation effects have been shown to dominate lower level AF effects that are driven by post-response residual activity (Gao et al., 2009). Our results also relate to the literature on belief-based decision biases in judgment and decision-making and behavioral economics. Numerous studies have documented that after seeing a sequence of identical stimuli (e.g. successful basketball shots or positive stock returns), humans have a tendency to extrapolate the past and believe that the streak will continue (Gilovich, Vallone, and Tversky, 1985; Greenwood and Shleifer, 2014). The origin of these extrapolative beliefs – also known as the "hot hand fallacy" – is still not completely understood, and multiple models have been proposed to explain it and its implications for financial markets (Rabin, 2000; Massey and Wu, 2005; Oskarsson et al., 2009; Asparouhova, Hertzel, and Lemmon, 2009; Barberis, Shleifer, and Vishny, 1998; Miller and Sanjurjo, 2014).[9]

Because of the common computational processes and the similar belief in continuing streaks across the two tasks, our results suggest that the extrapolative beliefs in economic decision-making may stem from low-level perceptual processes instead of deliberative analytical judgments. One possible interpretation is that the inability to maintain a constant initial point across trials in the PDT is driven by the fact that it is optimal to flexibly change the initial point from decision to decision in many other environments. This inability to maintain a constant initial point may then be inherited by the belief-formation mechanism deployed in economic decision-making, though we emphasize this is just a conjecture. Because subjects are explicitly told that the data generating process in the PDT does not contain any predictability, these perceptual pattern recognition processes may be difficult to suppress. Understanding whether extrapolative beliefs, and judgment biases in general, arise from deliberate

[9]While we focus on the tendency of humans to believe that a streak of identical stimuli will continue, there are also situations in which humans believe the streak will reverse, also known as the "gambler's fallacy." While the two effects may seem opposite of one another, theoretical work has shown that the gambler's fallacy can endogenously generate the hot hand fallacy through a reliance on the "law of small numbers" (Rabin, 2000). For a recent overview of both effects, see (Plonsky, K. Teodorescu, and Erev, 2015).

analytical processes or from bottom-up perceptual processes is important because the distinction has implications for policy-makers. If some biases are driven by low level processes, as our data suggests, then a policy that mandates additional information disclosure may not be effective at impacting decision-making (Brav and Heaton, 2002). While we examine only one specific decision bias in this study, future work may benefit from linking perceptual and economic decision-making via computational models to better understand the origins of other well-known behavioral biases (Tsetsos, Chater, and Usher, 2012).

APPENDIX

2.A Data preprocessing and order of experimental tasks

Preprocessing of reaction time data.

The PDT consisted of 4 blocks of 300 trials each. RTs and error rates systematically increased over the course of each block, likely due to subjects' fatigue (See Figures 2.9 and 2.10). To control for this, we removed a linear time trend, within each block, for each subject. All results and analyses in the text use this de-trended RT data and are robust to exclusion of the de-trending step.

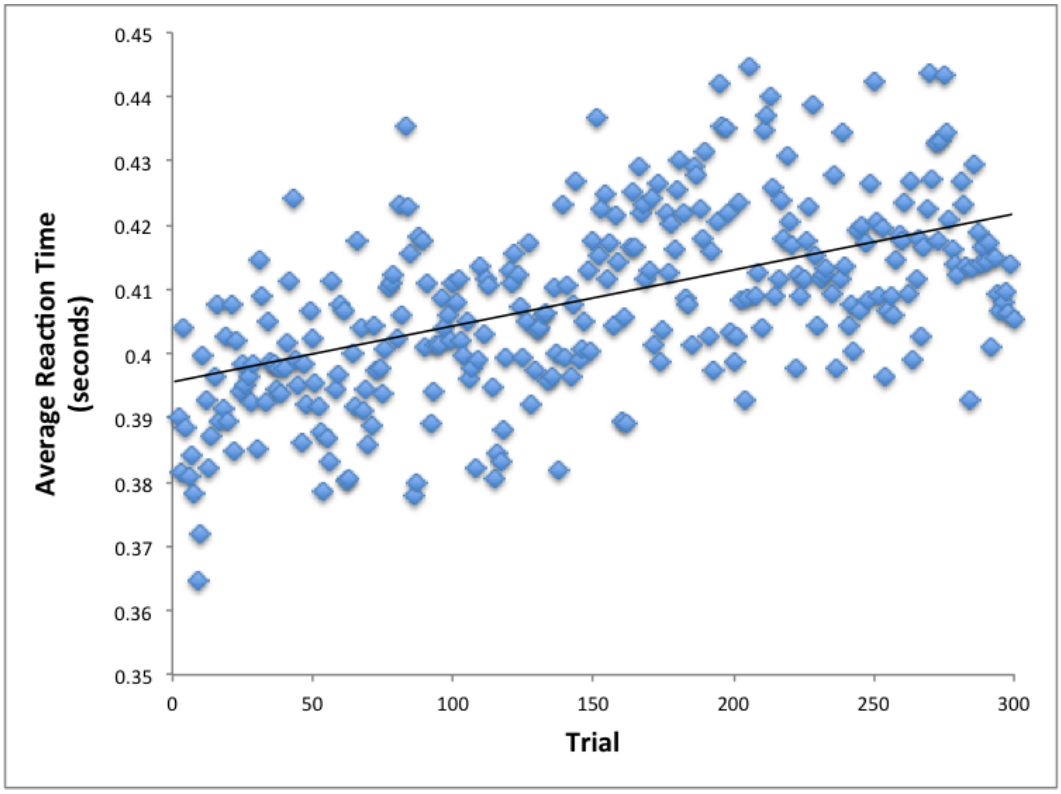


Figure 2.9: Average reaction times across subjects and across four blocks of trials (each data point is the average of four blocks across all subjects).

On the ordering of the tasks in the experimental session.

In our within-subjects design, the ordering of the tasks was not randomized and the PDT always took place first. We began the experiment with the PDT for all subjects because during pilot testing we observed a sharp fatigue effect in subjects' response times in this task (see Figures 2.9-2.11). We were therefore concerned that this fatigue effect would vary between subjects if the PDT was administered

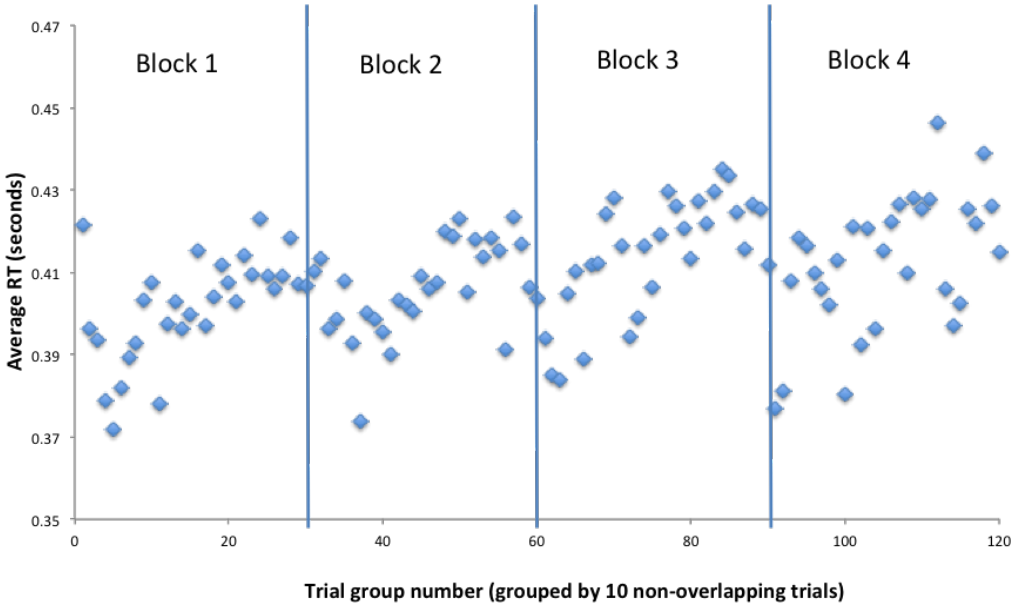


Figure 2.10: Average reaction times across subjects for each of the four blocks of trials (grouped by trials of 10).

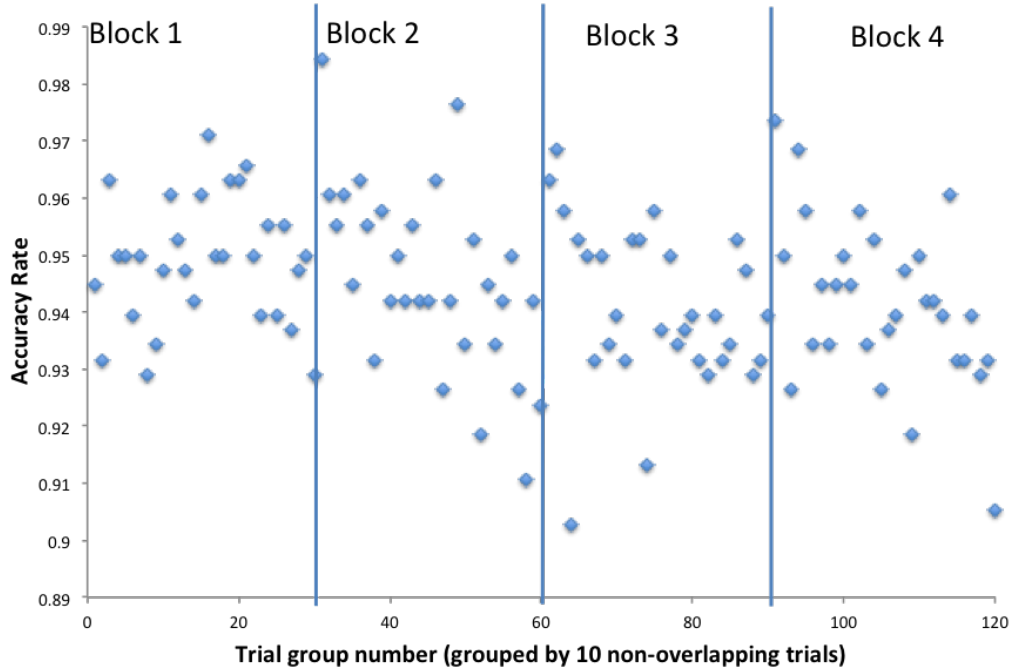


Figure 2.11: Average rate of correct responses across subjects for each of the four blocks of trials (grouped by trials of 10).

later in the session for a subset of subjects. Furthermore, because subjects in the PDT were explicitly informed that the probability of seeing either shape was 0.5, we believe that any possible spillover effects between the two tasks should bias us

against finding the extrapolation effect observed in the EDT (where subjects were not explicitly informed about the underlying random process).

2.B Robustness checks and extended statistical tests

The tables below summarize mixed model regressions estimating the effects of streak length on response times and error rates (PDT) and subjective beliefs (EDT).

Table 2.3: Mixed model linear regression (subject random intercepts and slopes), PDT response times (correct)

	Dependent variable:
	Adjusted RT (Correct trials)
Streak Length	0.002^{***}
	(0.001)
Continuation	0.020^{***}
	(0.004)
Streak Length x Continuation	-0.015^{***}
	(0.001)
Constant	0.394^{***}
	(0.008)
Observations	42,476
Log likelihood	36,134.610
AIC	$-72,663.220$
Note:	$^{*}p < 0.1$ $^{**}p < 0.05$ $^{***}p < 0.01$

Table 2.4: Mixed model logistic regression (subject random intercepts and slopes), PDT accuracy

	Dependent variable: Correct = 1
Streak Length	-0.136^{***}
	(0.028)
Continuation	-1.129^{***}
	(0.103)
Streak Length x Continuation	0.642^{***}
	(0.042)
Constant	3.240^{***}
	(0.100)
Observations	44,992
Log likelihood	$-9,327.701$
AIC	18,677.400
Note:	$^{*}p < 0.1$ $^{**}p < 0.05$ $^{***}p < 0.01$

Table 2.5: Mixed model linear regression (subject random intercepts and slopes), PDT response times (incorrect)

	Dependent variable: Adjusted RT (Incorrect trials)
Streak Length	-0.001^{***}
	(0.006)
Continuation	-0.037^{***}
	(0.020)
Streak Length x Continuation	0.011^{***}
	(0.010)
Constant	0.394^{***}
	(0.021)
Observations	2,516
Log likelihood	-274.151
AIC	572.303
Note:	$^{*}p < 0.1$ $^{**}p < 0.05$ $^{***}p < 0.01$

Table 2.6: Mixed model linear regression (subject random intercepts and slopes), EDT beliefs

	Dependent variable: Beliefs (continuation)
Streak Length	4.463^{***}
	(0.586)
Constant	48.613^{***}
	(0.669)
Observations	15,580
Log likelihood	$-71,203.040$
AIC	142,420.100
Note:	$^{*}p < 0.1$ $^{**}p < 0.05$ $^{***}p < 0.01$

2.C Derivation of prior decoding technique

The equation that we use to derive the prior probability from the DDM parameters is based on the basic DDM, but it can be extended to the full DDM in a straightforward manner by integrating the original equation against the across trial variability parameters. In particular, we start from equation (A3) from Ratcliff and Smith, 2004, which provides the probability of hitting the upper boundary in the basic DDM. In our setting, this corresponds to the probability of making the "correct" response:

$$q^{basic}(c, a, b, M, s) = \frac{e^{\frac{-2Mc}{s^2}} - e^{\frac{-2Mb}{s^2}}}{e^{\frac{-2Ma}{s^2}} - e^{\frac{-2Mb}{s^2}}}. \tag{2.8}$$

In this expression, c represents the initial point which always lies between the lower boundary, b, and the upper boundary, a. The prior probability of choosing the correct response can be computed by assuming that the drift rate, M, tends to 0. In this case, the stimulus contains no decision-relevant information, and therefore the probability of responding correctly is a function of the prior probability and the noise alone. Calculating the limit as the drift rate approaches zero by applying L'Hopital's rule, we find that,

$$\lim_{M \to 0} q^{basic}(c, a, b, M, s) = \frac{c - b}{a - b}. \tag{2.9}$$

Using this result, we can then compute the prior probability of hitting the upper boundary for the extended DDM by integrating $q^basic(c, a, b, M, s)$ against the across trial variability parameters, and then allowing M to get arbitrarily small. In particular, the probability of crossing the upper boundary under the extended DDM is:

$$q^{advanced}(c, a, b, M, s) = \iiint q^{basic}(c, a, b, M, s) f(c) g(M) h(T) dc dM dT, \tag{2.10}$$

where f, g, and h represent the probability density functions of the initial point, drift rate, and non-decision time, respectively. To compute the prior probability, we take the limit as the drift rate goes to 0.

$$\begin{aligned}
&\lim_{M\to 0} q^{advanced}(c,a,b,M,s) = \\
&\lim_{M\to 0}\iiint q^{basic}(c,a,b,M,s)f(c)g(M)h(T)dcdMdT = \\
&\lim_{M\to 0}\iiint \frac{c-b}{a-b}f(c)g(M)h(T)dcdMdT = \\
&\lim_{M\to 0}\int \frac{c-b}{a-b}f(c)dc = \\
&\frac{c-b}{a-b}.
\end{aligned} \tag{2.11}$$

The second equality is justified by the Dominated Convergence Theorem, which allows us interchange the order of the limit and the integration. The fifth equality is based on the facts that c is uniformly distributed and the integrand $\frac{(c-b)}{(a-b)}$, is linear in c. Finally, for the case at hand, since we set the lower boundary b to zero, the prior probability of reaching the upper boundary under the extended DDM is given by $\frac{c}{a}$. Figure 2.12 provides results from simulations for three different levels of $\frac{c}{a}$, and shows that the probability of choosing the correct alternative does converge to $\frac{c}{a}$ as the drift rate approaches zero.

2.D Instructions

Thank you for participating in this experiment. For your participation you have already made $5. During the rest of the experiment you have the chance to make more money. Your final payoff for participating depends on your decisions in Parts I, II.

Part I:

In this part of the task, you will see a sequence of shapes; each element of the sequence will be displayed one at a time. There are only two possible shapes: a white circle and a white square. Your task is to accurately classify which shape is currently being presented, as quickly as possible. If you see the circle, press the right arrow button, and if you see the square, press the left arrow button. The trials will be broken up into 4 separate blocks of 300 trials; after each block, you will have a 20 second break. Please use only one hand to enter both buttons.

For every shape you correctly classify, you will be paid 1 cent. If you classify all the shapes correctly, you will make 1200*0.01=$12.00. However, for every 0.05 second it takes you to respond, you will lose 0.1 cents. (you will have a maximum of 2 seconds/trial to respond). Therefore, to make the most money possible, you

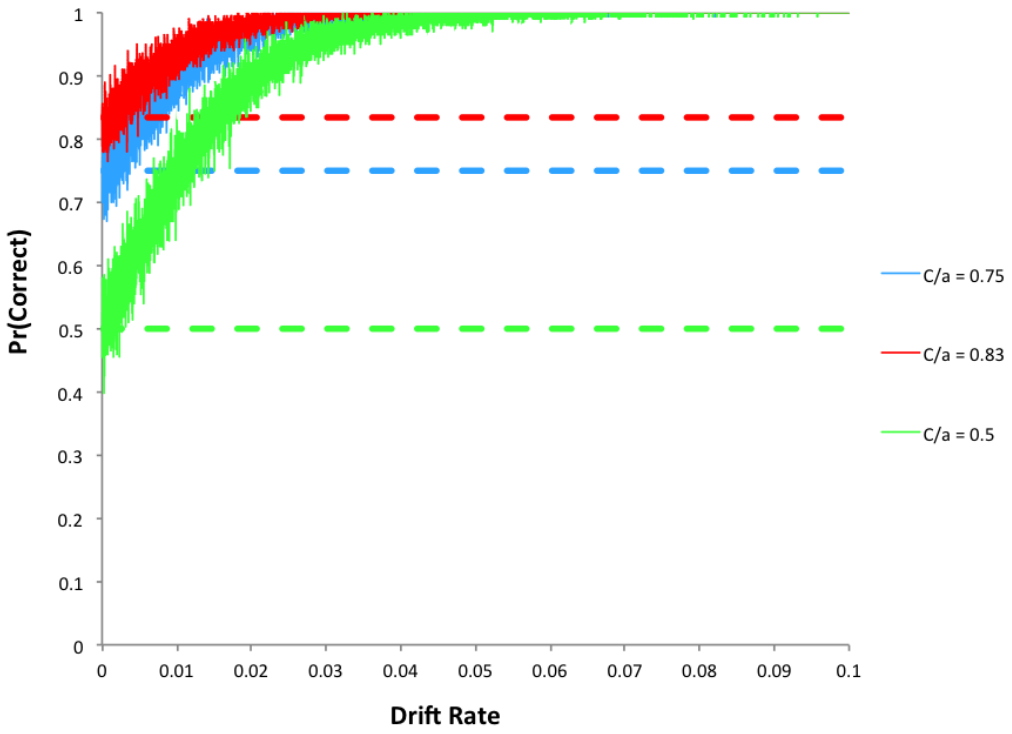


Figure 2.12: Simulations of choosing the "correct" alternative as a function of the drift rate. For a given level of within trial noise, as the drift rate tends to zero, the probability of choosing the correct alternative converges to $\frac{c}{a}$. Three different levels of $\frac{c}{a}$ are shown: $0.5, 0.75$, and 0.83.

should answer as quickly and as accurately as you can.

In each trial, the chance that you will see a circle is $\frac{1}{2}$, and the chance that you will see a square is $\frac{1}{2}$. Shapes on previous trials have no influence on the shape in the current trials; in other words the shape you see on the current trial is completely independent of all other shapes you've already seen. Before the real task starts, you will start with 5 practice trials.

Part II (instructions were given only after part I was completed)

We have studied large numbers of publicly traded companies, and constructed models of their performance patterns. Using these models, we created sequences to represent patterns of "surprises" (actual performance minus predicted performance). An upward movement indicates a "positive surprise," which results when the firm performs better than expected, and a downward movement indicates a "negative surprise" when the firm performs worse than expected.

In this task, you will see a sequence of 400 performance surprises from a typical

company, and your job is to estimate whether the next performance surprise will be positive or negative. For each of the 400 periods, you will see the performance surprises of the last 14 periods on the screen.

In each period, you will be asked to give a price at which you would be willing to buy a share of stock in this company. If you buy the stock and see a positive surprise, the stock will pay you $100. If you buy the stock and see a negative surprise, the stock will pay you $0. The important thing to understand is the following: the price you are willing to pay will, in general, not be the price you actually pay for the stock. Instead, the actual price of the stock will be drawn randomly between $0 and $100. If your willingness to pay is above this random price, you will pay the random price and receive a share of the company. If your willingness to pay is below the random price, you do not buy the share of the company. In order to make the most money under this rule, the best thing for you to do is set the price equal to probability you think there will be a positive surprise.

Examples

1. Suppose you believe that there will be a positive earnings surprise with 75% chance. You should then be willing to pay exactly $75 for this share; if the actual price is $50, then you will pay $50 for something that has a 75% chance of winning $100 which on average, will make you money. If instead the random price drawn was $90, the rule says that you will not buy this stock since $75<$90. This is good because you avoid paying $90 for something that has only a 75% chance of paying you $100.

2. Suppose you are certain (a 100% chance) that there will be positive performance surprise. Then you would be willing to pay any price between $0-$100 to buy this stock. The only way to guarantee that you buy this stock is to set your price exactly equal to $100. If you made a mistake and set the price of the stock to $90, then if the random price drawn is $92, you would not be able to buy the $92 stock, which has a 100% chance to pay $100.

3. Suppose you are certain that there will be negative performance surprise (0% chance of a positive surprise). Then you are not willing to pay any price to buy this stock. The only way to guarantee that you don't end up paying something for this stock is to set your price exactly equal to $0. If instead, you made a mistake and entered $10, then if the actual price drawn was $8, you would end up paying $8 for

a stock that has 0% chance of paying you.

In each period, you will be given $100 in experimental currency to buy a share of the stock. Since the maximum price you would ever pay for a share is $100, you will always have enough cash to buy a share of this stock, since you receive a new $100 endowment each period. Your payoff in each period will depend on the three things: your willingness to pay, the actual price, and whether there was a positive or negative surprise. To illustrate your payoffs consider the two scenarios.

If you believe there will be a positive surprise for sure, and your willingness to pay is $100, and the actual price drawn is $0, and there is actually a positive surprise, then you will end the period with $100 – $0 + $100= $200. That is, you will end the period with the $100 you started with, you don't pay any cost since the price was $0, and you earn $100 for buying the stock and having a positive earning surprise.

If you believe there will be a positive surprise with 60% chance, the actual price drawn is $30, and there is a positive surprise, then your total earnings this period will be $100-$30+$100=$170.

Your final earnings will be the sum of each of your individual period earnings, divided by 5,000. It is important to emphasize once more: the only way to maximize your final earnings is to enter your willingness to pay equal to the probability you think there will be a positive surprise.

After every 50 trials, you will see your accumulated payoff, and will be allowed to take a short break. Before the real task starts, you will start with 5 practice trials.

References

Achtziger, Anja and Carlos Alós-Ferrer (2013). "Fast or rational? A response-times study of Bayesian updating". In: *Management Science* 60.4, pp. 923–938.

Appelt, Kirstin C et al. (2011). "The Decision Making Individual Differences Inventory and guidelines for the study of individual differences in judgment and decision-making research". In: *Judgment and Decision Making* 6.3, p. 252.

Asparouhova, Elena, Michael Hertzel, and Michael Lemmon (2009). "Inference from streaks in random outcomes: Experimental evidence on beliefs in regime shifting and the law of small numbers". In: *Management Science* 55.11, pp. 1766–1782.

Barberis, Nicholas, Robin Greenwood, et al. (2015). "X-CAPM: An extrapolative capital asset pricing model". In: *Journal of Financial Economics* 115.1, pp. 1–24.

Barberis, Nicholas, Andrei Shleifer, and Robert Vishny (1998). "A model of investor sentiment". In: *Journal of financial economics* 49.3, pp. 307–343.

Becker, Gordon M, Morris H DeGroot, and Jacob Marschak (1964). "Measuring utility by a single-response sequential method". In: *Behavioral science* 9.3, pp. 226–232.

Behrens, Timothy EJ et al. (2007). "Learning the value of information in an uncertain world". In: *Nature neuroscience* 10.9, pp. 1214–1221.

Bloomfield, Robert and Jeffrey Hales (2002). "Predicting the next step of a random walk: experimental evidence of regime-shifting beliefs". In: *Journal of financial Economics* 65.3, pp. 397–414.

Bogacz, Rafal et al. (2006). "The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks." In: *Psychological review* 113.4, p. 700.

Brainard, David H (1997). "The psychophysics toolbox". In: *Spatial vision* 10, pp. 433–436.

Brav, Alon and John B Heaton (2002). "Competing theories of financial anomalies". In: *Review of Financial Studies* 15.2, pp. 575–606.

Brier, Glenn W (1950). "Verification of forecasts expressed in terms of probability". In: *Monthly weather review* 78.1, pp. 1–3.

Camerer, Colin and Teck Hua Ho (1999). "Experience-weighted attraction learning in normal form games". In: *Econometrica* 67.4, pp. 827–874.

Chater, Nick, Joshua B Tenenbaum, and Alan Yuille (2006). "Probabilistic models of cognition: Conceptual foundations". In: *Trends in cognitive sciences* 10.7, pp. 287–291.

Cho, Raymond Y et al. (2002). "Mechanisms underlying dependencies of performance on stimulus history in a two-alternative forced-choice task". In: *Cognitive, Affective, & Behavioral Neuroscience* 2.4, pp. 283–299.

Fehr, Ernst and Antonio Rangel (2011). "Neuroeconomic foundations of economic choice—recent advances". In: *The Journal of Economic Perspectives* 25.4, pp. 3–30.

Fox, Craig R and Yuval Rottenstreich (2003). "Partition priming in judgment under uncertainty". In: *Psychological Science* 14.3, pp. 195–200.

Gao, Juan et al. (2009). "Sequential effects in two-choice reaction time tasks: decomposition and synthesis of mechanisms". In: *Neural Computation* 21.9, pp. 2407–2436.

Gilovich, Thomas, Robert Vallone, and Amos Tversky (1985). "The hot hand in basketball: On the misperception of random sequences". In: *Cognitive psychology* 17.3, pp. 295–314.

Greenwood, Robin and Andrei Shleifer (2014). "Expectations of returns and expected returns". In: *Review of Financial Studies* 27.3, pp. 714–746.

Hong, Harrison and Jeremy C Stein (1999). "A unified theory of underreaction, momentum trading, and overreaction in asset markets". In: *The Journal of Finance* 54.6, pp. 2143–2184.

Huettel, Scott A, Peter B Mack, and Gregory McCarthy (2002). "Perceiving patterns in random series: dynamic processing of sequence in prefrontal cortex". In: *Nature neuroscience* 5.5, pp. 485–490.

Jones, Matt et al. (2013). "Sequential effects in response time reveal learning mechanisms and event representations." In: *Psychological review* 120.3, p. 628.

Karni, Edi (2009). "A mechanism for eliciting probabilities". In: *Econometrica* 77.2, pp. 603–606.

Klauer, Karl Christoph et al. (2007). "Process components of the Implicit Association Test: a diffusion-model analysis." In: *Journal of personality and social psychology* 93.3, p. 353.

Krajbich, Ian and Antonio Rangel (2011). "Multialternative drift-diffusion model predicts the relationship between visual fixations and choice in value-based decisions". In: *Proceedings of the National Academy of Sciences* 108.33, pp. 13852–13857.

Luce, Victor S et al. (1986). *Response Times: Their Role in Inferring Elementary Mental Organization: Their Role in Inferring Elementary Mental Organization*. Oxford University Press, USA.

Malmendier, Ulrike and Stefan Nagel (2009). *Depression babies: Do macroeconomic experiences affect risk-taking?* Tech. rep. National Bureau of Economic Research.

Malmendier, Ulrike and Stefan Nagel (2015). "Learning from inflation experiences*". In: *The Quarterly Journal of Economics*, qjv037.

Massey, Cade and George Wu (2005). "Detecting regime shifts: The causes of under-and overreaction". In: *Management Science* 51.6, pp. 932–947.

Miller, Joshua Benjamin and Adam Sanjurjo (2014). "A cold shower for the hot hand fallacy". In:

Mormann, Milica et al. (2012). "Relative visual saliency differences induce sizable bias in consumer choice". In: *Journal of Consumer Psychology* 22.1.

Mulder, Martijn J et al. (2012). "Bias in the brain: a diffusion model analysis of prior probability and potential payoff". In: *The Journal of Neuroscience* 32.7, pp. 2335–2343.

Murphy, Peter R, Joachim Vandekerckhove, and Sander Nieuwenhuis (2014). "Pupil-linked arousal determines variability in perceptual decision making". In: *PLOS Comput Biol* 10.9, e1003854.

Nassar, Matthew R et al. (2012). "Rational regulation of learning dynamics by pupil-linked arousal systems". In: *Nature neuroscience* 15.7, pp. 1040–1046.

Oaksford, Mike and Nick Chater (2009). "Precis of Bayesian rationality: The probabilistic approach to human reasoning". In: *Behavioral and Brain Sciences* 32.01, pp. 69–84.

Oskarsson, An T et al. (2009). "What's next? Judging sequences of binary events." In: *Psychological bulletin* 135.2, p. 262.

Platt, Michael L and Scott A Huettel (2008). "Risky business: the neuroeconomics of decision making under uncertainty". In: *Nature neuroscience* 11.4, pp. 398–403.

Plonsky, Ori, Kinneret Teodorescu, and Ido Erev (2015). "Reliance on small samples, the wavy recency effect, and similarity-based learning." In: *Psychological review* 122.4, p. 621.

Polania, Rafael et al. (2014). "Neural oscillations and synchronization differentially support evidence accumulation in perceptual and value-based decision making". In: *Neuron* 82.3, pp. 709–720.

Rabin, Matthew et al. (2000). *Inference by believers in the law of small numbers*. Institute of Business and Economic Research.

Ratcliff, Roger (1978). "A theory of memory retrieval." In: *Psychological review* 85.2, p. 59.

Ratcliff, Roger and Gail McKoon (2008). "The diffusion decision model: theory and data for two-choice decision tasks". In: *Neural computation* 20.4, pp. 873–922.

Ratcliff, Roger and Philip L Smith (2004). "A comparison of sequential sampling models for two-choice reaction time." In: *Psychological review* 111.2, p. 333.

Ravenzwaaij, Don van et al. (2012). “Do the dynamics of prior information depend on task context? An analysis of optimal performance and an empirical test”. In:

Roitman, Jamie D and Michael N Shadlen (2002). “Response of neurons in the lateral intraparietal area during a combined visual discrimination reaction time task”. In: *The Journal of neuroscience* 22.21, pp. 9475–9489.

Rorie, Alan E et al. (2010). “Integration of sensory and reward information during perceptual decision-making in lateral intraparietal cortex (LIP) of the macaque monkey”. In: *PloS one* 5.2, e9308.

Selten, Reinhard (1998). “Axiomatic characterization of the quadratic scoring rule”. In: *Experimental Economics* 1.1, pp. 43–62.

Soetens, Eric, Louis C Boer, and Johan E Hueting (1985). “Expectancy or automatic facilitation? Separating sequential effects in two-choice reaction time.” In: *Journal of Experimental Psychology: Human Perception and Performance* 11.5, p. 598.

Sugrue, Leo P, Greg S Corrado, and William T Newsome (2004). “Matching behavior and the representation of value in the parietal cortex”. In: *science* 304.5678, pp. 1782–1787.

Summerfield, Christopher and Konstantinos Tsetsos (2012). “Building bridges between perceptual and economic decision-making: neural and computational mechanisms”. In: *Frontiers in neuroscience* 6.70.

– (2015). “Do humans make good decisions?” In: *Trends in cognitive sciences* 19.1, pp. 27–34.

Teodorescu, Andrei R, Rani Moran, and Marius Usher (2015). “Absolutely relative or relatively absolute: violations of value invariance in human decision making”. In: *Psychonomic bulletin & review*, pp. 1–17.

Teodorescu, Andrei R and Marius Usher (2013). “Disentangling decision models: From independence to competition.” In: *Psychological Review* 120.1, p. 1.

Towal, R Blythe, Milica Mormann, and Christof Koch (2013). “Simultaneous modeling of visual saliency and value computation improves predictions of economic choice”. In: *Proceedings of the National Academy of Sciences* 110.40, E3858–E3867.

Townsend, James T and F Gregory Ashby (1985). *The Stochastic Modeling of Elementary Psychological Processes*.

Trueblood, Jennifer S et al. (2013). “Not just for consumers context effects are fundamental to decision making”. In: *Psychological science* 24.6, pp. 901–908.

Tsetsos, Konstantinos, Nick Chater, and Marius Usher (2012). “Salience driven value integration explains decision biases and preference reversal”. In: *Proceedings of the National Academy of Sciences* 109.24, pp. 9659–9664.

Usher, Marius and James L McClelland (2001). "The time course of perceptual choice: the leaky, competing accumulator model." In: *Psychological review* 108.3, p. 550.

Vandekerckhove, Joachim and Francis Tuerlinckx (2008). "Diffusion model analysis with MATLAB: A DMAT primer". In: *Behavior Research Methods* 40.1, pp. 61–72.

Webb, Ryan (2013). "Dynamic constraints on the distribution of stochastic choice: Drift Diffusion implies Random Utility". In: *unpublished, New York University*.

White, Corey N and Russell A Poldrack (2014). "Decomposing bias in different types of simple decisions." In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 40.2, p. 385.

Wilder, Matthew H et al. (2013). "The persistent impact of incidental experience". In: *Psychonomic bulletin & review* 20.6, pp. 1221–1231.

Wilder, Matthew, Matt Jones, and Michael C Mozer (2009). "Sequential effects reflect parallel learning of multiple environmental regularities". In: *Advances in neural information processing systems*, pp. 2053–2061.

Woodford, Michael (2014). "Stochastic choice: An optimizing neuroeconomic model". In: *The American Economic Review* 104.5, pp. 495–500.

Yu, Angela J and Jonathan D Cohen (2009). "Sequential effects: superstition or rational behavior?" In: *Advances in neural information processing systems*, pp. 1873–1880.

Zhang, Shunan, Crane He Huang, and J Yu Angela (2014). "Sequential effects: a Bayesian analysis of prior bias on reaction time and behavioral choice". In: *Proceedings of the 36th Annual Conference of the Cognitive Science Society (Québec City, QC:)*

Chapter 3

TESTOSTERONE IMPAIRS COGNITIVE REFLECTION IN MEN

ABSTRACT

The sex steroid testosterone regulates instinctive behaviors such as fighting and mating in non-humans. Correlational studies have linked testosterone with aggression and disorders associated with poor impulse control, but corresponding mechanisms are poorly understood and there is no evidence of causality. Building on a dual-process framework, we identified a mechanism for testosterone's behavioral effects in humans: reducing cognitive reflection. In the largest testosterone administration study to date, 243 men received either testosterone or placebo and took the Cognitive Reflection Test (CRT) that estimated their capacity to override incorrect intuitive judgments with deliberate correct responses. testosterone administration reduced CRT scores. The effect was robust to controlling for age, mood, math skills, treatment expectancy and 14 other hormones. The effects were enhanced in subjects with high cortisol and estradiol levels. Our findings suggest a unified mechanism underlying testosterone's varied behavioral effects in humans and provide novel, clear and testable predictions.

3.1 Introduction

The androgenic hormone testosterone (abbreviated "T") is produced in the male testes and in smaller quantities in female ovaries. T affects physiology, brain development, and behavior throughout life. T is released into the bloodstream and in the brain in response to external stimuli, such as the presence of an attractive mate or winning a competition, modulating physiological and cognitive processes in a context-sensitive manner (Mazur, 2005; Archer, 2006; Ronay and Hippel, 2010; Eisenegger, Haushofer, and Fehr, 2011). In many non-human species, T levels rise amid breeding season and facilitate instinctive behaviors such as intra-male fighting and mating (Edwards, 1969; Wingfield et al., 1990; Mazur, 2005; Archer, 2006). Laboratory studies have further shown that T administration causally induces aggression, mating, and behavioral disinhibition in rodents (Edwards, 1969; Wingfield et al., 1990; Bing et al., 1998; Archer, 2006).

A largely open question is how T affects human cognition and decision-making across the lifespan. T affects neurotransmitter and receptor production, as well as long- and shot-term changes in synaptic configuration, that might be involved in aging-related cognitive change (J. S. Janowsky, 2006). Studies in younger populations found correlations between endogenous T and physical aggression, sensation seeking and impulse control disorders such as drug abuse, bulimia, and borderline personality disorder (Daitzman and Zuckerman, 1980; Dabbs et al., 1995; Cotrufo et al., 2000; Martin et al., 2002; J. Janowsky, 2006; Reynolds et al., 2007; Campbell et al., 2010). Moreover, prefrontal brain regions involved in impulse control contain androgen receptors (Finley and Kritzer, 1999) and an imaging study showed that decreased prefrontal activity mediated the correlation of endogenous T with rejections of unfair ultimatum bargaining offers (Mehta and Beer, 2010), a behavior that can be interpreted as impulsive based on other behavioral studies (Grimm and Mengel, 2011). To date, all the evidence for T's influence on impulse control in humans is solely correlational. Due to the bi-directional influences between hormone levels and organisms' environment and behavior, cause and effect is conflated in correlational studies. Therefore, we test whether T causally influences impulsive cognition in humans through a randomized, placebo-controlled exogenous pharmacological manipulation.

Our study builds on the dual-process framework (Evans, 2003), according to which humans employ two types of information processing mechanisms in the course of decision-making. "System 1" (intuitive) processes occur automatically, rapidly,

and effortlessly, but might provide sub-optimal responses. "System 2" (deliberate) processes are relatively slow and computationally demanding, but more likely to produce optimal responses with greater accuracy. An important function of system 2 is monitoring system 1 responses and overriding them when needed (akin to 'checking your work' on an algebra problem). The dual system framework is not necessarily a reflection of actual algorithmic or neural implementation, and there are several different theories that might be more neurally plausible, e.g., model based vs. model free (Gläscher et al., 2010), or goal directed vs. habitual responses (Balleine and O'Doherty, 2010; Redgrave et al., 2010). Yet, it is a useful abstraction that provides sharp behavioral predictions in situations where relying on easy to compute heuristics might lead to sub-optimal outcomes.

Based on T's well-established role in instinctive behaviors in non-human animals and the correlational evidence of relation between T and impulsivity in humans, we propose that T biases human decision-making towards rapid, instinctive system 1 responses. We tested this hypothesis by randomly administering a single dose of either T or placebo to a sample of 243 males and measuring the treatment's influence on performance in the Cognitive Reflection Test (CRT, Frederick, 2005). The CRT is a widely used 3-item questionnaire that assesses one's capacity to monitor his or her intuitive judgments and override them when appropriate. CRT scores predict diverse behavioral outcomes, including preference of immediate gratification over greater delayed rewards and display of various decision-making biases such as the conjunction fallacy (Toplak, West, and Stanovich, 2011).

Here is an illustrative CRT question:

A bat and a ball cost $1.10 in total. The bat costs $1.00 more than the ball. How much does the ball cost? When faced with this question, an immediate incorrect answer (10 cents) automatically arises in most people's minds. Obtaining the correct answer (5 cents) requires actively checking the validity of the intuitive answer and engaging in deliberate, yet easy to perform arithmetic calculations (i.e., checking that the bat – ball difference is not $1 if the ball costs $.10 and the bat costs $1). We hypothesized that T administration would increase subjects' tendency to rely on their intuitive incorrect judgments, and therefore impair the CRT performance of subjects who received T relative to those who received a placebo. To rule out various confounding factors, namely T's potential influences on engagement, motivation or arithmetic skills, subjects also took part in an additional math task as a control. Subjects also provided pre- and post-treatment saliva samples that were assayed by

liquid chromatography tandem mass spectrometry (LC-MS/MS) as manipulation checks and to control for levels of other hormones that might influence the task (e.g., cortisol), Margittai et al., 2016.

3.2 Methods

Subjects

Two hundred and forty three males (mostly college students, see Table 3.2 for demographic details) were randomly administered either T (n=125) or placebo (n=118) topical gel under a double blind between-subjects protocol. Sample size was chosen to be as large as possible given the study's budget constraints, making it the largest T administration experiment conducted to date. The institutional review boards of Caltech and Claremont University approved the study, and all subjects gave informed consent.

Procedure

The timeline of our experimental procedure is illustrated in Figure 3.1. Subjects first arrived at the lab at 9:00am in the morning of their experimental session. They signed an informed consent form and proceeded to a designated room where their hands were scanned, to obtain digit ratio measurement - a possible proxy of prenatal T that was previously associated with the dependent variable (Bosch-Domenech, Branas-Garza, and Espin, 2014). Then, subjects were randomly assigned to private cubicles where they completed demographic and mood questionnaires and provided an initial baseline saliva sample. Afterwards, subjects proceeded to a designated room for T or placebo gel application. All subjects returned to the lab at 2:00pm (with no incidents of lateness), provided a second saliva sample and began the behavioral experiment at the same cubicle they were assigned in the morning session. The time frame between gel application and behavioral experiment was chosen so that tasks took place when the T group subjects experienced elevated and stable blood T levels following drug administration (Eisenegger, Eckardstein, et al., 2013).

The experiment consisted of a battery of seven behavioral tasks; none included feedback about the subjects' monetary payoffs (to avoid endogenous changes in T from changes in payoff). Only the final task included feedback regarding the subjects' performance relative to other participants (also to avoid outcome-related changes in T). The rationale for conducting a battery of tasks (compared to a single experiment) is maximizing the knowledge gained from each human subject undergoing a pharmacological manipulation, a practice which is standard (Zethraeus

et al., 2009; Kocoska-Maras et al., 2011) and looked favorably upon by Institutional Review Boards. Accordingly, we ensured that statistical tests for the CRT task alone survived correction for multiple comparisons (choosing only CRT out of the seven tasks for analysis) to avoid increased type-I error rate from multiple comparisons.

To maintain high-resolution monitoring of hormonal changes during the experiment and control for their influences, a total of four saliva samples were collected throughout the experiment (further details of collection frequency and time below). The accuracy and consistency of sampling times is crucial because the measured hormones have unique diurnal cycles which complicates comparing samples taken at different times of day. In order to standardize hormonal measurements among all subjects, we did not randomize the order of the behavioral tasks, in a similar fashion to previous studies (Zethraeus et al., 2009; Kocoska-Maras et al., 2011). The behavioral battery lasted approximately two hours. Both of the behavioral tasks reported here were computerized and occurred in the first hour of the experiment, between the second and third saliva samples. Following the experiment, subjects completed an exit survey, where they indicated their expectancies about which of the two treatments they had received, and then were privately paid in cash according to their performance.

Treatment administration

Participants were escorted in groups of 2-6 to a semi-private room where a research assistant provided a small plastic cup containing clear gel and stated it was equally likely to contain T or placebo (the cups were filled in advance by the lab manager, who did not interact with subjects and did not reveal the contents of the cup to the research assistant, so that the treatment was double-blind between assistant and subject). These cups contained either 10g of topical T 1% (2 x 50 mg packets Vogelxo® by Upsher-Smith) or volume equivalent of an inert placebo of similar texture and viscosity (80% alcogel, 20% Versagel®). We chose to administer T using topical gel, as this is the only T administration method for which the pharmacokinetics of a single dose administration (i.e., time-course of post-treatment T levels change) has been investigated in healthy young men (Eisenegger, Eckardstein, et al., 2013). The single-dose study demonstrated that plasma T levels peaked 3 hours following exogenous topical administration, and that T measurements stabilized at high levels during the time window between 4 and 7 hours following administration. Therefore we had all subjects return to the lab 4.5 hours after receiving gel, when androgen levels were higher and stable.

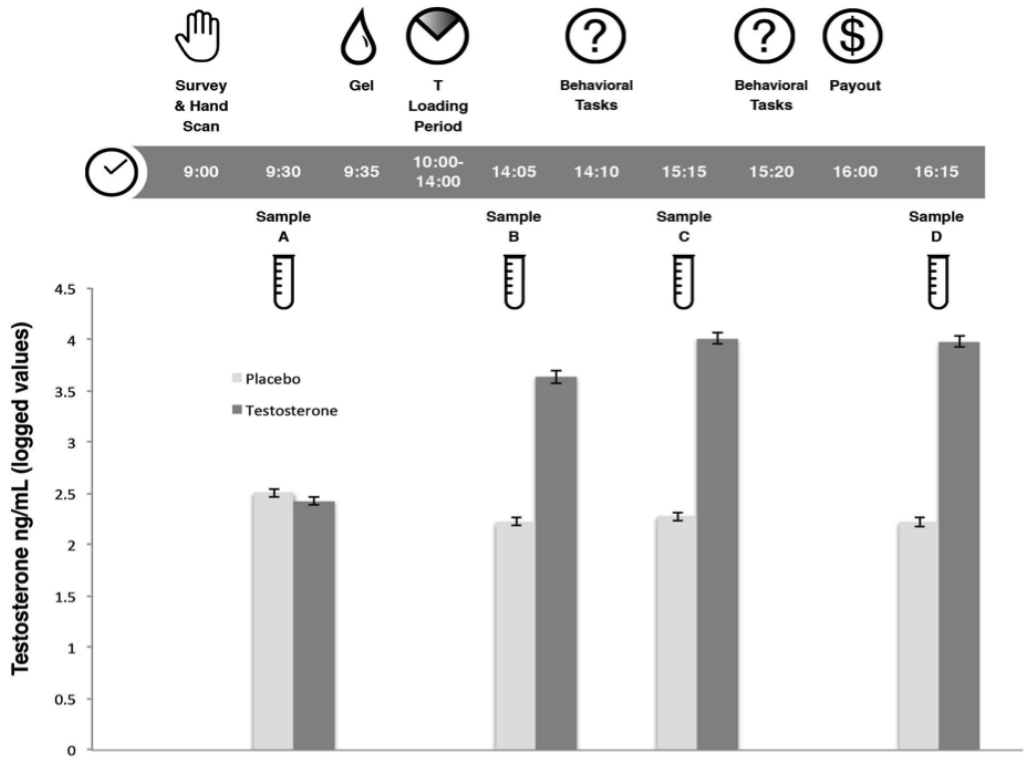


Figure 3.1: Experiment timeline and salivary testosterone levels. Subjects arrived at the lab at 9 am, had their hands scanned, filled an intake survey, and gave a baseline saliva sample "A" before application of either testosterone or placebo topical gel. After a four-hour loading period, subjects came back to the lab and took part in a battery of behavioral tasks. Three additional saliva samples ("B", "C" and "D") were collected during the experiment, all of which indicated elevated T levels in the treatment group compared to placebo. The CRT and math tasks took place between saliva sample B and C.

Subjects were instructed to remove upper body clothing and apply the entire contents of the gel container to their shoulders, upper arms, and chest as demonstrated by the research assistant. During application they were told to wait until the gel fully dried before putting clothes back on, refrain from bathing, or any activity that might cause excessive perspiration before the afternoon session, finish eating no later than 1:00pm, and return to the lab promptly at 1:55pm. After self-administering the gel under the supervision of the research assistant, participants were instructed to thoroughly wash their hands with warm water and soap, avoid touching any part of their body before thorough washing, and abstain from all skin-to-skin contact with females, as recommended by the gel manufacturers. All surfaces in the administration room were covered with medical grade isolation sheets and surfaces in the gel

application area were cleaned with alcohol swabs after each experimental session. The adjacent bathroom where the sink was located was also thoroughly wiped, as were doorknobs and handles.

Measures

Saliva sampling

Each subject provided four saliva samples at predetermined sampling times throughout the study: (1) Before treatment administration (all samples took place between 9:25 and 9:34 am) (2) upon return to the lab, just prior to starting the behavioral tasks (all samples took place between 1:55 and 2:15 pm); (3) in the middle of the behavioral tasks battery (between 3:02 and 3:38 pm) (4) a final sample following the one and only task involving performance feedback at the end of the experiment (between 4:10 and 4:44 pm). We chose to use saliva samples to avoid potential stress that might be induced by multiple blood draws throughout the experimental session. Each saliva sample was time stamped. No food or drinks were allowed into the laboratory, and the only water given to the participants was after their 3rd saliva draw (an hour before the 4th and final saliva draw).

Hormonal assays

Salivary steroids (estrone, estradiol, estriol, testosterone, androstenedione, DHEA, 5-alpha DHT, progesterone, 17OH-progesterone, 11-deoxycortisol, cortisol, cortisone, and corticosterone) were measured by LC-MS/MS using an AB Sciex Triple Quad 5500. Further details about the assay procedure are available in the appendix. A series of one-sample Kolmogorov-Smirnov tests for conformity to Gaussian (Table 3.3 in the appendix) indicated that all hormonal measurement distributions were better approximated by a Gaussian following a log-transformation, as indicated by higher p-values (i.e., the Gaussian normality hypotheses were less likely to be rejected after log-transformations). Thus, all hormonal measurements were log-transformed prior to data analysis in order to make their distributions closer to Gaussian.

Mood questionnaire

Subjects completed the PANAS-X scale (Watson and Clark, 1999), both pre-treatment (in the morning) and post-treatment (in the afternoon). Three subjects did not answer all of the negative affect items in their questionnaires, and five subjects did not complete all of the positive affect items; these subjects were excluded from

analyses that include these scales as control variables.

Digit ratio measurement

The ratio of second (index) finger length to fourth (ring) finger (abbreviated 2D:4D) is considered a proxy for pre-natal T exposure, and a previous study suggested that the measure correlates with CRT performance (Bosch-Domenech, Branas-Garza, and Espin, 2014). Subjects' 2D:4D ratios were measured by two independent raters using hand scans and digital calipers (correlation between the two raters was .95). The right hand digit ratio was not calculated for one subject due to a broken finger, and therefore he was excluded from all analyses that use the right hand digit ratio as control. Correlation between the digit ratios of the left and right hands was 0.64, p=0.0001. Regression models (tables 3.6, 3.7, 3.8) are reported using the right hand measurements. All of the results hold when replacing the right hand 2D:4D by either the left hand digit ratio or the averaged digit ratio of both hands.

Cognitive reflection test (CRT)

The CRT is designed to assess a specific cognitive function: the ability to suppress an intuitive and spontaneous ("system 1") incorrect answer in favor of a reflective and deliberative ("system 2") correct answer.

The test consists of the following three questions:

1. A bat and a ball cost $1.10 in total. The bat costs $1.00 more than the ball. How much does the ball cost?

2. If it takes 5 machines 5 minutes to make 5 widgets, how long would it take 100 machines to make 100 widgets?

3. In a lake, there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how long would it take for the patch to cover half of the lake?

Participants solved the CRT without time pressure, and were told they would be paid $1 for each correct answer and an additional bonus of $2 if they correctly solved all three questions. Thus, they could have earned as much as $5 in just a few minutes (to put this amount in perspective, the minimum wage in California, which is typical for student jobs, is $9 / hour). The CRT has never been conducted in this subject pool, mitigating the concern that subjects were previously exposed to the questions (Toplak, West, and Stanovich, 2014).

Math task

Participants completed a math task to control for their arithmetic skills, engagement levels, attention, and motivation. They had five minutes to correctly add as many sets of five two-digit numbers as possible. Subjects could use pen and paper but were not allowed to use a calculator. The two-digit numbers in each problem were randomly drawn and presented in the following way on the computer screen (participants entered their summation of the five numbers in the blank box on the right):

Table 3.1: math task question example.

21	35	48	29	83	

Once a participant submitted an answer, a new problem appeared. Participants received $1 for each correct answer and $0 for an incorrect answer.

Treatment expectancy

One previous study indicated an effect of subjects' beliefs about the treatment they had received on behavior (Eisenegger, Naef, et al., 2010). We therefore asked subjects to indicate their expectancy about whether they had received placebo or T using a 5-point scale. There were no significant differences between the groups on this expectancy measure (see Table 3.2). Two subjects did not report their treatment expectancy and therefore were excluded from all analyses in which this measure was used as a control.

3.3 Results

We observed elevated levels of T and its metabolites (e.g., dihydrotestosterone) in the saliva measurements of the T group but not in the placebo group (Figure 3.1). There were no treatment effects on either mood, treatment expectancy, or levels of all other measured hormones, ruling out these potential indirect treatment influences on the task; see appendix for further details.

We tested our hypothesis using linear regression models; full analysis detail and all models are summarized in the appendix. In line with our main hypothesis, the T group had significantly lower CRT scores compared to placebo, with 20% fewer correct answers (β=-0.43, 95% confidence interval (CI)= [-0.72 -0.16], t(241)=-3.07, p=0.002, Cohen's d: -0.42, CI = [-0.70 -0.15]; see Figure 3.2a). Moreover, incorrect intuitive answers were more common, and correct answers less common, in

the T group for each of the three CRT questions analyzed separately (see Figure 3.2c-e and appendix). Subjects who received T also gave incorrect answers more quickly, and correct answers more slowly, than subjects who received placebo (Appendix Table 3.11). These differences are consistent with T-induced bias toward system 1 intuitions and a degradation of system 2 processing speed. The negative influence of T administration on the CRT was stronger in subjects with either high cortisol or estradiol saliva levels (see Table 3.10). Previous research had suggested that these hormones moderate the behavioral influence of T (for reviews, see Liening and Josephs, 2010; Mehta, Mor, et al., 2015). More specifically, studies reported that T's influences on cognition (J. S. Janowsky, 2006) and aggression (Trainor, Kyomen, and Marler, 2006) are mediated by properly aromatized estradiol. Furthermore, the interaction between elevated endogenous T and high cortisol levels correlated with reactive aggression as a response a social provocation in females (Denson, Mehta, and Tan, 2013) and reduced earnings and increased conflict between financial and social motives in bargaining among MBA students (Mehta and Prasad, 2015).

Several factors other than reduced cognitive reflection might have lowered CRT scores following T treatment: it is possible that T affected participants' engagement, motivation or arithmetic skills. To control for these potential influences, subjects performed a separate arithmetic task of adding sequences of five two-digit numbers under time pressure (5 minutes) with the incentive of $1 for each correct answer. While arithmetic scores explained much of the between-subjects variance in CRT scores (β=0.08, CI= [0.04 0.11], t(240)=4.69, p<0.001), they were unaffected by T administration (β=0.04, CI= [-1.01 1.08], t(241)=0.07, p=0.94, Cohen's d: 0.01, CI = [-0.25 0.26]). Crucially, the effect of T on CRT scores remained highly significant after controlling for arithmetic performance, age, treatment expectancy, affective state, 2D:4D digit ratio (a potential proxy for pre-natal T exposure that has been previously associated with CRT performance (Bosch-Domenech, Branas-Garza, and Espin, 2014), and the levels of all other measurable hormones that were not affected by the pharmacological manipulation (Appendix, Table 3.6). Further analysis corroborated that CRT scores were influenced by levels of T, rather than by other metabolites that were affected by T treatment (appendix Table 3.7).

3.4 Discussion

We have demonstrated a causal effect of T on human cognition and decision-making. We now relate this effect to previous findings in the literature. First, there is extensive evidence that T increases instinctive responses with sensitivity to context.

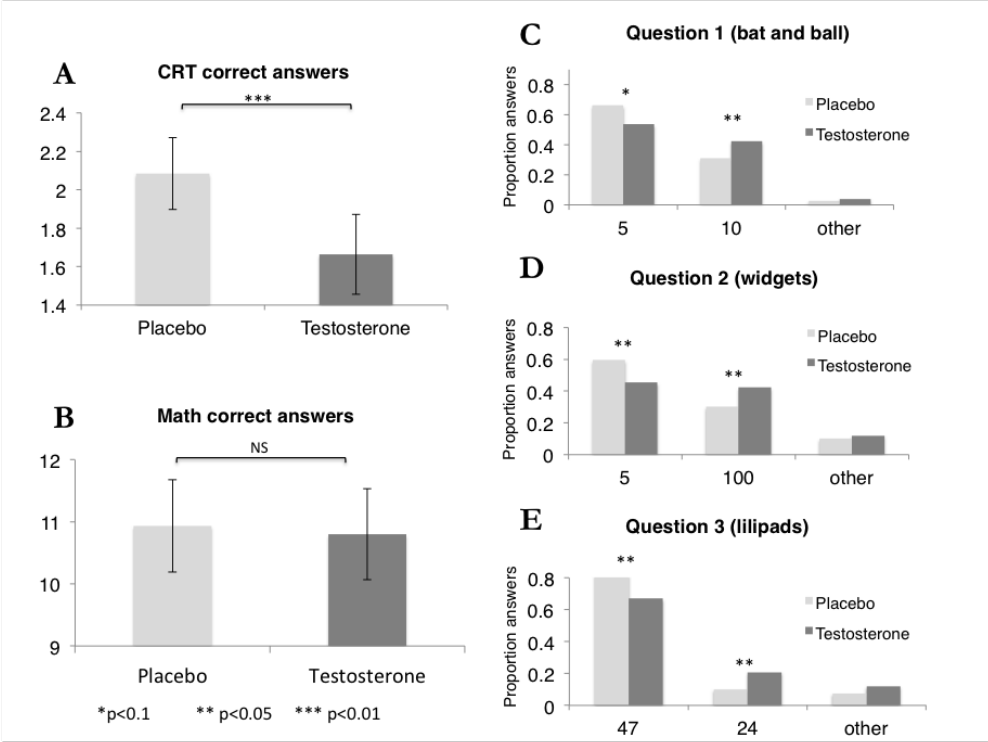


Figure 3.2: Testosterone's influence on CRT and math performance: behavioral results. (a) Mean CRT scores under placebo and testosterone treatments. (b) Mean arithmetic scores under placebo and testosterone treatment (c-e) proportions of answers given to each of the CRT questions separately. The left bar represents the correct, deliberate answer; the middle bar represents the incorrect intuitive answer; the right bar represent incorrect answers that are different from the intuitive one. Error bars denote 95% confidence intervals.

In non-human species, T levels typically rise during breeding season to facilitate instinctive behaviors such as mating and intra-male aggression (Edwards, 1969; Wingfield et al., 1990; Mazur, 2005; Archer, 2006; Eisenegger, Haushofer, and Fehr, 2011). In humans, the analogous effects are release of T and its precursors during competition, challenge, presence of an attractive mate, and in anticipation of sexual activity (Mazur, 2005; Archer, 2006; Eisenegger, Haushofer, and Fehr, 2011; Miller and Maner, 2009).

Our result fits the T-initiates-instinctive-behavior neurobiological pathway if one thinks of intuitive CRT responses as unsuppressed cognitive instincts. In this account, T's effect on cognition is an evolutionary vestige (or repurposing) that blunts more careful deliberation in favor of rough and rapid processing. A second,

more specific suggestion is that T affects cognition by inducing over-confidence, which is empirically linked to status enhancement (Von Hippel and Trivers, 2011). T is clearly associated with aggression and dominance in many non-human species (Archer, 2006), in whom aggressive behavior is typically the only way to promote hierarchical status. In humans, however, dominance or status can be established without physical aggression via displays of resources, talent, and culturally valued behaviors (Eisenegger, Naef, et al., 2010; Eisenegger, Haushofer, and Fehr, 2011). In this account, T induces a status-seeking motivational state that elevates confidence in the intuitive system 1 response, leading to faster and more frequent commission of errors.

Collateral support for this hypothesis comes from three sources. Early propositions and recent experimental studies suggest that judgmental over-confidence enhances status in long- and short-term groups (Anderson et al., 2012; Burks et al., 2013), and invoking status motives increases over-confidence (Kennedy, Anderson, and Moore, 2013). These effects can persist even when confident individuals are shown to be wrong (Anderson et al., 2012). A large-sample study with truckers showed that over-confidence about cognitive skill is higher in those who are higher in "social potency", an MMPI scale associated with dominance-seeking (Burks et al., 2013). This evidence suggests a causal pathway in which T increases status-seeking motivation, which is behaviorally implemented by acting over-confidently and reducing cognitive reflection. This proposition is further supported by a recent T administration study, showing that T made females less likely to incorporate the opinions of others in a cooperative perceptual decision-makingtask (Wright et al., 2012).

At the population level, many studies indicate that men (who have much higher T levels than women) are overconfident compared to women about the accuracy of their judgments and their relative standing on positive traits (Lundeberg, Fox, and Punćcohaŕ, 1994; Barber and Odean, 2001); these differences are even manifested in different actions by male and female CEOs (Huang and Kisgen, 2013). Thus, there appear to be population-wide correlations between T (higher in men) and overconfidence.

Most speculatively, there is evidence of correlation between T and certain kinds of financial trading performance that require rapid intuitions. One study found correlation between endogenous T and daily profit (Coates and Herbert, 2008). A second study found a correlation between prenatal T, proxied by 2D:4D digit ratios,

and profitability and longevity of high frequency traders (HFT, Coates, Gurnell, and Rustichini, 2009). HFT requires rapid processing of visuospatial information to detect temporary mispricing between markets on the scale of seconds to minutes. HFT is very likely a domain in which rapid system 1 responses are optimal. Indeed, the co-existence of systems 1 and 2 strongly suggests that system 1 responses are not always wrong or suboptimal (keeping in mind that the CRT was specifically designed to show system 1 flaws). In HFT, traders who deliberate too long will see the mispricing disappear as faster traders profitably erase it.

The hypothesis that T reduces cognitive reflection to enhance status has many testable implications. Conditions known to elevate T, such as winning contests and presence of attractive mates, should reduce cognitive reflection. In tasks where rapid intuitions are useful (e.g., HFT) increased T will boost performance, and in tasks where deliberation is needed, T will reduce performance. Finally, our study has important public health implications. Western society has experienced an exogenous T 'shock' over the past decade from a rapidly growing T replacement therapy industry, with annual sales estimated at over $2B USD in 2013 (Von Drehle, 2014). Demand for T prescriptions has also become a trend on Wall Street, where financial professionals have come to believe that high T levels turn them into "alpha males", yield greater financial gains, and increase professional status (Wallace, 2012). The possibility that T might have deleterious influences on judgments and decision-making, should be investigated further and taken into account by users, therapists, and policy makers.

APPENDIX

3.A Subjects

There were n=243 male-only participants. Most (217, 89%) were students from a southern Californian college. Non-student participants were community members from surrounding cities. n=125 of subjects were randomly assigned to receive a standard dose of T and n=118 received placebos of matched viscosity in a double blind exogenous administration paradigm.

Pre-screening criteria excluded everyone with relevant medical and psychological conditions (5α-reductase deficiency, Klinefelter's syndrome, brain tumor, cancer, psychiatric diagnosis/diagnoses, high blood pressure, liver disease, kidney disease, angina, cancer, hepatitis, renal/kidney impairment, history of epileptic seizures, and hypersensitivity to soy/ alcohol), subjects using prescription drugs that may interfere with the study (oxyphenbutazone, insulin, corticosteroids, opioids), subjects who self reported consuming illegal drugs or excessive alcohol in the last 24 hours, and non-native English speakers.

Personal, demographic, and treatment expectancy characteristics of the two treatment groups are summarized in Table 3.2 (note that 5 subjects did not report their age and were therefore excluded from all analyses in which age is used as a control variable). The right column of Table 3.2 also reports the p-value of two sample t-tests for differences between T and placebo group characteristics (a check on whether random assignment resulted in balance on all such variables). Two subjects (one from each treatment group) self reported taking T treatment on a regular basis; all analyses include these subjects and are robust to excluding them. In order to reduce the potential effect of a female experimenter's presence on T-related behaviors, male researchers conducted all of the experimental sessions.

3.B Hormonal assay procedure

Salivary steroids (estrone, estradiol, estriol, testosterone, androstenedione, DHEA, 5-α DHT, progesterone, 17OH-progesterone, 11-deoxycortisol, cortisol, cortisone, and corticosterone) were measured by LC-MS/MS using an AB Sciex Triple Quad 5500. Internal standards were added to 1 ml of saliva and the steroids then extracted by C18 column chromatography with 0.1 M NH4OH wash followed by 10% acetone. Steroids were eluted from the SPE with 10% methanol in acetone and dried under nitrogen. The dried samples were subjected to derivatization—the process of trans-

Table 3.2: Self-reported demographic data summary (standard errors in parentheses)

	All	T	Placebo	p-values for t-test of difference
N	243	118	125	
Age	23.63	24.42	22.78	0.08
	(0.46)	(0.77)	(0.49)	
Left-handed (proportion)	0.074	0.064	0.085	0.54
	(0.02)	(0.02)	(0.03)	
Heterosexual (proportion)	0.90	0.91	0.89	0.56
	(0.02)	(0.03)	(0.03)	
Treatment expectancy[1]	2.76	2.67	2.85	0.16
	(0.06)	(0.08)	(0.09)	
Married (proportion)	0.08	0.09	0.08	0.74
	(0.02)	(0.03)	(0.03)	
In a relationship (proportion)	0.38	0.34	0.42	0.20
	(0.03)	(0.05)	(0.04)	
Has children	0.06	0.08	0.04	0.23
	(0.02)	(0.02)	(0.02)	
Personal monthly income[2]	2.05	2.02	2.07	0.84
	(0.11)	(0.14)	(0.16)	

forming a compound into a derivative product of similar chemical structure—with pyridine-3-sulfonyl chloride for the estrogens (estrone (E1), estradiol (E2), and estriol (E3)) as outlined by Xi and Spink (2008). 40 μL sodium bicarbonate (50mM, pH 10) and 40 µL pyridine-3-sulfonyl chloride (3 mg/mL in acetonitrile) were added to the dried samples, and incubated at 60oC for 10 minutes. After derivatization, the samples were diluted with 80 µL of water and injected for LC-MS/MS analysis with analytical separation performed on an Agilent Poroshell 120 EC-C8 column and ionization by atmospheric pressure chemical ionization (APCI) in the positive ionization mode.

Table 3.3 lists each analyte along with its validation results for the lower limit of quantitation (LLOQ is jargon for the lowest level of detection with coefficients of variation (CVs) < 20% over the linear range), linear range, and the inter-assay precision from the highest concentration to the LLOQ within the linear range. When salivary hormone levels of participants were below their LLOQ, we assigned values halfway between zero and their respective LLOQ (note that the true quantities of the

Table 3.3: Detection levels, precision and normality tests of hormonal assays

Analyte	LLOQ	Range	Precision	Proportion undetected, pre-treatment sample A	Proportion undetected, first post-treatment sample B	K-S test p-value	K-S test (log) p-value
Estrone pg/mL	0.5	0.5 - 510	8.7 - 13.7%	0.132	0.257	<0.01	0.56
Estradiol pg/mL	0.3	0.3 - 510	4.3 - 18.7%	0.128	0.329	0.06	0.88
Testosterone pg/mL	3.0	3.0 - 5100	3.0 - 18.1%	0	0.008	$<10^{-20}$	<0.01
Androstenedione pg/mL	5.0	5.0 - 2300	5.2 - 6.6%	0	0.008	$<10^{-20}$	0.008
DHEA pg/mL	20.0	20.0 - 1800	4.1 - 15.2%	0.004	0.012	0.002	0.98
DHT pg/mL	10.0	10.0 - 920	3.6 - 17.7%	0.786	0.473	$<10^{-11}$	0.02
Progesterone pg/mL	10.0	10.0 - 10000	4.8 - 10.8%	0.794	0.753	<0.01	0.03
17OH-Progesterone pg/mL	5.0	5.0 - 630	3.9 - 13.8%	0.004	0.061	0.003	0.98
11-Deoxycortisol pg/mL	5.0	5.0 - 410	6.8 - 16.6%	0.132	0.473	<0.01	0.04
Cortisol ng/mL	0.1	0.1 - 52	5.1 - 17.9%	0	0.008	<0.01	0.92
Cortisone ng/mL	0.1	0.1 - 81	4.1 - 14.9%	0	0.008	0.07	0.59
Corticosterone pg/mL	5.0	5.0 - 1800	4.6 - 17.5%	0.313	0.312	<0.01	0.08
Aldosterone pg/mL	10.0	10.0 - 560	8.9 - 18.8%	0.272	0.272	<0.06	0.39
Melatonin pg/mL	2.5	2.5-10000	5.2 - 15.9%	0.502	0.500	0.07	0.14

hormone in the sample are never zero, even when they do not reach the detection threshold)

3.C Hormonal changes following treatment and manipulation check

As expected, there were significant post-treatment differences between groups with respect to all hormones influenced by T treatment, either as an upstream (androstenedione) or downstream (5-α DHT) metabolite of T (Horton and Tait, 1966). There was also a decrease in progesterone 170H resulting from an increase in T (which is common, according to personal communication from ZRT Laboratories chief scientist Dr. David Zava). The changes in saliva T measures were similar in magnitude to those reported in previous studies following topical gel administration of T and progesterone (e.g. Mayo et al., 2004; Du et al., 2013).

We observed no significant differences between treatment groups in hormones that were not expected to change following short-term T treatment (e.g., aldosterone, cortisol, cortisone, melatonin) in all four saliva measurements throughout the experiment (i.e., the pre-treatment and the three post-treatment measurements). The

pre-treatment and first post-treatment mean hormonal saliva levels are summarized in table 3.4; note that differences between morning and afternoon hormonal levels were affected by diurnal cycles in both treatment groups (Nomura et al., 1997; Hurwitz, Cohen, and Williams, 2004; Hucklebridge et al., 2005). From assays conducted during the first 13 (out of 17) sessions of the study, we noted that 72 out of 184 pre-treatment baseline saliva samples (in both treatment groups) presented measurements with higher T level that are expected in normal young men (greater than 400 pg/mL). All other measurements (including T metabolites) were hormonally typical. The effects of T on the CRT were robust to excluding the subjects with abnormal measurements (see below).

We traced the cause of these abnormal measurements to T gel transfer to common surfaces (e.g., door knobs, mouse pads). Crucially, the high measurements were caused by local spread of T into saliva tubes, but physiological levels were unaffected by superficial contact with the dry nuisance T gel, as (a) we observed normal pre-treatment levels of T metabolites, namely DHT and androstenedione in all subjects; (b) none of the placebo group participants showed abnormally high values of T metabolites in any of the post-treatment measurements; (c) only five out of 118 subjects from the placebo group showed consistently elevated T measurements in all of the three post-treatment saliva samples; (d) previous investigations found that interpersonal T transfer is highly unlikely even with skin-to-skin contact, (Rolf et al., 2002). Thus, we found convergent evidence that biofluid levels were unaffected by superficial contact. This conclusion was supported by ZRT Laboratories chief scientist Dr. David Zava.

In response to this finding during the course of the experimental period, we identified all surfaces and objects through which T could spread in the facility and improved sterile isolation protocol to eliminate the spread of the dried T gel. This protocol included thorough cleaning of keyboards, computer mice, chair backs, displays, and all doorknobs with a bleach-alcohol solution after each session as well as asking subjects to carefully wipe hands with a wet tissue before collecting each saliva sample. New pens were used for each session while all previously used pens were removed from the testing area. Clipboards and other miscellaneous objects that participants did or could interact with were cleaned, and an aerosol "air sanitizer" that bonds to VOCs (volatile organic compounds) was sprayed into the air. Following the adoption of this strict sterilization protocol, we found a drastic reduction in incidence of high T samples in the pre-treatment measurements, to a

total of five participants out of 58 in the following four sessions (sessions 14-17).

Finally, we conducted additional robustness checks by examining the effects of T on the CRT when (a) excluding subjects with pre-treatment saliva T of greater than 400 pg/ml from both treatment groups; (b) excluding placebo subjects with post-treatment saliva T (sample B) greater than 400 pg/ml; (c) excluding all subjects in either condition (a) or (b); and (d) repeating the analysis with a more conservative cutoff of 250 pg/ml. We found that the effect of T administration on the CRT was highly significant (all p's<0.02) regardless of the exclusion criteria used.

3.D Results

Mood questionnaire

Table 3.5 shows a modest decrease in both affect measures over time (morning vs. afternoon), and no treatment or time x treatment interaction, indicated by the output of 2-way analysis of variance (ANOVA) with an interaction term, ruling out this indirect way in which T might affect cognition and behavior. Three subjects did not answer all of the negative affect items in their questionnaires, and five subjects did not complete all of the positive affect items; these subjects were excluded from analyses that include these scales as control variables.

Cognitive reflection test

CRT scores were comparable to those previously found in equivalent samples (Brañas-Garza, Kujal, and Lenkei, 2015), although at the high end of the range. This is likely due to high analytical skill in the sampled college population (conducted in one of the top ranked schools in the US) and the use of monetary incentives (the task is typically non-incentivized). We tested our main hypothesis by estimating linear regression models with the three-item total CRT score as the dependent variable (DV). All of the analyses were conducted using the function 'lm' implemented in 'R' and the results are summarized in table 3.6. Model A1 included only treatment (testosterone=1, placebo=0) as an independent variable (IV); Model A2 also included the math task performance. Model A3 also included age, positive and negative affect (measured using the PANAS-X scale), treatment expectancy and the right hand digit ratios (the results hold when the left hand or the average between the two hands are used). Model A4 included all of the IVs of model A3 with the addition of all of the hormonal levels that were not affected by the treatment, as measured from the first post-treatment saliva sample (i.e., the second overall sample); all of the results hold when the measurements are replaced with the second post-treatment

Table 3.4: Hormone panel data measurements log(pg/mL) summary statistics (standard errors in parentheses)

	Placebo		**Testosterone**		**Two-tailed p-value from t-test of T-Placebo equality**	
Sampling time[1]	**9am**	**2pm**	**9am**	**2pm**	**9am**	**2pm**
Testosterone	**5.743**	**5.111**	**5.580**	**8.373**	**0.267**	**1.06E-13**
	(0.094)	**0.085**	**0.084**	**0.151**		
Androstenedione	**4.510**	**4.205**	**4.525**	**5.462**	**0.634**	**3.11E-09**
	(0.039)	**0.044**	**0.034**	**0.084**		
DHT	**1.984**	**1.867**	**1.905**	**3.482**	**0.745**	**2.38E-06**
	0.069	**0.051**	**0.060**	**0.114**		
Progesterone	1.937	2.002	1.829	1.883	0.36	0.41
	0.058	0.064	0.052	0.055		
Progesterone170H	**3.245**	**2.675**	**3.217**	**2.463**	**0.792**	**0.008**
	0.050	**0.058**	**0.049**	**0.058**		
Estrone	-0.088	-0.557	-0.007	-0.389	0.29	0.42
	0.063	0.066	0.064	0.056		
Estradiol	-0.743	-1.158	-0.766	-1.066	0.86	0.44
	0.052	0.059	0.054	0.054		
DHEA	5.198	4.570	5.116	4.557	0.30	0.76
	0.053	0.058	0.051	0.054		
Deoxycortisol11	2.579	1.650	2.568	1.584	0.66	0.35
	0.079	0.072	0.083	0.064		
Cortisol	1.047	0.062	1.045	0.077	0.68	0.81
	0.058	0.065	0.057	0.058		
Cortisone	2.539	1.952	2.539	2.003	0.70	0.76
	0.030	0.060	0.034	0.050		
Corticosterone	2.442	1.274	2.646	1.290	0.37	0.76
	0.126	0.065	0.123	0.060		
Aldosterone	2.640	2.516	2.634	2.395	0.82	0.14
	0.067	0.071	0.068	0.066		
Melatonin	1.045	0.276	1.221	0.353	0.27	0.23
	0.093	0.029	0.101	0.051		

Table 3.5: Positive and negative affect (PANAS-X) summary statistics

	All		Testosterone		Placebo		ANOVA: p-values		
Time	**Morning**	**Afternoon**	**Morning**	**Afternoon**	**Morning**	**Afternoon**	**T**	**Time**	**T x time**
Positive affect	2.72	2.61	2.72	2.63	2.72	2.60	0.85	0.16	0.85
	(0.05)	(0.06)	(0.06)	(0.08)	(0.07)	(0.09)			
Negative affect	1.53	1.45	1.53	1.46	1.53	1.43	0.77	0.13	0.84
	(0.04)	(0.04)	(0.06)	(0.05)	(0.05)	(0.05)			

saliva sample (i.e., the third overall measurement; see table 3.8).

In models (B1-B4), summarized in table 3.7, we repeated the analyses of models (A1)-(A4), where the binary treatment variable was replaced by the measurements of the hormones that are affected by the treatment (T, DHT, androstenedione, and progesterone 170H).

Finally, models (C1-C2) in Table 3.8 replicate the results of models (A4) and (B4) using the hormonal measurements extracted from the second post-treatment (and third overall) saliva sample.

Table 3.6: Linear regression, dependent variable: CRT score. Hormonal measurements are log transformed and taken from the first post-treatment saliva sample.

	(A1)	(A2)	(A3)	(A4)
Treatment	-0.438***	-0.441***	-0.356**	-0.350**
	(0.142)	(0.136)	(0.142)	(0.145)
Math		0.079***	0.078***	0.074***
		(0.017)	(0.017)	(0.018)
Negative affect			0.058	0.002
			(0.132)	(0.137)
Positive affect			-0.142*	-0.124
			(0.075)	(0.076)
Age			-0.030***	-0.032***
			(0.010)	(0.010)
Treatment expectancy			0.062	0.085
			(0.076)	(0.077)
Digit ratio (right)			-1.456	-1.276
			(2.029)	(2.063)
Estrone				0.028
				(0.115)
Estradiol				0.158
				(0.121)
DHEA				-0.135
				(0.151)
Progesterone				0.003
				(0.110)
Deoxycortisol11				0.114
				(0.130)
Cortisol				-0.431*
				(0.225)
Cortisone				0.339*
				(0.200)
Corticosterone				0.039
				(0.121)
Aldosterone				0.193*
				(0.098)
Melatonin				0.010
				(0.156)
Constant	2.102***	1.248***	3.425*	2.770
	(0.102)	(0.203)	(1.965)	(2.158)
Observations	243	243	229	229
R^2	0.038	0.122	0.150	0.195
Adjusted R^2	0.034	0.115	0.123	0.130
Residual Std. Error	1.109 (df = 241)	1.062 (df = 240)	1.050 (df = 221)	1.046 (df = 211)
F Statistic	9.447*** (df = 1; 241)	16.665*** (df = 2; 240)	5.582*** (df = 7; 221)	3.003*** (df = 17; 211)
Note:				*p < 0.1 **p < 0.05 *** p < 0.01

Table 3.7: Linear regression, dependent variable: CRT score. Hormonal measurements are log transformed and taken from the first post-treatment saliva sample.

	(B1)	(B2)	(B3)	(B4)
Testosterone	-0.205***	-0.202***	-0.185***	-0.198***
	(0.068)	(0.065)	(0.067)	(0.069)
Androstenedione	0.360**	0.306**	0.284*	0.230
	(0.159)	(0.153)	(0.158)	(0.172)
DHT	-0.054	-0.010	0.016	0.070
	(0.085)	(0.082)	(0.083)	(0.087)
Progesterone170H	0.009	-0.002	-0.028	-0.081
	(0.117)	(0.112)	(0.116)	(0.129)
Math		0.077***	0.076***	0.074***
		(0.017)	(0.018)	(0.018)
Negative affect			0.017	-0.039
			(0.133)	(0.137)
Positive affect			-0.149**	-0.131*
			(0.075)	(0.076)
Age			-0.028***	-0.031***
			(0.010)	(0.010)
Treatment expectancy			0.066	0.089
			(0.077)	(0.078)
Digit ratio (right)			-0.523	-0.232
			(2.042)	(2.072)
Estrone				0.043
				(0.116)
Estradiol				0.167
				(0.122)
DHEA				-0.133
				(0.152)
Progesterone				0.007
				(0.110)
Deoxycortisol11				0.105
				(0.132)
Cortisol				-0.480**
				(0.227)
Cortisone				0.421*
				(0.220)
Corticosterone				0.062
				(0.122)
Aldosterone				0.209**
				(0.098)
Melatonin				0.047
				(0.158)
Constant	1.649***	0.969**	2.312	1.662
	(0.452)	(0.458)	(2.038)	(2.193)
Observations	243	243	229	229
R^2	0.050	0.128	0.158	0.208
Adjusted R^2	0.034	0.109	0.119	0.132
Residual Std. Error	1.110 (df = 238)	1.065 (df = 237)	1.052 (df = 218)	1.045 (df = 208)
F Statistic	3.107** (df = 4; 238)	6.949*** (df = 5; 237)	4.084*** (df = 10; 218)	2.739*** (df = 20; 208)
Note:				*p < 0.1 **p < 0.05 ***p < 0.01

Table 3.8: Linear regression, dependent variable: CRT score. Hormonal measurements are log transformed and taken from the second post-treatment saliva sample.

	(C1)	(C2)
Treatment	-0.335**	
	(0.149)	
Testosterone		-0.156**
		(0.064)
Androstenedione		0.064
		(0.158)
DHT		0.097
		(0.103)
Progesterone170H		-0.232
		(0.141)
Math	0.077***	0.077***
	(0.018)	(0.018)
Negative affect	0.050	0.056
	(0.137)	(0.136)
Positive affect	-0.112	-0.117
	(0.077)	(0.077)
Age	-0.029***	-0.033***
	(0.011)	(0.011)
Treatment expectancy	0.054	0.046
	(0.078)	(0.077)
Digit ratio (right)	-1.945	-1.122
	(2.093)	(2.096)
Estrone	-0.045	-0.016
	(0.110)	(0.111)
Estradiol	0.135	0.166
	(0.125)	(0.126)
DHEA	-0.135	-0.102
	(0.132)	(0.133)
Progesterone	0.128	0.124
	(0.128)	(0.128)
Deoxycortisol11	0.207	0.295**
	(0.128)	(0.135)
Cortisol	-0.216	-0.318*
	(0.169)	(0.175)
Cortisone	0.043	0.151
	(0.182)	(0.192)
Corticosterone	0.277*	0.298**
	(0.147)	(0.147)
Aldosterone	0.064	0.048
	(0.150)	(0.150)
Melatonin	-0.015	-0.002
	(0.166)	(0.167)
Constant	3.467	3.292
	(2.209)	(2.254)
Observations	229	229
R^2	0.191	0.213
Adjusted R^2	0.126	0.137
Residual Std. Error	1.048 (df = 211)	1.042 (df = 208)
F Statistic	2.936*** (df = 17; 211)	2.813*** (df = 20; 208)
Note:		*p**p***p<0.01

Table 3.9: CRT score response frequencies and statistics by question

	Testosterone		Placebo		Logistic regression stats: (intuitive=1)		Logistic regression stats (deliberate=1)	
Question	**% intuitive**	**% deliberate**	**% intuitive**	**% deliberate**	**z**	**p-value**	**z**	**p-value**
Bat and ball	0.42	0.53	0.31	0.66	2.05	0.04	-2.69	0.009
Widgets	0.42	0.45	0.30	0.59	2.057	0.04	-2.64	0.008
Lilipads	0.20	0.67	0.10	0.82	2.24	0.02	-2.77	0.006

CRT, question level

We further examined the effect of T on each of the three CRT questions separately. For each question, we classified the responses as either (a) an intuitive incorrect answer, i.e., 10 cents in the "bat and the ball" question, 100 minutes in the "widgets" question, 24 in the "lily pads" question; (b) the reflective, correct answer, i.e., 5 cents in the "bat and the ball" question, 5 minutes in the "widgets" question, 47 in the "lily pads" question; or (c) another incorrect answer, i.e., different than in (a) or (b).

We estimated two logistic regressions for each question, one that included a binary DV that was equal "1" for incorrect intuitive answers and the other included a binary DV that was equal "1" for correct answers. The analyses revealed that the likelihood of the incorrect intuitive response was significantly greater in the T group for each one of the three questions and that the proportion of correct answers was greater in the placebo group for each of the CRT questions in isolation (see Figure 3.1 and Table 3.9). Intriguingly, both of the subjects who self-reported taking T supplements regularly (one from each group) scored 0 out of 3 in the CRT, and all of their answers were the incorrect intuitive ones. Although the latter finding suggests that long term T treatment might have larger effects on CRT performance compared to a single dose, the small number of such subjects does not allow for making inferences that can be considered more than anecdotal. Moreover, as the long-term treatment was not assigned at random, causality cannot be inferred from these two data points (e.g., it is possible that subjects with low CRT scores are more likely to use T supplements, rather than vice versa).

Dual hormone interactions

To formally test whether T's causal effects on the CRT are moderated by estradiol and cortisol, we estimated five additional linear regression models with the CRT score as the DV, summarized in table 3.10. Model (D1) included T, estradiol and T x estradiol interaction as DV. Model D2 included the same IVs as model (D1), with the addition of controls for all of the other hormonal measurements and the other controls used in the main analysis. Models (D3) and (D4) repeated the analyses of (D1) and (D2), this time including cortisol and T x cortisol interaction terms. Finally, model (D5) included both interactions in addition to all other control variables. These analyses (table 3.10) revealed that the coefficients of T, T x estradiol and T x cortisol were all negative and reliably different from zero in all of the models. These results imply that T's negative effect on the CRT was enhanced in subjects with higher levels of both estradiol and cortisol. Moreover, the model that included both interaction terms (D5) had the highest predictive power compared to all other models, demonstrated by highest value of adjusted R^2.

Table 3.10: Linear regression with dual hormone interactions. Dependent variable: CRT score. Hormonal measurements are log transformed and taken from the second post-treatment saliva sample.

	(D1)	(D2)	(D3)	(D4)	(D5)
Testosterone	-0.300***	-0.353***	-0.085**	-0.184***	-0.336**
	(0.067)	(0.086)	(0.034)	(0.068)	(0.085)
Estradiol	1.421***	1.287***		0.160	1.260***
	(0.371)	(0.406)		(0.121)	(0.402)
Testosterone x Estradiol	-0.194***	-0.168***			-0.165***
	(0.053)	(0.058)			(0.058)
Cortisol		-0.384*	0.632***	0.274	0.347
		(0.226)	(0.220)	(0.391)	(0.386)
Testosterone x Cortisol			-0.116***	-0.098**	-0.095**
			(0.033)	(0.042)	(0.041)
Androstenedione		0.191		0.202	0.164
		(0.169)		(0.170)	(0.168)
DHT		0.063		0.090	0.082
		(0.085)		(0.086)	(0.085)
Progesterone170H		-0.043		-0.062	-0.024
		(0.128)		(0.128)	(0.127)
Math		0.074***		0.075***	0.075***
		(0.018)		(0.018)	(0.017)
Negative affect		-0.051		-0.047	-0.059
		(0.135)		(0.136)	(0.134)
Positive affect		-0.124*		-0.143*	-0.136*
		(0.075)		(0.076)	(0.075)
Age		-0.026**		-0.030***	-0.030***
		(0.010)		(0.010)	(0.010)
Treatment expectancy			0.100	0.089	0.100
			(0.076)	(0.077)	(0.076)
Digit ratio (right)			-0.283	0.027	0.031
			(2.037)	(2.053)	(2.018)
Estrone			0.063	0.031	0.051
			(0.115)	(0.115)	(0.114)
DHEA			-0.174	-0.115	-0.155
			(0.150)	(0.151)	(0.149)
Progesterone			-0.053	0.011	0.048
			(0.110)	(0.109)	(0.109)
Deoxycortisol11			0.117	0.150	0.160
			(0.130)	(0.132)	(0.130)
Cortisone			0.249	0.091	-0.068
			(0.224)	(0.259)	(0.261)
Corticosterone			0.059	0.056	0.054
			(0.120)	(0.120)	(0.118)
Aldosterone			0.182*	0.204**	0.178*
			(0.097)	(0.097)	(0.096)
Melatonin			0.015	0.091	0.059
			(0.156)	(0.158)	(0.155)
Constant	4.054***	3.434	2.500***	1.882	3.613
	(0.477)	(2.241)	(0.238)	(2.172)	(2.219)
Observations	243	229	243	229	229
R^2	0.083	0.239	0.074	0.229	0.259
Adjusted R^2	0.071	0.162	0.063	0.151	0.179
Residual Std. Error	1.088 (df = 239)	1.026 (df = 207)	1.093 (df = 239)	1.033 (df = 207)	1.016 (df = 206)
F Statistic	7.195*** (df = 3; 239)	3.099*** (df = 21; 207)	6.391*** (df = 3; 239)	2.929*** (df = 21; 207)	3.267*** (df = 22; 206)

Note: *p < 0.1 **p < 0.05 ***p < 0.01. Sample sizes vary due to exclusion of subjects who did not respond on various measures.

Response times

The T group responded 6 seconds slower on average when making correct answers (T: 50.97s, placebo: 44.23s) and 7 second faster on average when providing incorrect answers (T: 51.35s, placebo: 58.46s). A Kolmogorov-Smirnov test revealed that the response times (RT) were highly non-Gaussian ($p<10\text{-}18$). Therefore the values were log-transformed for normalization purpose before statistical tests (post-transformation Kolmogorov-Smirnov test, $p=0.25$). To formally examine the treatment's effect on RT, we estimated a linear mixed model regression with log(RT) as the dependent variable (DV), and treatment (binary variable), error indicator (incorrect=1, correct=0) and the interaction between those binary treatment and error dummies as independent variables (fixed effects). Random effects of subject and question number were also included (see table 3.11). The main treatment coefficient was insignificant, implying that T subjects did not differ in their general response times relative to placebo. However, the interaction between treatment and incorrect answers was negative and significant ($p=.06$). That is, T group subjects adopted their incorrect intuitions more rapidly when providing incorrect answers in the CRT.

Table 3.11: Mixed model linear regression. Dependent variable: log(response times), with subject and question random intercept

	(E1)	(E2)
Incorrect		0.089
		(0.086)
Treatment	0.054	0.125
	(0.065)	(0.078)
Incorrect x Treatment		-0.189*
		(0.115)
Constant	3.605***	3.578***
	(0.071)	(0.076)
Observations	729	729
Log Likelihood	-776.8	-778.6
Akaike Inf. Crit.	1,563.	1,571
Bayesian Inf. Crit.	1,586	1,603.
Note:	*p < 0.1 **p < 0.05 ***p < 0.01	

References

Anderson, Cameron et al. (2012). "A status-enhancement account of overconfidence." In: *Journal of personality and social psychology* 103.4, p. 718.

Archer, John (2006). "Testosterone and human aggression: an evaluation of the challenge hypothesis". In: *Neuroscience & Biobehavioral Reviews* 30.3, pp. 319–345.

Balleine, Bernard W and John P O'Doherty (2010). "Human and rodent homologies in action control: corticostriatal determinants of goal-directed and habitual action". In: *Neuropsychopharmacology* 35.1, pp. 48–69.

Barber, Brad M and Terrance Odean (2001). "Boys will be boys: Gender, overconfidence, and common stock investment". In: *Quarterly journal of Economics*, pp. 261–292.

Bing, Ola et al. (1998). "High doses of testosterone increase anticonflict behaviour in rat". In: *European Neuropsychopharmacology* 8.4, pp. 321–323.

Bosch-Domenech, Antoni, Pablo Branas-Garza, and Antonio M Espin (2014). "Can exposure to prenatal sex hormones (2D: 4D) predict cognitive reflection?" In: *Psychoneuroendocrinology* 43, pp. 1–10.

Brañas-Garza, Pablo, Praveen Kujal, and Balint Lenkei (2015). "Cognitive Reflection Test: Whom, how, when". In:

Burks, Stephen V et al. (2013). "Overconfidence and social signalling". In: *The Review of Economic Studies* 80.3, pp. 949–983.

Campbell, Benjamin C et al. (2010). "Testosterone exposure, dopaminergic reward, and sensation-seeking in young men". In: *Physiology & behavior* 99.4, pp. 451–456.

Coates, John M, Mark Gurnell, and Aldo Rustichini (2009). "Second-to-fourth digit ratio predicts success among high-frequency financial traders". In: *Proceedings of the National Academy of Sciences* 106.2, pp. 623–628.

Coates, John M and Joe Herbert (2008). "Endogenous steroids and financial risk taking on a London trading floor". In: *Proceedings of the national academy of sciences* 105.16, pp. 6167–6172.

Cotrufo, Paolo et al. (2000). "Aggressive behavioral characteristics and endogenous hormones in women with bulimia nervosa". In: *Neuropsychobiology* 42.2, pp. 58–61.

Dabbs, James M et al. (1995). "Testosterone, crime, and misbehavior among 692 male prison inmates". In: *Personality and individual Differences* 18.5, pp. 627–633.

Daitzman, Reid and Marvin Zuckerman (1980). "Disinhibitory sensation seeking, personality and gonadal hormones". In: *Personality and Individual Differences* 1.2, pp. 103–110.

Denson, Thomas F, Pranjal H Mehta, and Daniela Ho Tan (2013). "Endogenous testosterone and cortisol jointly influence reactive aggression in women". In: *Psychoneuroendocrinology* 38.3, pp. 416–424.

Du, Joanna Y et al. (2013). "Percutaneous progesterone delivery via cream or gel application in postmenopausal women: a randomized cross-over study of progesterone levels in serum, whole blood, saliva, and capillary blood". In: *Menopause* 20.11, pp. 1169–1175.

Edwards, David A (1969). "Early androgen stimulation and aggressive behavior in male and female mice". In: *Physiology & Behavior* 4.3, pp. 333–338.

Eisenegger, Christoph, Arnold von Eckardstein, et al. (2013). "Pharmacokinetics of testosterone and estradiol gel preparations in healthy young men". In: *Psychoneuroendocrinology* 38.2, pp. 171–178.

Eisenegger, Christoph, Johannes Haushofer, and Ernst Fehr (2011). "The role of testosterone in social interaction". In: *Trends in cognitive sciences* 15.6, pp. 263–271.

Eisenegger, Christoph, Michael Naef, et al. (2010). "Prejudice and truth about the effect of testosterone on human bargaining behaviour". In: *Nature* 463.7279, pp. 356–359.

Evans, Jonathan St BT (2003). "In two minds: dual-process accounts of reasoning". In: *Trends in cognitive sciences* 7.10, pp. 454–459.

Finley, SK and MF Kritzer (1999). "Immunoreactivity for intracellular androgen receptors in identified subpopulations of neurons, astrocytes and oligodendrocytes in primate prefrontal cortex". In: *Journal of neurobiology* 40.4, pp. 446–457.

Frederick, Shane (2005). "Cognitive reflection and decision making". In: *The Journal of Economic Perspectives* 19.4, pp. 25–42.

Gläscher, Jan et al. (2010). "States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning". In: *Neuron* 66.4, pp. 585–595.

Grimm, Veronika and Friederike Mengel (2011). "Let me sleep on it: Delay reduces rejection rates in ultimatum games". In: *Economics Letters* 111.2, pp. 113–115.

Horton, R and JF Tait (1966). "Androstenedione production and interconversion rates measured in peripheral blood and studies on the possible site of its conversion to testosterone." In: *Journal of Clinical Investigation* 45.3, p. 301.

Huang, Jiekun and Darren J Kisgen (2013). "Gender and corporate finance: Are male executives overconfident relative to female executives?" In: *Journal of Financial Economics* 108.3, pp. 822–839.

Hucklebridge, Frank et al. (2005). "The diurnal patterns of the adrenal steroids cortisol and dehydroepiandrosterone (DHEA) in relation to awakening". In: *Psychoneuroendocrinology* 30.1, pp. 51–57.

Hurwitz, Shelley, Richard J Cohen, and Gordon H Williams (2004). "Diurnal variation of aldosterone and plasma renin activity: timing relation to melatonin and cortisol and consistency after prolonged bed rest". In: *Journal of Applied Physiology* 96.4, pp. 1406–1414.

Janowsky, Jeri S (2006). "Thinking with your gonads: testosterone and cognition". In: *Trends in cognitive sciences* 10.2, pp. 77–82.

Janowsky, JS (2006). "The role of androgens in cognition and brain aging in men". In: *Neuroscience* 138.3, pp. 1015–1020.

Kennedy, Jessica A, Cameron Anderson, and Don A Moore (2013). "When overconfidence is revealed to others: Testing the status-enhancement theory of overconfidence". In: *Organizational Behavior and Human Decision Processes* 122.2, pp. 266–279.

Kocoska-Maras, Ljiljana et al. (2011). "A randomized trial of the effect of testosterone and estrogen on verbal fluency, verbal memory, and spatial ability in healthy postmenopausal women". In: *Fertility and sterility* 95.1, pp. 152–157.

Liening, Scott H and Robert A Josephs (2010). "It is not just about testosterone: physiological mediators and moderators of testosterone's behavioral effects". In: *Social and Personality Psychology Compass* 4.11, pp. 982–994.

Lundeberg, Mary A, Paul W Fox, and Judith Punćcohaŕ (1994). "Highly confident but wrong: gender differences and similarities in confidence judgments." In: *Journal of educational psychology* 86.1, p. 114.

Margittai, Zsofia et al. (2016). "Exogenous cortisol causes a shift from deliberative to intuitive thinking". In: *Psychoneuroendocrinology* 64, pp. 131–135.

Martin, Catherine A et al. (2002). "Sensation seeking, puberty, and nicotine, alcohol, and marijuana use in adolescence". In: *Journal of the American academy of child & adolescent psychiatry* 41.12, pp. 1495–1502.

Mayo, A et al. (2004). "Transdermal testosterone application: pharmacokinetics and effects on pubertal status, short-term growth, and bone turnover". In: *The Journal of Clinical Endocrinology & Metabolism* 89.2, pp. 681–687.

Mazur, Allan (2005). *Biosociology of dominance and deference*. Rowman & Littlefield Publishers.

Mehta, Pranjal H and Jennifer Beer (2010). "Neural mechanisms of the testosterone–aggression relation: The role of orbitofrontal cortex". In: *Journal of Cognitive Neuroscience* 22.10, pp. 2357–2368.

Mehta, Pranjal H, Shira Mor, et al. (2015). "Dual-hormone changes are related to bargaining performance". In: *Psychological science* 26.6, pp. 866–876.

Mehta, Pranjal H and Smrithi Prasad (2015). "The dual-hormone hypothesis: a brief review and future research agenda". In: *Current opinion in behavioral sciences* 3, pp. 163–168.

Miller, Saul L and Jon K Maner (2009). “Scent of a woman men’s testosterone responses to olfactory ovulation cues”. In: *Psychological Science*.

Nomura, Shinji et al. (1997). “Circadian rhythms in plasma cortisone and cortisol and the cortisone/cortisol ratio”. In: *Clinica Chimica Acta* 266.2, pp. 83–91.

Redgrave, Peter et al. (2010). “Goal-directed and habitual control in the basal ganglia: implications for Parkinson’s disease”. In: *Nature Reviews Neuroscience* 11.11, pp. 760–772.

Reynolds, Maureen D et al. (2007). “Testosterone levels and sexual maturation predict substance use disorders in adolescent boys: A prospective study”. In: *Biological psychiatry* 61.11, pp. 1223–1227.

Rolf, C et al. (2002). “Interpersonal testosterone transfer after topical application of a newly developed testosterone gel preparation”. In: *Clinical endocrinology* 56.5, pp. 637–641.

Ronay, Richard and William von Hippel (2010). “The presence of an attractive woman elevates testosterone and physical risk taking in young men”. In: *Social Psychological and Personality Science* 1.1, pp. 57–64.

Toplak, Maggie E, Richard F West, and Keith E Stanovich (2011). “The Cognitive Reflection Test as a predictor of performance on heuristics-and-biases tasks”. In: *Memory & Cognition* 39.7, pp. 1275–1289.

– (2014). “Assessing miserly information processing: An expansion of the Cognitive Reflection Test”. In: *Thinking & Reasoning* 20.2, pp. 147–168.

Trainor, Brian C, Helen H Kyomen, and Catherine A Marler (2006). “Estrogenic encounters: how interactions between aromatase and the environment modulate aggression”. In: *Frontiers in neuroendocrinology* 27.2, pp. 170–179.

Von Drehle, D (2014). “Manopause!? Aging, insecurity and the $2 billion testosterone industry”. In: *Time, US Edition* 184, pp. 36–43.

Von Hippel, William and Robert Trivers (2011). “The evolution and psychology of self-deception”. In: *Behavioral and Brain Sciences* 34.01, pp. 1–16.

Wallace, Charles (2012). “Keep taking the testosterone”. In: *Financial Times.*

Watson, David and Lee Anna Clark (1999). “The PANAS-X: Manual for the positive and negative affect schedule-expanded form”. In:

Wingfield, John C et al. (1990). “The" challenge hypothesis": theoretical implications for patterns of testosterone secretion, mating systems, and breeding strategies”. In: *American Naturalist*, pp. 829–846.

Wright, Nicholas D et al. (2012). “Testosterone disrupts human collaboration by increasing egocentric choices”. In: *Proceedings of the Royal Society of London B: Biological Sciences*, rspb20112523.

Zethraeus, Niklas et al. (2009). “A randomized trial of the effect of estrogen and testosterone on economic behavior”. In: *Proceedings of the National Academy of Sciences* 106.16, pp. 6535–6538.

Chapter 4

UNSTRUCTURED BARGAINING WITH PRIVATE INFORMATION: THEORY AND EXPERIMENT

ABSTRACT

We study dynamic unstructured bargaining with deadlines and one-sided private information about the amount available to share (the “pie size"). Using mechanism design theory, we show that given the players’ incentives, the equilibrium incidence of bargaining failures (“strikes”) should increase with the pie size, and we derive a condition under which strikes are efficient. In our setting, no equilibrium satisfies both equality and efficiency in all pie sizes. We derive two equilibria that resolve the trade-off between equality and efficiency by either favoring equality or favoring efficiency. Using a novel experimental paradigm, we confirm that strike incidence is decreasing in the pie size. Subjects reach equal splits in small pie games (in which strikes are efficient), while most payoffs are close to either the efficient or the equal equilibrium prediction when the pie is large. We employ a machine learning approach to show that bargaining process features recorded early in the game improve out of sample prediction of disagreements at the deadline. The process feature predictions are as accurate as predictions from pie sizes only, and adding process and pie data together improves predictions even more.

4.1 Introduction

bargaining is everywhere in economic activity: from price haggling in flea markets, to wage negotiations between unions and firms, to high-stakes diplomacy. Even in competitive, large-scale markets, sequences of market trades often result from individual buyer-seller partners bargaining over a range of mutually-agreeable contract terms, knowing their outside options from the market. Bargaining failures such as holdouts and strikes - due to disputes over what each side should get - are also common and reduce welfare.

Strikes are surprising because in almost every case, the bargain that was eventually struck after a costly strike could have been agreed to much earlier in the bargaining, which would have saved lost profits, legal bills, and many other collateral costs. Then why do strikes happen? The standard approach in the game theory of private-information bargaining is that the willingness to endure a strike is the only way for one side to credibly convince the bargaining partner that their existing offer is inadequate. Although making a deal appears to be a better outcome for both sides, when players' incentives and information are taken into account strikes are not only efficient but can also be unavoidable (Kennan and R. Wilson, 1990).

Private information bargaining theories, and tests of these theories, have developed in two ways:

(1) The most popular way is bargaining theories based on highly structured settings, e.g., Ståhl (1972) or Rubinstein (1982); for a review see Ausubel, Cramton, and Deneckere (2002). "Structure" means that the rules of how bargaining proceeds are clearly specified in the theory. The rules typically define when bargaining must be completed (either a deadline or an infinite horizon), who can offer or counteroffer and at what time, when offers are accepted, whether communication is allowed (and in what form), and so on. Theoretical predictions of outcomes and payoffs depend sensitively on these structural features (see Cramton, 1984; Fudenberg, Levine, and Tirole, 1985; Rubinstein, 1985; Grossman and Perry, 1986; Gul and Sonnenschein, 1988; Ausubel and Deneckere, 1993). Following the burst of progress in game theory on structured private-information bargaining, a large experimental literature emerged testing these theories (Ochs and Roth, 1989; Camerer et al., 1993; Mitzkewitz and Nagel, 1993; Güth, Huck, and P. Ockenfels, 1996; Kagel, Kim, and Moser, 1996; Güth and Van Damme, 1998; Rapoport, Daniel, and Seale, 1998; Kagel and Wolfe, 2001; Srivastava, 2001; Croson, Boles, and Murnighan, 2003; Johnson et al., 2002; Kriss, Nagel, and Weber, 2013).

The clear assumptions about structure in the theory made experimental design and theory-testing straightforward.

(2) The less popular way of theorizing and experimentation in economics is based on unstructured bargaining. Our paper returns to this less popular route, exploring unstructured bargaining with one-sided private information in an experiment.

There are three good reasons to study unstructured bargaining.

First, most natural two-player bargaining is *not* highly structured. Conventional methods for conducting bargaining do emerge in natural settings, but these methods are rarely constrained, because there are no penalties for deviating from conventions. Studying unstructured bargaining is of particular importance, as strategic behavior may substantially differ between static and dynamic environments that allow continuous-time interaction (Friedman and Oprea, 2012). There may also be clear empirical regularities in unstructured bargaining– such as deadline effects (Roth, Murnighan, and Schoumaker, 1988; Gächter and Riedl, 2005) – that are evident in the data but not predicted by theory. Establishing these regularities can *lead* theorizing, rather than test theory.

Second, unstructured bargaining creates a large amount of interesting data during the bargaining process. Players can make offers at any time, retract offers, and so on. Of course, theories can gain precision by ignoring these process data. However, if process variables are systematically associated with outcomes, these empirical regularities both challenge simple equilibrium theories and invite new theory development. Indeed, we use process data in a new way: To predict which bargaining trials will results in deals and strikes. We use a penalized regression approach from machine learning, to select those features from a large number of process features and make out of sample, cross-validated predictions (guarding against overfitting). The process features can predict strikes about as accurately as the pie sizes can; adding both process and pie size together makes even better predictions.

Process data are also useful because practical negotiation advice often consists of simple heuristics about how to bargain well (Pruitt, 2013). For example, negotiation researchers have long ago postulated that initial offers might serve as bargaining anchors and that various psychological manipulations, such as perspective taking, could potentially bias bargaining outcomes (Kristensen and Gärling, 1997; Galinsky and Mussweiler, 2001; Van Poucke and Buelens, 2002; Mason et al., 2013; Ames

and Mason, 2015). Advice like this can be easily tested by carefully controlled experimental designs that allows structure-free bargaining while keeping the process fully tractable, such as our paradigm.

Third, even when bargaining is unstructured, theory can still be applied to make clear interesting predictions. A natural intuition is that when bargaining methods are unstructured, no clear predictions can be made, as if the lack of structure in the bargaining protocol must imply a lack of structure (or precision) in predictions. This intuition is just not right. In the case we study, clear predictions about unstructured bargaining do emerge, thanks to the wonderful "revelation principle" (Myerson, 1979; Myerson, 1984). This principle has the useful property of implying empirical predictions for noncooperative equilibria, independently of the bargaining protocol, based purely on the information structure. For example, the application of the revelation principle in our setting leads to the prediction that strikes will become less common as the amount of surplus the players are bargaining over grows. This type of prediction is non-obvious and can be easily tested. Furthermore, if additional assumptions are made about equilibrium offers, and combined with the revelation principle, then exact numerical predictions about offers and strike rates can be made. That is, even if the bargaining protocol lacks structure, predictions can have plenty of restricted "structure" thanks to the beautiful game theory.

4.2 Background

The experimental literature on bargaining is vast, so below we only focus studies closely related to ours.[1] Before theoretical breakthroughs in understanding structured bargaining, most experiments used unstructured communication. The main focus of interest was process-free solution concepts such as the Nash bargaining solution (Nash Jr, 1950), and important extensions (e.g. Kalai and Smorodinsky, 1975). We will refer to the amount of surplus available to share as the "pie". Many bargains (Nydegger and Owen, 1974; Roth and Michael W Malouf, 1979) led to an equal split of the pie. Roth suggested that "bargainers sought to identify initial bargaining positions that had some special reasons for being credible... that served as *focal points* that then influenced the subsequent conduct of negotiation" (Roth, 1985). Under informational asymmetries, disagreements may arise due to coordination difficulties. Several papers by Roth and colleagues then explored what happens when players bargain over points which have different financial value to players

[1] For reviews, see Kennan and R. B. Wilson, 1993; Ausubel, Cramton, and Deneckere, 2002; Thompson, J. Wang, and Gunia, 2010

(Roth and Michael W Malouf, 1979; Roth, Michael WK Malouf, and Murnighan, 1981; Roth and Murnighan, 1982; Roth, 1985). In theory, there should be no disagreements in these games but a modest percentage of trials (10-20%) did result in disagreement. Many of the disagreements could be traced to self-serving differences between which of two focal points should be adopted– whether to allocate points equally, or to allocate the money, resulting from points, equally. Focal points have remained an important theme in more recent work (Schelling, 1960; Roth, 1985; Kristensen and Gärling, 1997; Janssen, 2001; Binmore and Samuelson, 2006; Janssen, 2006; Bardsley et al., 2010; Isoni et al., 2013a; Isoni et al., 2013b; Hargreaves Heap, Rojo Arjona, and Sugden, 2014). Roth, Murnighan, and Schoumaker, 1988, also drew attention to the fact that the large majority of agreements are made just before a (known) deadline, an observation called the "deadline effect."

Several experiments have observed what happens in unstructured bargaining with *two-sided* private information (K. Valley et al., 2002). The typical finding is that in face-to-face and unstructured communication via message-passing, there are *fewer* disagreements than predicted by theory.[2] However, when players bargaining can only make a single offer, disagreements are more common, and the key predictions of theory hold surprisingly well (Radner and Schotter, 1989; Rapoport, Erev, and Zwick, 1995; Rapoport and Fuller, 1995; Daniel, Seale, and Rapoport, 1998).

The closest precursor to our design is Forsythe, Kennan, and Sopher (henceforth FKS), who studied unstructured bargaining with one-sided private information about the sizes of two possible pies (Forsythe, Kennan, and Sopher, 1991). They used mechanism design to identify properties shared by all Bayesian equilibria of any bargaining game, using the revelation principle (Myerson, 1979; Myerson, 1984). This approach gives a "strike condition" predicting when disagreements would be ex-ante efficient. They tested their theory by conducting several experimental treatments, with free-form communication. The results qualitatively match the theory. We generalize their earlier model to capture any finite number of pie sizes. Because there are several different pie sizes, equilibria which maximize efficiency or equality create different predictions, which we test. Our experimental design uses 6 pie sizes with rapid bargaining (10 seconds per trial), where bargaining occurs only through visible offers and counter-offers, with no other restrictions. They also did not analyze their process data at all, whereas we use machine learning analysis

[2] A comparable finding in sender-receiver games is that senders willingly share more private information than is selfishly rational; see Crawford, 2003; Cai and J. T.-Y. Wang, 2006; J. T.-y. Wang, Spezio, and Camerer, 2010.

of the process features to predict strikes on a trial-by-trial basis.

From the literature studying structured bargaining, Mitzkewitz and Nagel, 1993 (henceforth MN) is a closely related design. They study ultimatum bargaining with incomplete information. MN use the same distribution over pie sizes in ultimatum bargaining that we employ in unstructured bargaining. The pattern of payoffs and disagreements in our results is similar to that of MN's "offer" game, in which the informed player makes an ultimatum proposal. Our results generalize their conclusion that fairness and equality concerns matter in asymmetric information ultimatum bargaining to a less structured environment.

Another branch of literature that is related to our study is the experimental work investigating how humans resolve tradeoffs between equality and efficiency. While this question is still under a (heated) debate (Kritikos and Bolle, 2001; Charness and Rabin, 2002; Engelmann and Strobel, 2004; Engelmann and Strobel, 2006; Fehr, Naef, and Schmidt, 2006; Bolton and A. Ockenfels, 2006; El Harbi et al., 2015), it is largely accepted that people are heterogeneous with respect to how they prioritize these factors.[3]

A few recent papers have investigated highly structured strategic interactions (De Bruyn and Bolton, 2008; Blanco, Engelmann, and Normann, 2011; López-Pérez, Pintér, and Kiss, 2013; Jacquemet and Zylbersztejn, 2014), and some have examined free form bargaining with full information (Herreiner and Puppe, 2004; Galeotti, Montero, and Poulsen, 2015). We extend this literature by deriving theoretical predictions and test empirically how humans resolve the equality-efficiency tradeoff in a dynamic strategic environment with informational asymmetry.

Finally, our study closely relates to negotiation research (Pruitt, 2013), a branch of social psychology and organizational behavior research. In contrast to economic theories that typically describe behavior in equilibrium (i.e., when players best respond to each other's actions), negotiation theories assume that bargainers are not in equilibrium and focus on prescriptive models, in which adopting certain strategies improves negotiation outcomes. Negotiation researchers take into account the process of bargaining by studying psychological constructs such as aspirations, defined as "the highest valued outcome at which the negotiator places a non-negligible likelihood that that value would be accepted by the other party" (White and Neale, 1994).

[3]For example, economics students are inclined to favor efficiency over equality, females are more egalitarian than males, and political preferences do not seem to have an effect (Engelmann and Strobel, 2004; Fehr, Naef, and Schmidt, 2006).

Aspirations play an important role in determining the bargainers' initial offers, and were shown to influence bargaining outcome variables such as disagreement rates and surplus division (Yukl, 1974; White and Neale, 1994; White, K. L. Valley, et al., 1994; Kristensen and Gärling, 1997; Galinsky and Mussweiler, 2001; Van Poucke and Buelens, 2002; Buelens and Van Poucke, 2004; Mason et al., 2013; Ames and Mason, 2015).

The reminder of this paper is organized as follows. In section 4.3, we use mechanism design theory to derive general qualitative properties of bargaining in equilibrium. We show that in our setting, no equilibrium satisfies both equality and efficiency in all sates of the world, and propose two equilibria that solve this tradeoff by either favoring the former or the latter. We present a novel experimental design in section 4.4, and summarize its general results in section 4.5. We use machine learning to examine how bargaining process data can be associated with bargaining outcome variables in section 4.6, and conclude in section 4.7.

4.3 Theory

In this section we develop a theory that provides testable predictions of disagreement rates and surplus division. Our model combines two methods to analyze bargaining: mechanism design and focal points. We extend the model of strikes developed in Kennan, 1986 and Forsythe, Kennan, and Sopher, 1991 to an arbitrary finite number of states. This extension yields non-obvious predictions of the frequency of disagreement (the strike rate) in each state, using only the game structure, rationality, and incentive-compatibility constraints. Assuming ex-ante efficiency allows further predictions. We then suggest a focal point approach to the problem of equilibrium selection. Combining these two approaches yields testable predictions about both strike rates and payoffs in each state.

Game and notation

Two players must agree on how to split a surplus (or "pie"), a random variable denoted by π. The informed player knows the actual size of the pie. The uninformed player knows that the informed player knows the pie size. States of the world are indexed by $k \in \{1, 2, \ldots, K\}$, and the pie size in state k is π_k. Without loss of generality, we assume $\pi_k > \pi_j$ when $k > j$. The probability distribution of pie sizes $\Pr(\pi_k) = p_k$ is commonly known. The players have a finite amount of time T, which is commonly known, to reach an agreement. They bargain over the payoff of the uninformed player, denoted by w, by continuously communicating their bids.

Players cannot commit to a particular bargaining position. In case of agreement on an uniformed player's payoff w, the informed player gets $y = \pi - w$. If no deal is made by time T, both players' payoffs are zero.

The direct bargaining mechanism

By the revelation principle (Myerson, 1979; Myerson, 1984), for any Nash equilibrium in the bargaining game, there exists a payoff-equivalent equilibrium of a simplified game ("a direct mechanism") in which the informed player truthfully reveals the pie size to a neutral "mediator" who determines the payoffs and the probability of a strike based on that report (Forsythe, Kennan, and Sopher, 1991). Following FKS, we assume that bargainers negotiate inscrutably over the set of direct mechanisms of the following type.

In the direct mechanism, the informed player announces the true size of the pie, π_k. The pie is then decreased by a known fraction, $1 - \gamma_k$, which can be interpreted as the strike probability in state k, leaving an expected pie size of $\gamma_k \pi_k$. We refer to γ_k as the deal probability and $1 - \gamma_k$ as the strike probability. The uninformed bargainer receives x_k, and the informed player gets the rest of the pie, $\gamma_k \pi_k - x_k$. To make predictions regarding observed behavior, we rely on the fact that the payoff x_k in the direct mechanism is tantamount to the expected payoff of the uninformed player in state k of the bargaining game: $x_k = \gamma_k w_k$ such that w_k is the uninformed payoff conditional upon a deal in state k. A mechanism therefore involves $2K$ parameters, $\{\gamma_k, x_k\}_{k=1}^{K}$.

Individual rationality (IR)

Individual rationality requires that both players prefer to participate in the mechanism. Therefore, the IR requirement is that for all k

$$\gamma_k \pi_k - x_k \geq 0 \tag{4.1}$$

$$x_k \geq 0. \tag{4.2}$$

Incentive compatibility (IC)

A mechanism is IC if it is optimal for the informed player to tell the truth, i.e., her expected payoff is (weakly) maximized when she announces the true size of the pie. This requires

$$\gamma_k \pi_k - x_k \geq \gamma_j \pi_k - x_j \text{ for all } k, \text{ for all } j \neq k. \tag{4.3}$$

The IR and IC conditions together lead to the following result.

Lemma 1. *If the bargaining mechanism satisfies IR and IC:*

1. *Deal rates are monotonically increasing in the pie size π_k.*
2. *The uninformed player's payoffs are monotonically increasing in the pie size.*
3. *The uninformed player's payoff is identical for all states in which the deal probability is 1.*

Proof: See the Appendix, section 4.A

Efficiency

In our setting a mechanism is efficient (more precisely, is "interim-incentive efficient", Holmström and Myerson (1983)) if it is Pareto optimal for the set of $K + 1$ agents: the K informed players in each of the different states k, and the uninformed player.

Lemma 2. *The strike condition: For IR and IC mechanisms, strikes in state k are ex-ante efficient if*

$$\frac{\pi_k}{\pi_{k+1}} < \frac{(1 - \sum_{j=1}^{k} p_j)}{(1 - \sum_{j=1}^{k-1} p_j)} = \frac{\Pr(\pi \geq \pi_{k+1})}{\Pr(\pi \geq \pi_k)}. \tag{4.4}$$

Proof: See section 4.A of the Appendix.

The relations between pie size ratios and conditional probabilities of pie size in Eq. 4.4 are called "strike conditions".

By Lemma 1 (result 1), if there exists a cutoff state, π_c, in which $\gamma_c = 1$ (no strikes), then strikes are inefficient in all states π_k such that $k \geq c$. Furthermore, as the uninformed player's payoff must be the same in all states where no disagreements occur (result 3), this implies that if strikes are inefficient in more than a single state, there exists no equilibrium where both efficiency and payoff equality hold for all states. Thus, there is a built-in tension between efficiency and equality under some informational settings.

Equilibrium selection using focal points

In theory, the IR and IC constraints limit the scope of possible bargaining outcomes and predict when strikes are likely to occur. This is remarkable considering that the bargaining protocol is unstructured. However, these conditions do not precisely pin down the numerical strike rates $1 - \gamma_k$ and the equilibrium payoffs (conditional on a deal being reached) w_k for each state. There are many such sets of parameter values that will satisfy IR and IC, and are equilibrium outcomes.

To make a more precise prediction, we incorporate an equilibrium selection approach that relies on the extensive literature emphasizing the importance of focal points in bargaining games (Schelling, 1960; Roth, 1985; Kristensen and Gärling, 1997; Janssen, 2001; Binmore and Samuelson, 2006; Janssen, 2006; Bardsley et al., 2010; Isoni et al., 2013a; Isoni et al., 2013b).

Absent other salient features of bargaining, the natural focal point is an equal split (i.e., $w_k = \pi_k/2$). Indeed, equal splits often emerge in bargaining experiments (e.g. Roth and Michael W Malouf, 1979; Roth and Murnighan, 1982).[4] Note that equal sharing is also common in sharecropping contracts (Young and Burke, 2001), corporate budget allocations to divisions (Bardolet, Fox, and Lovallo, 2011) and bequests to heirs (Menchik, 1980; Behrman and Rosenzweig, 2004). Regardless of the source of equal sharing, here we simply use this regularity as a basis for generating *precise* numerical predictions of the strike rates.

In practice, we propose that the equilibrium payoff of the uninformed player, conditional on a deal, will equal half of the pie size ($w_k = \pi_k/2$) as long as an equal split satisfies the IR and IC conditions (Lemma 1). We use this premise to calibrate our model and derive two competing predictions that resolve the tension between efficiency and equality (discussed in section 4.3) by either prioritizing the former or the latter.

The efficient equilibrium

To prioritize efficiency over equality, we set the deal rate to 1 whenever the strike condition (lemma 2) does not hold (i.e., whenever strikes are inefficient). Then, we split the pie equally given this constraint. Suppose that strikes are inefficient for all pies that are greater than π_c. As discussed above, this implies that the uninformed

[4]Many possible explanations have been proposed to the prevalence of equal-splitting, including social norms (Andreoni and Bernheim, 2007), pure dislike of unequal distributions (Fehr and Schmidt, 1999) or beliefs about the preferences of one's partner (Chmura et al., 2005).

player's payoff must be the same for all $\pi_k \geq \pi_c$ (lemma 1, result 3). In order to yield a clear prediction about the equilibrium uninformed payoffs w_k^*, we divide the pie equally in lower-value pie states given this constraint:

$$w_k^* = \begin{cases} \frac{\pi_k}{2} & \forall \pi_k \leq \pi_c \\ \frac{\pi_c}{2} & \forall \pi_k > \pi_c. \end{cases} \tag{4.5}$$

In our experiment, π takes on values which are the integer dollar amounts between \$1-6 with equal likelihood. It follows numerically that the strike condition (lemma 2) holds for pies of size 1 and 2. When $\pi = 3$, the two sides of the inequality are equal so the strike rate is indeterminate. When $\pi \geq 4$ there should be no strikes. Combining this constraint with the focal principle of equal splitting implies that an equal split of $\pi = 4$ (i.e., the uninformed player's payoff is 2) can be an equilibrium, but then the same amount (\$2) must also be the equilibrium payoff of the uninformed player for the larger pie sizes 5 and 6.

The efficiency constraint (Eq. 4.4) and the use of focal payoffs (Eq. 4.5) enable us to pin down the exact numerical strike rates for all pie sizes. We set $\gamma_4 = \gamma_5 = \gamma_6 = 1$, as required by the strike condition when pie sizes are uniformly distributed over $\{\$1, 2, 3, 4, 5, 6\}$. Noting again that the uninformed player's payoff in each state x_k in the direct bargaining mechanism is equal to the payoff in case of a deal times the strike rate, we fix $x_k = \gamma_k(0.5\pi_k)$ for all $k < 4$, and $x_k = 2$ for all $k \geq 4$. Consequently, we can use the IC condition (Eq. 4.3) to make explicit predictions of the strike rates in the efficient equilibrium:

$$\begin{cases} \gamma_j \leq \frac{0.5\pi_k}{\pi_k - 0.5\pi_j}\gamma_k & \forall\ k \leq 4,\ j \neq k \\ \gamma_k = 1 & \forall k \geq 4. \end{cases} \tag{4.6}$$

Solving this set of inequalities numerically (see Section 4.A of the Appendix) and picking the highest possible values of γ_k (for maximal efficiency) yields the prediction of

$$[\gamma_1,\ \gamma_2,\ \gamma_3,\ \gamma_4,\ \gamma_5,\ \gamma_6] = [0.4,\ 0.6,\ 0.8,\ 1,\ 1,\ 1]. \tag{4.7}$$

The equal split equilibrium

As discussed in section 4.3, some efficiency must be sacrificed in order to achieve equality for every pie size. In the equal split equilibrium, we first impose equal splits and only then maximize efficiency given this constraint:

$$w_k^* = \frac{\pi_k}{2}. \tag{4.8}$$

As the deal rates are increasing with the pie size (Lemma 1.1), and as the uninformed payoff must be identical in all states where there are no strikes (Lemma 1.3), full equality implies that efficiency (i.e., no strikes) can only be achieved in the largest pie (for formal proof, see section 4.A of the appendix). Thus, to pin down exact numerical predictions of deal rates in the equal equilibrium, we set $\gamma_6 = 1$. Then, we use the IC inequalities (Eq. 4.3) to make explicit predictions of the strike rates:

$$\begin{cases} \gamma_j \leq \frac{0.5\pi_k}{\pi_k - 0.5\pi_j}\gamma_k & \forall\ j \neq k \\ \gamma_k = 1 & k = 6. \end{cases} \tag{4.9}$$

Solving this set of inequalities numerically and picking the highest possible values of γ_k (for maximal efficiency) yields the prediction of

$$[\gamma_1,\ \gamma_2,\ \gamma_3,\ \gamma_4,\ \gamma_5,\ \gamma_6] = [0.3583,\ 0.5250,\ 0.6917,\ 0.8167,\ 0.9167,\ 1]. \tag{4.10}$$

4.4 Experiment

In this section, we present a novel experimental paradigm of dynamic bargaining, which allows both parties to communicate offers whenever they please, while keeping their behavior tractable.

Design

Our design is a continuous-time bargaining game with one-sided private information. At the start of each session, participants were randomly divided into two equally-sized type groups, informed and uninformed. The types were fixed for the session's 120 bargaining periods. Each period had the following steps:

1. Each player was randomly matched with a partner from the other group in a stranger protocol (to prevent sequential effects such as reputation building).

2. In each game, an integer pie size, $\pi \in \{\$1, 2, 3, 4, 5, 6\}$, was drawn from a commonly known discrete uniform distribution:

$$\Pr(\pi_k) = \frac{1}{6} \quad \forall \pi \in \{\$1, 2, 3, 4, 5, 6\}.$$

3. The informed player was told the true value of π for that period.

4. Each pair bargained over the uninformed player's payoff, denoted by w. Players communicated their monetary offers, in multiples of \$0.2, using mouse clicks on a graphical interface that was designed for this purpose by z-tree software (Fischbacher, 2007)[5] (see Figure 4.1). The offer values were between \$0 and \$6.

5. During the first two seconds of bargaining, both players fixed their initial offers, without seeing the offers of their partner (see Figure 4.1a).

6. Once the initial offers were set, players bargained continuously for 10 seconds using mouse clicks (see Figure 4.1b).

7. When players' positions matched each other, visual feedback was given to both of them in the form of a vertical stripe connecting their offer lines (see Figure 1c). If none of the players changed their position for the next 1.5 seconds following the offer-match feedback, a deal was made. Thus, in order to make a deal, the latest time in which players' bids could match was $t = 8.5$ seconds.

8. If no deal had been made within 10 seconds of bargaining, both players' payoffs from that period were \$0.

9. After each game, both players were told their payoffs and the actual pie size (see Figure 4.1d).

Methods

We conducted eight experiment sessions, five at the Caltech SSEL and three at the UCLA CASSEL labs. There were a total of N=110 subjects (mean age: 21.3 SD:

[5] A video demonstration of the task is available on `https://www.youtube.com/watch?v=y7pKh1EJsvM&`.

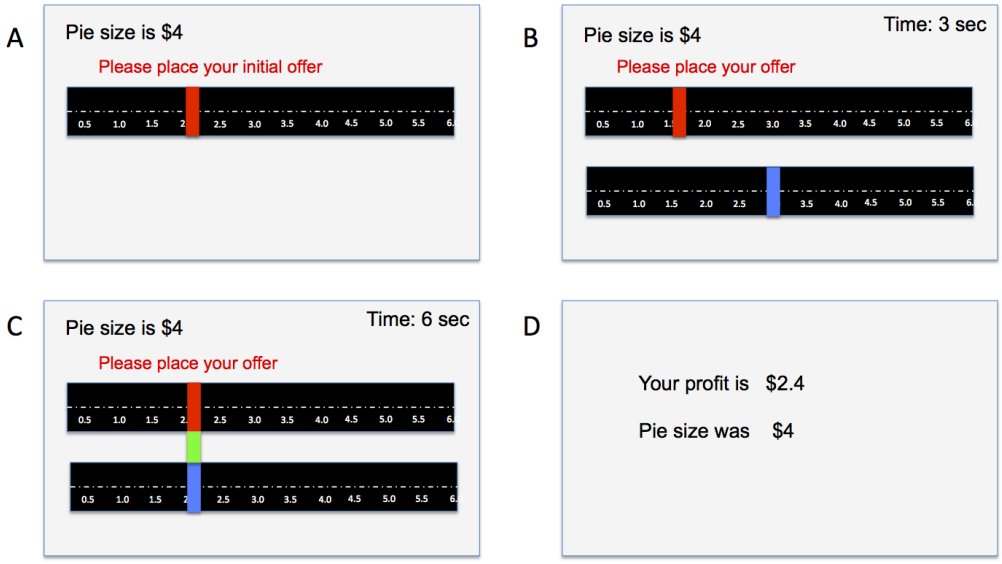


Figure 4.1: Bargaining interface. (a) initial offer screen: in the first two seconds of bargaining, players set their initial position, oblivious to the initial position of their partner. The pie size at the top left corner appears only for the informed type. (b) Players communicate their offers using mouse click on the interface. (c) When demands match, feedback in the form of a green vertical stripe appears on the screen. If no changes are made in the following 1.5 seconds, a deal is made. (d) Following the game, both players are notified regarding their payoffs and the pie size.

2.4; 47 females). The number of subjects varied slightly across sessions due to show-up differences (see Appendix 4.B for details)[6]. In the beginning of each session, subjects were randomly assigned to isolated computer workstations and were handed printed versions of the instructions (see Appendix 4.D). The instructions were also read aloud by the experimenter (who was the same person in all sessions). All of the participants completed a short quiz to check their understanding of the task. Subjects played 15 practice rounds in order to become familiar with the game and the interactive interface before the actual play of 120 periods. Participants' payoffs were based on their profits in randomly chosen 15% of the periods, plus a show-up fee of $5. Each session lasted approximately 90 minutes.

[6]There is a negative correlation ($r = -0.49$) between session size and overall deal rate, which is largely due to smaller high-deal rate sessions being conducted at Caltech (controlling for location reduces the correlation to -0.09). The difference between (regression-predicted) average deal rates in the smallest and largest session sizes is also not large in magnitude, dropping from 65% to 58%.

Table 4.1: Average payoffs* and deal rates by pie size**, standard errors in parantheses

Pie size	1	2	3	4	5	6	Mean
Informed payoff	0.37	0.95	1.56	2.23	3.07	3.87	2.01
	(0.03)	(0.04)	(0.04)	(0.03)	(0.05)	(0.06)	
Uninformed payoff	0.63	1.05	1.44	1.77	1.93	2.13	1.49
	(0.03)	(0.04)	(0.04)	(0.03)	(0.05)	(0.06)	
deal rate	0.42	0.48	0.54	0.69	0.73	0.81	0.61
	(0.06)	(0.05)	(0.03)	(0.02)	(0.02)	(0.02)	
Surplus Loss	0.58	1.04	1.39	1.25	1.36	1.16	1.13
	(0.06)	(0.10)	(0.10)	(0.10)	(0.10)	(0.11)	
Information value* **	-0.11	-0.05	0.05	0.31	0.83	1.39	0.40
	(0.03)	(0.03)	(0.04)	(0.04)	(0.07)	(0.10)	

* Averages are calculated for deal games only.

** Means and standard errors are calculated by treating each session's mean as a single observation.

*** Information value = the mean difference between the informed and uninformed payoffs.

4.5 Results

Main findings

The data are each subject's bargaining positions and the outcomes of 120 periods.[7] We first note that strike rates and offer amounts were not significantly different in the two subject groups (Caltech vs. UCLA). Strike rates do appear to decline somewhat with experience, but we report results across all periods and include controls for period number (see Appendix 4.B for details). Therefore, we pool these data across subject groups.

We observed the following empirical regularities:

Result 1. *deal rates are increasing with the pie size.*

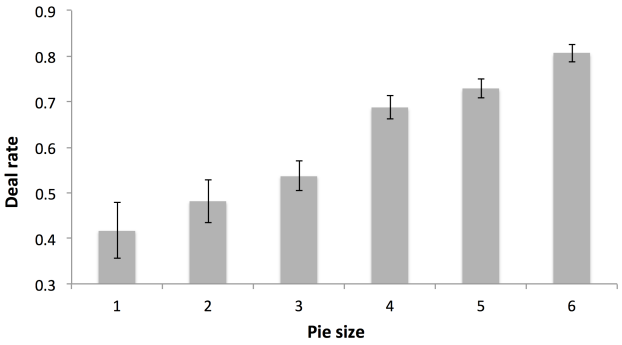
The mean deal rates for pie are summarized in Table 4.1 and Figure 4.2A. While the *probability* of disagreement decreased with the pie size, the mean amount of surplus lost due to strikes (Table 4.1) was positively correlated with the pie, as relatively small amounts of money are lost when strikes occur in small pie games.

Result 2. *When the pie is small or medium ($\pi \leq \$4$), the modes of the uninformed players' payoffs distribution are half of the pie; in large pie games ($\pi > \$4$) the*

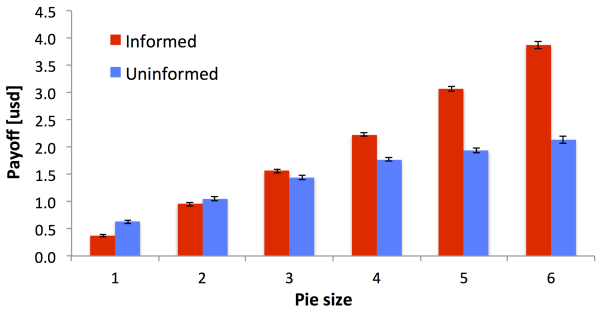
[7] A small fraction (less than 2.5 percent) of the games were excluded from analysis, due to a software bug in the first sessions conducted

Figure 4.2: Deal rates and mean payoffs across pie sizes. Standard errors are calculated at the session level.

(a) deal rates by pie size



(b) Mean payoffs by pie size and subject type, periods ending in a deal. Standard errors are calculated at the session level.



modes are \$2 *and there are local maxima at the half of the pie.*

The distributions of uninformed players' payoffs are in Figure 4.3. The mean payoffs (conditional upon a deal being reached) are in Figure 4.2B.

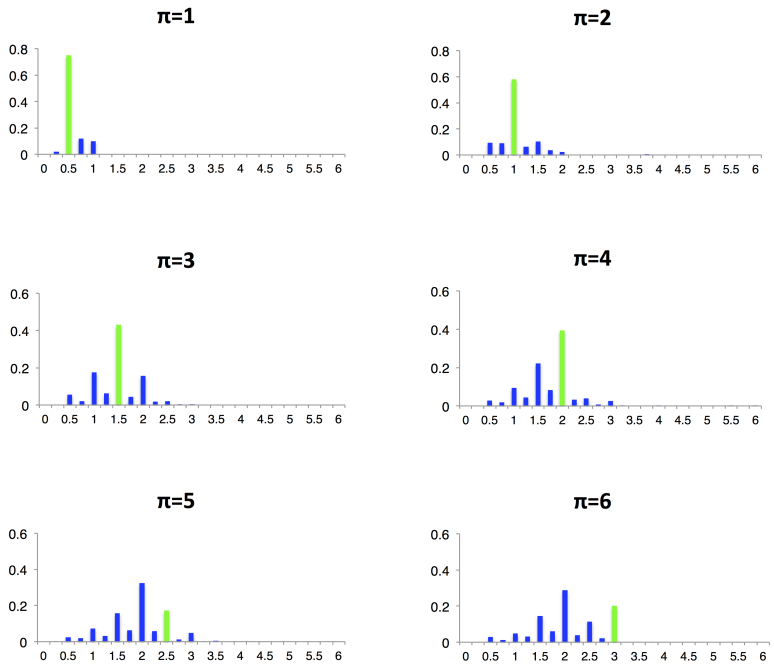
Result 3. *The informed players' offers increase, and the uninformed players' demands decrease with time (within a trial).*

Result 3 is illustrated by the plots of mean bargaining positions shown in Figure 4.4.

Result 4. *Most deals are made close to the deadline.*

More than half of the deals were made in the last two seconds of bargaining. Figure 4.5 shows the cumulative distribution function (CDF) of deals over time, which sharply increased as the deadline approached for all pies. Generally, deals were reached sooner when the pie was larger. This result is in line with the "deadline

Figure 4.3: Uninformed player's payoff relative frequencies (deal games, binned in a $0.25 resolution). The green bar locates the half of the pie in each distribution.



effect" reported in previous studies of unstructured bargaining with full information (Roth, Murnighan, and Schoumaker, 1988; Gächter and Riedl, 2005).

Comparison with focal equilibria

We now turn to testing the qualitative and quantitative predictions derived from the bargaining theory. In particular, we test the predictions of the efficient and equal-split equilibria. For convenience, we refer to the informed and uninformed players' bargaining positions as 'offers' and 'demands', respectively.

Payoff distributions

Overall, 82% of the payoffs, conditional upon a deal being reached, match values that are halves of one of the six possible pies.[8] Equal splits are the most prevalent

[8]In our experimental interface, players communicated their bids in integer multiples of 0.2, and therefore could not make offers of exactly 0.5, 1.5 or 2.5. We consider offers that are within 0.1 of these values (that are as close as one could get to them) as matching half of integer pies.

Figure 4.4: Mean bargaining position for all pie sizes (all periods pooled)

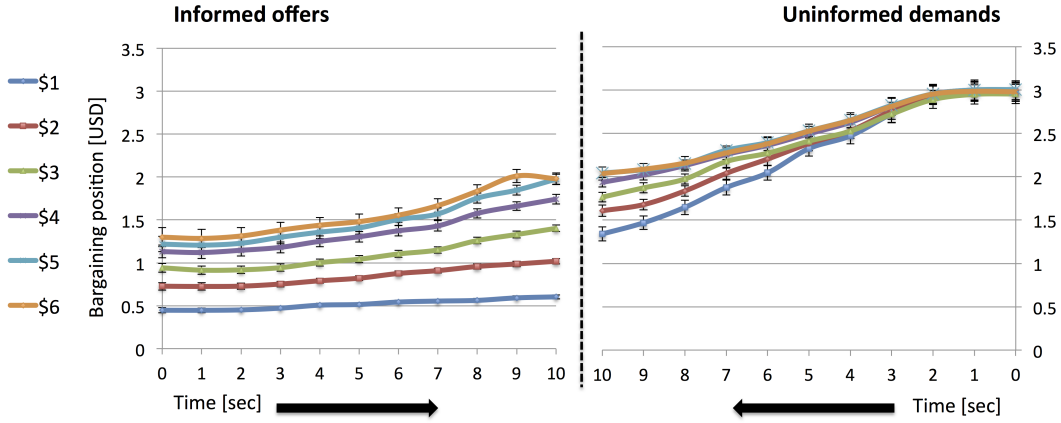
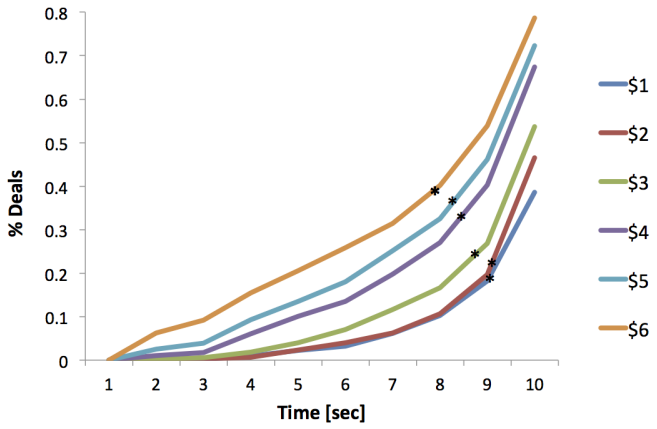


Figure 4.5: Cumulative distribution of deal times by pie size. Median deal times are marked by an asterisk



outcomes (54%) of small and medium pie games ($\pi \leq 4$), where the predicted payoffs of the efficient and equal equilibria coincide (Figure 4.3, top two rows). These results confirm that equality concerns did influence bargaining outcomes, generalizing the experimental literature studying complete information bargaining (Nydegger and Owen, 1974; Roth and Michael W Malouf, 1979) and ultimatum bargaining with private information (Mitzkewitz and Nagel, 1993) to an unstructured environment with informational asymmetry.

In large pie games ($\pi \geq 5$), equality and efficiency are in discord. The payoff distributions of these games (Figure 4.3, bottom row) have modes at the efficient (but unequal) uninformed payoff of 2 (31% of payoffs), and local maxima (19% of payoffs) at the equal (but inefficient) payoffs of 2.5 ($\pi = 5$) and 3 ($\pi = 6$). Thus, about half of the bargaining payoffs match one of the two equilibria.

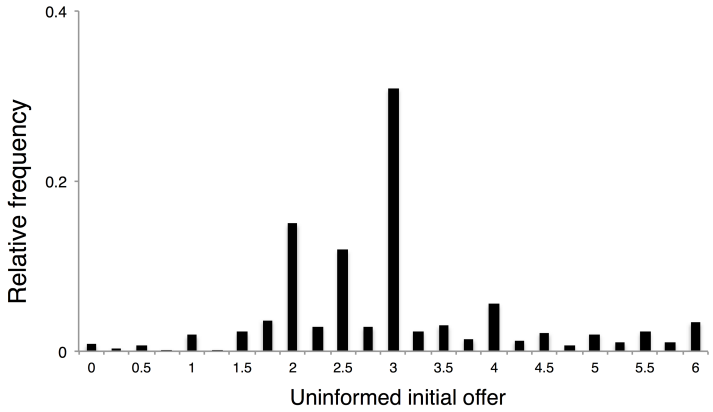
To further illuminate the role of equality and efficiency in large pie games, we investigate the uninformed players' *initial* bargaining positions (Figure 4.6). Many negotiation researchers see these initial demands as reflecting the players' aspirational payoffs, i.e., the most desirable payoffs that they can achieve, according to their beliefs (Yukl, 1974; White and Neale, 1994; Kristensen and Gärling, 1997; Galinsky and Mussweiler, 2001; Van Poucke and Buelens, 2002). The mode of the initial demands distribution (pooled across all pie sizes[9]) was 3 (31%)– matching the highest possible equal equilibrium payoff. An additional local maxima at 2 (19%) matched the highest possible payoff in the efficient equilibrium. Thus, the majority of the uninformed players' initial demands *exactly* match their maximal payoffs in either the efficient or equal equilibria, with a greater proportion matching the equal-split equilibrium.

Deal rates

Empirical deal rates rise smoothly with increasing pie size, in line with the qualitative prediction derived using the IC condition (lemma 1). However, strikes occur in all pie sizes, in contrast to the efficiency condition (Lemma 2); see Table 4.1, Figure 4.2. This finding is not surprising in light of the uninformed players' initial bargaining positions (discussed just above), most of which reflect the equal, rather than efficient equilibrium payoffs.

[9]Pooling the demands makes sense because the uninformed players have no information regarding the realization of the pie that might be deduced from the behavior of the informed player at the initial offer stage.

Figure 4.6: Uninformed player's initial demands (pooled across all games, binned in a $0.25 resolution).



strikes are common (19%), even at the largest pie size of 6– in contrast to the predictions of both equilibria. It is important to note that in some interesting models, and under certain experimental conditions, strikes can occur even with complete information (e.g. Roth and Michael W Malouf, 1979; Roth, Michael WK Malouf, and Murnighan, 1981; Roth and Murnighan, 1982; Roth, 1985; Haller and Holden, 1990; Herreiner and Puppe, 2004; Gächter and Riedl, 2005; Gachter and Riedl, 2006; Embrey, Hyndman, and Riedl, 2014). If the forces operating in such models and environments also apply in our private-information settings, the strike rates could be larger than those predicted by the mechanism design approach.

One factor that might account for disagreement rates that are higher than predicted is false revelations made by the informed players (i.e., offers that are too low). To assess the role of this factor, we estimated three logistic regression models with the dependent measure $deal = 1$ (i.e., $strike = 0$), that included subject-level dummy variables (for both informed and uninformed players) and control for period.[10] We estimated a model that includes the pie size alone (Model A), the final offer made by the informed player alone (B) [11], and both pie size and final offers (C). Our analysis (Table 4.2) reveals that Model B, which includes the final offers, fits the data better than Model A which includes the pie size, as implied by a lower Akaike Information Criterion (AIC) score. Furthermore, when including both the pie size

[10] The regression effects are robust to inclusion/exclusion of these controls.

[11] All bargaining positions lacked the players' ability to commit, with the exception of (a) positions at the deadline (i.e., 8.5 seconds into the bargaining process); (b) positions at the time a deal is made, 1.5 seconds after the positions' initial match had occurred. We refer to these bargaining positions as "final offers". We successfully replicated all of the analyses while setting the time of the final offers to 8 and 9 seconds into the bargaining process.

and the final offers in the model, the marginal effect of the latter was almost 6 times greater.[12] These results show that bargaining process (in this case the information revealed by the informed player's behavior) plays an important role in determining whether a deal is reached, beyond the actual realization of the pie size. As private information might be unobservable to an econometrician in more realistic settings, this finding has important practical implications, that we further discuss in the following sections.

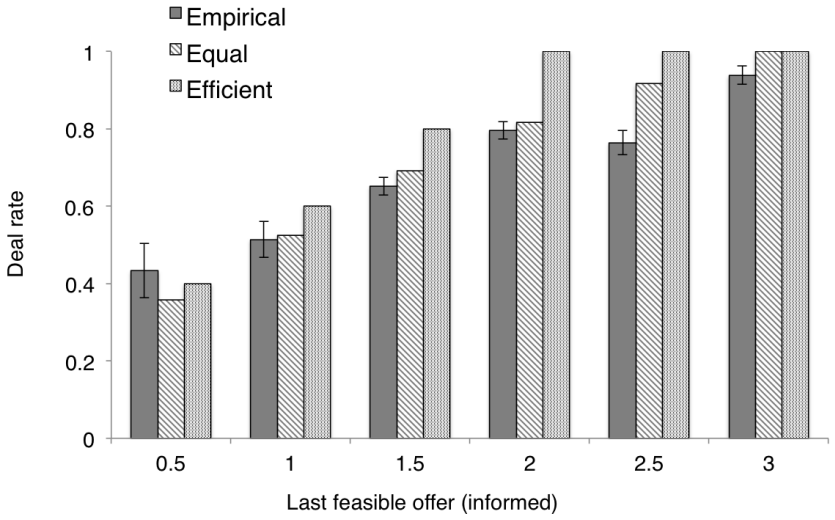
We estimated the empirical deal rates (across all pie sizes) as a function of the informed player's final offers (see Figure 4.7) and found that the empirical likelihood of reaching a deal was 74% and 94% when the final offers of the informed players matched the halves of the large pies ($2.5 and $3). The deal rate was 79% when the final offer was $2, lower than the efficient strike condition prediction (of no strikes), but very close to the deal rate predicted by the equal split equilibrium. In smaller pie-sizes ($1, 2, 3), disagreement rates were also closer to the prediction of the equal split equilibrium (Figure 4.7).

In summary, we find support for the qualitative prediction that deal rates increase with the pie size. Disagreement rates match the equal split equilibrium better than the efficient one. Further investigation of the initial demands suggests that the uninformed players aspire to an equal split in all pies; therefore, striking in high stake games might implement the uninformed players' strategy to enforce equal splits. The payoff distribution modes in small pie games are at the equal split (in accordance both equilibria). In large pie games, where there is a conflict between efficiency and equality, the payoff distributions are bi-modal, with the global mode matching efficient equilibrium and a second local mode matching the equal split equilibrium.

Our results demonstrate that theoretical predictions, derived from mechanism design models which assume risk-neutral, selfish players, can take us a long way even in unstructured settings, but also reveal the limitations of this approach. Bargaining outcomes match a mix of two equilibrium patterns and some game outcomes match neither equilibria. Furthermore, theoretical predictions critically depend on the pie size– which is private information that would typically be unobservable to econometricians in field data. In the next section, we use bargaining process data to overcome some of these limitations.

[12]The regression results are robust to the inclusion of quadratic terms for the pie size and period (which effects are statistically insignificant) and to variation in the definition of final offer, by setting its time to $t = 8$ and $t = 9$.

Figure 4.7: Empirical and theoretical deal rates for informed player's final offer matching pie halves (standard errors are clustered at the session level)



.

Table 4.2: Logistic regression - simple predictors of deals.

	A Mfx / SE	B Mfx / SE	C Mfx / SE
Pie	0.1027***		0.0398***
	(0.0052)		(0.0065)
Period	0.0007***	0.0002	0.0003
	(0.0002)	(0.0002)	(0.0002)
Final Offer		0.3090***	0.2322***
		(0.0168)	(0.0217)
Observations	6432	6432	6432
AIC	7295.074	7110.591	7066.761
BIC	7999.054	7814.571	7777.510

* p < 0.1, ** p < 0.05, *** p < 0.01. Robust standard errors in parentheses.

Note: Mfx: marginal effects. All models include subject controls. Standard errors clustered at the level of subject pairs.

4.6 Using process data

Our unstructured paradigm records bargaining process data that could be associated with outcome variables. This process data may be used to predict disagreements before the deadline has arrived. For example, suppose that at the 5 second mark, neither player has changed her offer for more than 3 seconds. This mutual stubbornness might be associated with an eventual strike. Our approach is to consider a large number of such candidate observable features in search of a small set that is predictive, using cross-validation (Stone, 1974) to control for overfitting. This machine learning approach has been used in many, many applications in computer science and neuroscience, and is beginning to be more widely used in economics (Krajbich et al., 2009; Belloni, Chen, et al., 2012; Einav and Levin, 2014; Varian, 2014; Smith et al., 2014; Mullainathan, 2014; Bajari et al., 2015).

One possibility is that there is little predictive information in such features, after controlling for overfitting. Indeed, if players know what the predictive features are, they should alter their behavior in order to avoid costly disagreements, erasing the features' predictive power.[13] Another possibility is that there are numerous small influences on disagreement that the players simply do not notice and which may be picked up by our modeling.

Predicting disagreements using bargaining process data

We chose 34 behavioral features recorded during bargaining and randomly split the entire set of trials into ten groups. Examples of features are the current difference between the offer and demand and the time since the last position change. The full list is in an Appendix 4.C. For each of the 10 holdout groups, we trained a model to classify trials into disagreements or deals, using the remaining 90% of the data, by estimating a logistic regression with a Least Absolute Shrinkage and Selection Operator (LASSO) penalty (Tibshirani, 1996).[14] By applying these trained models, we then made out-of-sample predictions of the binary bargaining outcomes for each of the 10 holdout samples.[15]

[13] By the revelation principle, every equilibrium in our setting has a payoff-equivalent equilibrium of the direct mechanism. As the direct mechanism is "process free", process features should not have predictive power in equilibrium after controlling for pie size.

[14] A LASSO-penalized logistic regression maximizes the standard logistic regression log-likelihood function minus a penalty term equal to the the sum of their absolute values of the regression coefficients (their L_1 norm) to overcome potential overfitting of the training data. The procedure includes a pre-processing stage of standardizing the dependent variables to have mean 0 and standard deviation 1.

[15] We use cross-validation to determine the weight placed on the penalty term in the LASSO regression. In our setting cross-validation involves partitioning the training data into k subsets,

As noted above, the pie size is a strong predictor of disagreements. The challenges for our machine learning approach are two-fold. First, we investigate whether process features have predictive power similar to the pie size when studied alone. In other words, we test whether process data allows predicting bargaining outcomes when the pie size, which is private information, is treated as if it was unobservable. Second, we investigate whether process features add predictive power when used *together* with the pie size.

To assess the predictive power of process data, we estimated three strike prediction models at eight different points in the bargaining process, separated by 1 second intervals (i.e. $1, 2, \ldots, 8$ seconds after bargaining started). One model relies only on the pie size, the second uses only process features, and the third uses both pie size and process features.[16]

We evaluate our results using "Receiver Operating Characteristic" (ROC) curves (Hanley and McNeil, 1982; Bradley, 1997). ROC is a standard tool in signal detection theory, used for quantifying the performance of a binary classifier under different trade-offs between type I and type II errors. A familiar example is a household smoke alarm: the alarm can be tuned to be very sensitive, indicating a fire when a burnt toast creates too much smoke. Or it can be tuned to be insensitive, ignoring the smoke from burnt toast, but also possibly ignoring smoke from a genuine fire caused by a half-lit cigar accidentally knocked onto a copy of the Daily Prophet newspaper.

The use of an ROC curve reflects the fact that one can always create more true positives (in our example, predicting more strikes) but doing so comes at the cost of then predicting more false positives (predicting strikes that don't happen). When using these methods, one would often like to know the tradeoff between correctly detecting true positives more accurately and also reducing the probability of false positives. A curve mapping all pairs of true and false positive levels therefore allows choosing an optimal policy for every given relative cost of the two types of errors.

To calculate the ROC, we subjected the out-of sample predicted deal probabilities (calculated by applying the estimated logistic LASSO regression weights to the out-

holding out one of the subsets, and calculating coefficient values (models) over a range of penalty weights. For each penalty weight, the model's out-of-sample predictive performance is calculated on the hold-out sample. The process is then repeated by holding out each of the other $k - 1$ subsets, and the final penalty weight is chosen as the value of the penalty that results in the best out-of-sample predictive performance over all k hold-out samples.

[16] We included only trials that were still in progress (when a deal has not yet been achieved), and excluded trials in which the offer and demand were equal at the relevant time stamp.

of-sample process data) to different decision thresholds, i.e., for a decision threshold $\tau \in [0..1]$, all predicted values less than τ were classified as "strike' where predicted values greater than or equal to τ were classified as "deal".[17] Every point on the ROC, therefore, represents a decision threshold, such that its coordinates represent the empirical false positive and true positive rates, calculated using the threshold.

For a random classifier, the true positive and false positive rates are identical (the 45-degree line in Figure 4.8). A good classifier increases the true positive rate (moving up on the y-axis) and also *decreases* the false positive rate (moving left on the x-axis). The difference between the ROC and the 45-degree line, in the upper-left direction, also known as the "area under the curve' (AUC, Bradley, 1997) is an index of how well the classifier does.[18]

The ROC analysis shows that process data does better than random for every time stamp (for illustration, see Figure 4.8). Furthermore, the mean out of sample prediction accuracy of the classifier, using solely process features, is as high as a classifier using solely the pie size, for times greater than 5 seconds into the bargaining process. Combining pie size and process features improves accuracy further: a classifier using both pie size and process data outperforms the classifier using the pie size alone as early as 2 seconds into the bargaining process (Figure 4.8).

Which bargaining process features predict disagreements?

To further investigate which behavioral process features predict disagreements, we used a "post-LASSO" procedure (Belloni and Chernozhukov, 2009; Belloni, Chen, et al., 2012).[19] Figure 4.9 summarizes the marginal effects of the most predictive process features (z-scored for every intra-trial, i.e, within period, time point), such that an "interaction" represents a multiplification of two variables. The marginal effects of all process features investigated are reported in appendix 4.C.

Not surprisingly, the most predictive process features are the current informed player's offer (positively associated with a deal) and the current difference between

[17] We used decision threshold between 0 and 1 on a grid with a resolution of 0.01.

[18] The AUC is closely related to the Mann-Whitney-Wilcoxon U-statistic (Hanley and McNeil, 1982).

[19] The "post-LASSO" procedure consisted of three steps. First, we optimized the LASSO tuning parameter λ using 10-fold cross validation on the entire data set. Second, we conducted model selection by fitting a logistic LASSO regression using the optimized tuning parameter to the data. Finally, we fitted an ordinary logistic regression to the data, using the features with non-zero LASSO coefficients from the second stage.

Figure 4.8: Strike prediction using bargaining process data, Receiver Operating Characteristic (ROC). The dashed lines represent the false and true positive rates of a random classifier.

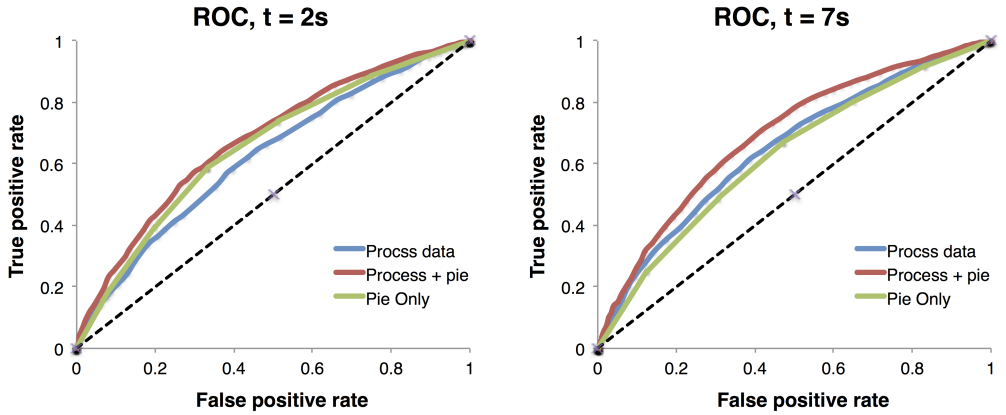
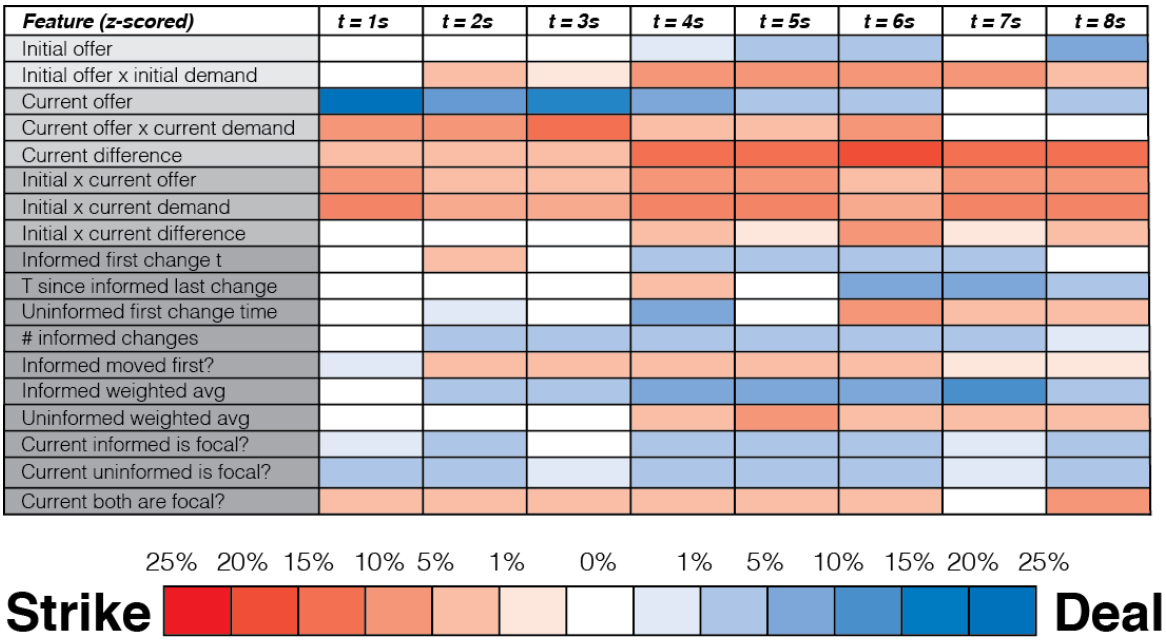


Figure 4.9: Bargaining process features selected by the classifier for outcome prediction (deal=1) and their estimated marginal effects. (Pie sizes are excluded.)



the players' bargaining positions (positively associated with a strike). More surprisingly, the players' *initial* bargaining positions contain predictive information regarding the chance of reaching a deal, even as the deadline approaches, and even after controlling for current offers. The informed player's initial offer is positively associated with a chance of a deal, and the effect is moderated by the uninformed player's initial demand, as implied by a negative interaction between the two factors. Thus, initial offers are mostly associated with deals when the initial demands are low. There was also an intriguing negative interaction between the initial and current offers: the *current* offer becomes particularly associated with a deal when the *initial* offer is low. This result is consistent with an idea from negotiation research that initial offers serve as reference points in bargaining. When initial offers are low, they make later, more generous offers seem more attractive, and increase the chances of a deal (Galinsky and Mussweiler, 2001).

Our analyses further revealed a rich set of behavioral features that reliably predicted disagreements throughout the bargaining process, even after controlling for the current bargaining positions. For example, an increased activity on the informed player side (i.e., many position changes) is a precursor of an upcoming deal, as early as two seconds into the bargaining process. The use of focal points (i.e., offers and demands that match halves of the integer pies) was positively associated with an upcoming deal, unless both players' positions match *different* focal points, as implied by a negative marginal interaction effect. This finding suggests that disagreements may arise as a result of a coordination failure when players use different focal points to communicate their claims, in line with Roth's focal theory of bargaining (Roth, 1985). However, one must keep in mind that despite the substantial predictive power of all process data features jointly, the marginal effect of most features alone was relatively small (less than 5%, see appendix 4.C for all marginal effects).

Using bargaining process data for statistical mediation

The statistical value of process measures is important for studying bargaining in naturalistic settings. The accuracy of features alone for predicting strikes (even without pie size) suggests that it could be possible to use this type of analysis to do statistical mediation. That is, an important and often overlooked body of theory in mechanism design showing that if the designer has an independent measure of private information (which the informed player cannot manipulate or hide), efficiency can be enhanced by conditioning mechanism outcomes on this independent measure.

Intuitively, suppose in our setting the pie size is $6. For the IC constraint to bind, the mechanism must impose strikes when a lower pie size is (untruthfully) reported, to prevent an informed player from misreporting that the pie is worth less than $6. But what if there were another indicator measure of pie size which is sufficiently accurate and not manipulable? Then the mechanism could combine this indicator with the reported pie size, penalizing the informed player if her report and the indicator disagree.

A proof of principle that such a mechanism can work was offered by Krajbich et al., 2009. They used neural measures of private value for a public good in a threshold public goods game. In their domain, it was shown that the mechanism satisfies the voluntary participation (IR) constraint provided the mechanism is sufficiently accurate and agents are not too risk-averse.

In future work, process measures could be used as indicators of likely strikes, or as indicators of pie sizes, to create behaviorally-enhanced mechanisms which avoid disagreements. Such a process-informed mechanism can, in principle, reduce strikes and improve efficiency, while also satisfying voluntary participation constraints so that bargainers will agree to use them.

4.7 Conclusion

Much of the recent literature on bargaining has studied structured bargaining. We reiterate here our motivations for studying unstructured bargaining in dynamic and uncertain environments. First, much real-world bargaining is unstructured and involves private information; unstructured bargaining generates process data that can be used to predict strikes ahead of time; and theory can be used to make precise predictions even with minimal structure.

In this paper we study dynamic unstructured bargaining in a game with one-sided private information. We combine mechanism design theory with an equilibrium selection approach that builds on a well documented empirical regularity: the appeal of an equal split as a bargaining focal point. Our approach is agnostic regarding the driving force behind equal splits. A large theoretical literature attempts to address the question of why equal splits are focal; equal splits might result, for example, from inequality aversion, concerns about fairness, or social norms. Another explanation might be lying aversion, and our experimental design, which incorporates feedback after each round of bargaining, may encourage truthful revelation. However, our design also involves random, anonymous re-matching of bargaining partners after

each game, which might be expected to act in the opposite direction.

Our theoretical model predicts that the rate of bargaining failures will be decreasing in the pie size. The additional assumption of interim incentive efficiency implies that the distribution of surplus will favor the informed player when the pie size crosses a threshold. We find support for both of these hypotheses in our data. However, we also observe an interesting departure from the "efficient" benchmark: bargaining failures arise even at the highest pie levels and even after many rounds of play, and the surplus is divided equally in many high stake games, in contrast to the efficient equilibrium prediction.

In theory, the uninformed players' payoffs must be identical in all pies where no disagreements occur, generating an inherent trade-off between efficiency and equality. We propose two ways to resolve this tension, by either favoring efficiency and dividing the pie equally given the efficiency constraint ("efficient" equilibrium) or by imposing equal splits and only then maximizing efficiently ("equal split" equilibrium). While the modes of the distributions of the informed player's final offers more closely match the efficient equilibrium, deal rates more closely match the the equal-split equilibrium. Further, the uninformed players' initial offers reflect aspirations of equal splits in the largest pie. The latter two patterns suggest that the uninformed players use disagreements as means to impose equal splits despite the loss of efficiency.

Although our results show that theoretical predictions go a long way even in an unstructured setting, they also highlight their limitations. The data qualitatively match the mix of the two equilibria patterns, but some games do not match either. Further, the theoretical prediction depends on the realization of the pie size, which is private information, and therefore might not be observable in many realistic circumstances. We propose to overcome this obstacle by analyzing bargaining process data.

Our machine learning approach shows that process data is incrementally informative for predicting strikes when the pie size is included in the model, and is as just as informative as knowing the pie size when the latter is unobservable (before the deadline has arrived). These results suggest that some bargaining failures may result from process "mistakes" that could have been avoided if players had behaved differently. Process data may be used to avert strikes and other inefficient disagreements by offering 'course corrections' in the bargaining process. Bargaining process data could potentially be much richer, and therefore substantially more

informative, than the series of cursor locations that our exploratory investigation has focused on. Our results should therefore be considered as a lower bound regarding the predictive power of process data. Incorporating bargaining features such as verbal communication, non-verbal gestures (e.g., facial expressions, body language), and physiological responses (e.g., skin conductance, pupil dilation, brain activity) are likely to improve predictive performance. These biomarkers could be informative for understanding the origins of costly disagreements in bargaining.

Finally, we acknowledge that our laboratory bargaining institution deliberately omits many features of natural bargaining. Lifelike bargaining is often face-to-face, has little anonymity, uses natural language, includes repetition and resulting reputations, and typically has two-sided private information. Adding more lifelike features can also be easily done step-by-step, as part of a research program reviving interest in unstructured bargaining. Typically, adding natural institutional properties makes it harder to Figure out theoretically what behavior will result. The opposite is true when machine learning is used: adding more natural institutional properties simply adds more “features" that can be used for prediction.

APPENDIX

4.A Mathematical Appendix

Proof of Lemma 1

As noted in the main text (Equation 4.3) individual rationality (IR) and incentive compatibility (IC) for the informed player imply that:

$$\gamma_k \pi_k - x_k \geq \gamma_j \pi_k - x_j \text{ for all } j \neq k.$$

We restate Lemma 1 here:

Lemma 1. *If the bargaining mechanism satisfies IR and IC:*

1. *deal rates are monotonically increasing in the pie size k.*
2. *The uninformed player's payoffs are monotonically increasing in the pie size.*
3. *The uninformed player's payoff is identical for all states in which the deal probability is 1.*

We first show that γ_k is decreasing in k (Lemma 1.1), and then rely on Lemma 1.1 for the proofs of Lemmas 1.2 and 1.3.

Proof. Consider π_k and π_{k+1}. Incentive compatibility requires

$$\begin{aligned} \gamma_k \pi_k - x_k &\geq \gamma_{k+1} \pi_k - x_{k+1} \\ \gamma_{k+1} \pi_{k+1} - x_{k+1} &\geq \gamma_k \pi_{k+1} - x_k. \end{aligned}$$

These two equations imply that

$$(\gamma_{k+1} - \gamma_k)\pi_{k+1} \geq x_{k+1} - x_k \geq (\gamma_{k+1} - \gamma_k)\pi_k \tag{4.11}$$

and therefore

$$(\gamma_{k+1} - \gamma_k)(\pi_{k+1} - \pi_k) \geq 0. \tag{4.12}$$

By definition, $\pi_{k+1} \geq \pi_k$, so then $\gamma_{k+1} \geq \gamma_k$, and therefore the disagreement (or strike) rate $1 - \gamma_k$ is monotonically decreasing in the pie size.

The remaining results follow directly from Equations 4.11 and Lemma 1.1. By Lemma 1.1, $(\gamma_{k+1} - \gamma_k)\pi_k \geq 0$, so by Equation 4.11 $x_{k+1} - x_k \geq 0$, and therefore the uninformed player's payoffs are monotonically increasing in the pie size (Lemma 1.2). Furthermore, replacing $\gamma_k = \gamma_{k+1} = 1$ in the righthand inequality of Equation 4.11, it immediately follows that $x_k = x_{k+1}$ (Lemma 1.3). □

Proof of Lemma 2: The strike Condition

A mechanism is interim-efficient if it is Pareto optimal for the set of $K + 1$ agents: the informed player for each pie size, and the uninformed player.

Following FKS, we first show that strikes in the "best" pie size π_K are never efficient for the class of direct mechanisms that we consider. That is, if the mechanism $\mu = \{\gamma_k, x_k\}_{k=1}^K$ is efficient, then it must be the case that $\gamma_K = 1$.

If μ is an efficient mechanism, then the incentive compatibility conditions must hold and so by Lemma 1 $\gamma_K \geq \gamma_k$ for all $k \leq K$. If $\gamma_K = 1 - \delta < 1$, we can define a new mechanism μ^* with $\gamma_K^* = 1$, $\gamma_k^* = \gamma_k + \delta$, for all $k < K$, and $x_k^* = x_k$, for all k. The mechanism μ^* does not affect the uninformed player's expected payoff, but it increases the informed player's payoff by $\delta\pi_K$ in state K and by $\delta\pi_k$ in states $1, \ldots, K - 1$, so the original mechanism cannot be efficient.

Next, if γ_k, $k < K$, can be increased without violating the IC constraint, the uninformed bargainer is unaffected as is the informed bargainer in states $j \neq k$, while player I_k, the informed bargainer in state k, is made better off. Therefore, efficiency requires that right-hand side of Equation 4.11 holds at equality:

$$x_{k+1} - x_k = (\gamma_{k+1} - \gamma_k)\pi_{k+1}. \tag{4.13}$$

We make use of Equation 4.13 to derive Lemma 2, the strike condition, below.

Lemma 2. *The strike condition: For IR and IC mechanisms, strikes in state k are ex-ante efficient if*

$$\frac{\pi_k}{\pi_{k+1}} < \frac{(1 - \sum_{j=1}^{k} p_j)}{(1 - \sum_{j=1}^{k-1} p_j)} = \frac{\Pr(\pi \geq \pi_{k+1})}{\Pr(\pi \geq \pi_k)}. \tag{4.14}$$

Proof. To derive the strike condition, consider mechanisms $\mu = \{\gamma_k, x_k\}_{k=1}^K$ and $\mu^* = \{\gamma_k + \delta_k, x_k + d_k\}_{k=1}^K$ which satisfy IR and IC, and assume that both satisfy 4.13. Since μ^* satisfies 4.13, we have

$$(x_{k+1} + d_{k+1}) - (x_k + d_k) = ((\gamma_{k+1} + \delta_{k+1}) - (\gamma_k + \delta_k))\pi_{k+1}. \tag{4.15}$$

By subtracting 4.13 from 4.15, we find a useful condition that

$$d_{k+1} - d_k = (\delta_{k+1} - \delta_k)\pi_{k+1}. \tag{4.16}$$

Next, assume that strikes are not efficient in states $k+1, \ldots, K$, so that $\gamma_j = 1$ if $j > k$, but assume that that $\gamma_k < 1$. This implies that $d_{k+1} = \ldots = d_K$.

Let ΔV_k and ΔU represent the difference in payoffs between μ^* and μ for the informed player in state k and the uninformed player, respectively. If $\mu*$ dominates μ, then $\Delta V_k \geq 0$ for all k, and $\Delta U \geq 0$, and at least one of these inequalities is strict.

First, consider the K conditions for the informed player:

$$\begin{aligned}
\Delta V_1 &= \delta_1 \pi_1 - d_1 \geq 0 \\
&\vdots \quad\quad \vdots \\
\Delta V_j &= \delta_j \pi_j - d_j \geq 0, j < k \\
\Delta V_k &= \delta_k \pi_k - d_k \geq 0 \\
\Delta V_j &= \delta_k \pi_{k+1} - d_k \geq 0, j > k.
\end{aligned}$$

Multiplying the conditions for players $I_1, \ldots, I_k$ by p_j and summing them up gives

$$\sum_{j=1}^{k} p_j \pi_j \delta_j \geq \sum_{j=1}^{k} p_j d_j.$$

Multiplying the equation for player k by $(1 - \sum_{j=1}^{k-1} p_j)$ gives

$$(1 - \sum_{j=1}^{k-1} p_j)\delta_k \pi_k \geq (1 - \sum_{j=1}^{k-1} p_j) d_k.$$

Adding up these two conditions gives:

$$\sum_{j=1}^{k} p_j \pi_j \delta_j + (1 - \sum_{j=1}^{k-1} p_j)\delta_k \pi_k \geq \sum_{j=1}^{k} p_j d_j + (1 - \sum_{j=1}^{k-1} p_j) d_k. \tag{4.17}$$

Next we consider the uninformed player. If μ^* dominates μ, it must be the case that the uninformed player's payoff from μ^* is at least as large as in μ:

$$\Delta U = \sum_{j=1}^{K} p_j d_j = \sum_{j=1}^{k} p_j d_j + (1 - \sum_{j=1}^{k} p_j) d_{k+1} \quad \geq \quad 0$$

$$\begin{aligned}
\sum_{j=1}^{k} p_j d_j + (1 - \sum_{j=1}^{k} p_j)(d_k - \delta_k \pi_{k+1}) &\geq 0 \\
\sum_{j=1}^{k-1} p_j d_j + p_k d_k + (1 - \sum_{j=1}^{k} p_j) d_k - (1 - \sum_{j=1}^{k} p_j)\delta_k \pi_{k+1} &\geq 0 \\
\sum_{j=1}^{k-1} p_j d_j + (1 - \sum_{j=1}^{k-1} p_j) d_k - (1 - \sum_{j=1}^{k} p_j)\delta_k \pi_{k+1} &\geq 0 \\
\sum_{j=1}^{k-1} p_j d_j + (1 - \sum_{j=1}^{k-1} p_j) d_k &\geq (1 - \sum_{j=1}^{k} p_j)\delta_k \pi_{k+1} \qquad (4.18)
\end{aligned}$$

Combining Equations 4.17 and 4.18 gives

$$\sum_{j=1}^{k-1} p_j \pi_j \delta_j + (1 - \sum_{j=1}^{k-1} p_j)\delta_k \pi_k \geq \sum_{j=1}^{k-1} p_j d_j + (1 - \sum_{j=1}^{k-1} p_j)(d_k) \geq (1 - \sum_{j=1}^{k} p_j)\delta_k \pi_{k+1}. \tag{4.19}$$

And this implies that

$$\sum_{j=1}^{k-1} p_j \pi_j \delta_j + (1 - \sum_{j=1}^{k-1} p_j)\delta_k \pi_k \geq (1 - \sum_{j=1}^{k} p_j)\delta_k \pi_{k+1}. \tag{4.20}$$

To examine whether strikes are efficient in state k, suppose μ^* and μ have identical strike rates in all states $j < k$. Then δ_j equals 0 for all $j < k$, implying that

$$(1 - \sum_{j=1}^{k-1} p_j)\delta_k \pi_k \geq (1 - \sum_{j=1}^{k} p_j)\delta_k \pi_{k+1}. \tag{4.21}$$

Then $\delta_k > 0$ implies that strikes are inefficient in state k if

$$\frac{\pi_k}{\pi_{k+1}} \geq \frac{(1 - \sum_{j=1}^{k} p_j)}{(1 - \sum_{j=1}^{k-1} p_j)},$$

implying that strikes are efficient in state k if

$$\frac{\pi_k}{\pi_{k+1}} < \frac{(1 - \sum_{j=1}^{k} p_j)}{(1 - \sum_{j=1}^{k-1} p_j)},$$

or alternatively

$$\frac{\pi_k}{\pi_{k+1}} < \frac{\Pr(\pi \geq \pi_{k+1})}{\Pr(\pi \geq \pi_k)}. \tag{4.22}$$

□

Calculating strike rates using focal points: efficient equilibrium

The strike condition implies that disagreement is inefficient when the pie size is $4, 5$ or 6, so we first fix $\gamma_4 = \gamma_5 = \gamma_6 = 1$. Based on the strike condition and the equal split principle, payoffs conditional on a deal are $x_6 = x_5 = x_4 = 2$, $x_3 = 1.5$, $x_2 = 1$ and $x_1 = 0.5$. As no disagreement should occur for $\pi \in \{\$4, 5, 6\}$ and as the predicted equilibrium payoff is $x_4 = \$2$ for these pies, it follows that the informed player's payoff for $\pi \in \{\$5, 6\}$ is always greater than for $\pi = \$4$. Therefore, we set $x_k = 0.5\gamma_k\pi_k$ and then solve the IC inequalities for $\pi \leq \$4$:

$$\gamma_j \leq \frac{0.5\pi_k}{\pi_k - 0.5\pi_j}\gamma_k \text{ for all } k \leq 4, j \neq k. \tag{4.23}$$

Solving the inequalities for $k = 4$ and $j = 3, 2, 1$ yields

$$\gamma_3 \leq \frac{2}{2.5} \tag{4.24}$$

$$\gamma_2 \leq \frac{2}{3} \tag{4.25}$$

$$\gamma_1 \leq \frac{2}{3.5}. \tag{4.26}$$

Solving the inequalities for $k = 3$ and $j = 4, 2, 1$ yields

$$\gamma_3 > \frac{2}{3} \tag{4.27}$$

$$\gamma_2 < \frac{1.5}{2}\gamma_3 \tag{4.28}$$

$$\gamma_1 < \frac{1.5}{2.5}\gamma_3. \tag{4.29}$$

Solving the inequalities for $k = 2$ and $j = 4, 3, 1$ yields

$$\gamma_2 > 0 \tag{4.30}$$

$$\gamma_2 > 0.5\gamma_3 \tag{4.31}$$

$$\gamma_2 > 1.5\gamma_1. \tag{4.32}$$

Finally, for $k = 1$ it is always optimal to report the truth if $\gamma_1 > 0$, as offers exceeding 1 would generate a non-positive payoffs.

Maximal efficiency requires the largest possible values of $\gamma_1, \gamma_2, \gamma_3$ which are compatible with the IC inequalities. The only upper constraint on γ_3 is equation (4.24); thus, we set $\gamma_3 = \frac{2}{2.5} = 0.8$. The lowest upper constraint on γ_2 is equation (4.28); accordingly we set $\gamma_2 = \frac{1.5}{2}\gamma_3 = 0.6$. The value of γ_1 is constrained by equation(4.28), to be less than $.8 * \frac{1.5}{2.5} = .48$ and is constrained by equation(4.32) to be $\gamma_1 < \gamma_2/1.5 = 0.4$.Therefore, the maximal value is $\gamma_1 = .4$.

Calculating strike rates using focal points: equal split equilibrium

We first show that we can increase efficiency for every equal-split equilibrium that has strikes in the "best" pie size π_K. That is, if an equal-payoff mechanism $\mu = \{\gamma_k, x_k\}_{k=1}^K$ maximizes efficiency (given equality constraints), then it must be the case that $\gamma_K = 1$. Note that the analogous proof for the efficient equilibrium (see Appendix 4.A) does not hold for the qual split equilibrium, as in the latter the uninformed player's expected payoff immediately depends on the deal rate.

If μ is a maximal efficiency equal payoff mechanism, then the incentive compatibility conditions must hold and so by Lemma 1 $\gamma_K \geq \gamma_k$ for all $k \leq K$. If $\gamma_K = 1 - \delta < 1$, we can define a new equal-payoff mechanism μ^* with $\gamma_K^* = 1$, $\gamma_k^* = \gamma_k + \delta\gamma_k$, for all $k < K$, $x_K^* = x_k + \frac{\delta}{2}$, and $x_k^* = x_k + \frac{\delta\gamma_k}{2}$ for all k. The mechanism μ^* increases both players' expected payoffs by $\frac{\delta}{2}\pi_K$ in state K and by $\frac{\delta\gamma_k}{2}\pi_k$ in states $1, \ldots, K-1$. Further, μ^* is an equal-split incentive compatible (i.e., complies with Eq. 4.9), as the addition to both player's expected payoffs is equal (compared to μ) and the ratio between γ_k and γ_j is identical to μ for all k, j. Thus the original mechanism cannot be efficient maximizing equal-split equilibrium.

It follows that for maximizing efficiency (given equal splits) we must set $\gamma_6 = 1$. Then, we set for all k $w_k = 0.5\pi_k$ and solve the IC inequalities for all $\pi < 6$:

$$\gamma_j \leq \frac{0.5\pi_k}{\pi_k - 0.5\pi_j}\gamma_k \text{ for all } k < 6, j \neq k. \tag{4.33}$$

Numerically solving this set of inequalities and taking the highest possible deal rates (for maximal efficiency, in a similar manner to the solution of the efficient equilibrium), we get exact numerical predictions of the deal rate for every given pie size in the equal split equilibrium:

$$[\gamma_6,\ \gamma_5,\ \gamma_4,\ \gamma_3,\ \gamma_2,\ \gamma_1] = [1,\ 0.9167,\ 0.8167,\ 0.6917,\ 0.5250,\ 0.3583] \tag{4.34}$$

4.B Pooling data

Caltech SSEL vs. UCLA CASSEL

Summary information of all of the experimental sessions (location, number of subjects and gender by role) is recapitulated in Table 4.3. For comparing the sessions taking places at Caltech vs. UCLA, we first calculated the mean deal rates and payoffs (in case of a deal) for each subject and pie size, and contrasted the group averages (see Table 4.4). Qualitatively, deal rates and payoff were monotonically increasing with the pie for both groups. The most significant difference observed

Table 4.3: Session information, I-Informed, U-Uninformed

Session No.	Location	Date	N	I Male	I Female	U Male	U Female
1	Caltech	12/1/2011	10	3	2	3	2
2	Caltech	12/8/2011	10	2	3	2	3
3	Caltech	1/9/2012	8	3	1	2	2
4	Caltech	1/11/2012	16	5	3	5	3
5	Caltech	2/28/2012	8	3	1	1	3
6	UCLA	5/11/2012	18	6	3	6	3
7	UCLA	5/11/2012	20	4	6	6	4
8	UCLA	5/11/2012	20	6	4	6	4
Total			**110**	**32**	**23**	**31**	**24**

Table 4.4: Average payoffs (case of deal) and deal rates by pie size, Caltech vs. UCLA

Pie size	Venue	1	2	3	4	5	6
deal rates	Caltech	0.43	0.50	0.56	0.71	0.75	0.84
	UCLA	0.34	0.42	0.51	0.63	0.71	0.76
	p-value*	0.14	0.29	0.42	0.11	0.36	0.03
Payoff, informed	Caltech	0.39	0.98	1.60	2.23	3.02	3.83
	UCLA	0.36	0.95	1.55	2.31	3.19	4.06
	p-value*	0.67	0.61	0.56	0.40	0.05	0.05
Payoff, uninformed	Caltech	0.61	1.05	1.45	1.82	2.01	2.19
	UCLA	0.66	1.12	1.50	1.75	1.85	2.01
	p-value*	0.44	0.21	0.40	0.37	0.04	0.03

c *Two-sided t-tests, uncorrected for multiple comparisons.

between the groups was a 9 percent increase of deal rates in the largest pie ($6) at Caltech sessions. We used a 2-sided t-test to compare Caltech and UCLA subjects; while for some of the pies we found statistically significant differences at the 0.05 level, none of the differences survived correction for multiple hypothesis (p_{max} = 0.096 using the Bonferroni correction, for deal rates at $6 pie).

First vs. second half of the trials

To compare the first and second halves of bargaining periods, we calculated the mean deal rates and payoffs (in case of a deal) for each subject at any given pie size, and contrasted the averages of the first and second halves of the periods (see Table

4.5). Qualitatively, deal rates and payoff were monotonically increasing with the pie for both groups. The largest difference observed was 8 percent increase of efficiency (deal rates) in the second half compared to the first one, when the pie was $6. We further used a 2-sided t-test to compare the two halves. While for some of the pies we found statistically significant differences at the 0.05 level (in particular, deal rates were higher and informed players' payoffs in case of a deal were lower at the Caltech pool), none of the differences survived correction for multiple hypothesis ($p_{max} = 0.24$ using Bonferroni correction).

Table 4.5: Average payoffs (case of deal) and deal rates by pie size, first vs. second half of the trials

Pie size		**1**	**2**	**3**	**4**	**5**	**6**
deal rates	First 60	0.38	0.47	0.49	0.63	0.72	0.76
	Last 60	0.39	0.45	0.58	0.70	0.73	0.84
	p-value*	0.97	0.61	0.07	0.09	0.68	0.02
Payoff, informed	First 60	0.43	1.02	1.63	2.32	3.17	4.03
	Last 60	0.31	0.91	1.52	2.23	3.05	3.89
	p-value*	0.08	0.03	0.08	0.15	0.10	0.13
Payoff, uninformed	First 60	0.60	1.04	1.41	1.74	1.88	2.00
	Last 60	0.68	1.13	1.53	1.82	1.99	2.17
	p-value*	0.14	0.07	0.04	0.15	0.10	0.02

*Two-sided t-tests, uncorrected for multiple comparisons.

4.C List of process features and associated marginal effects

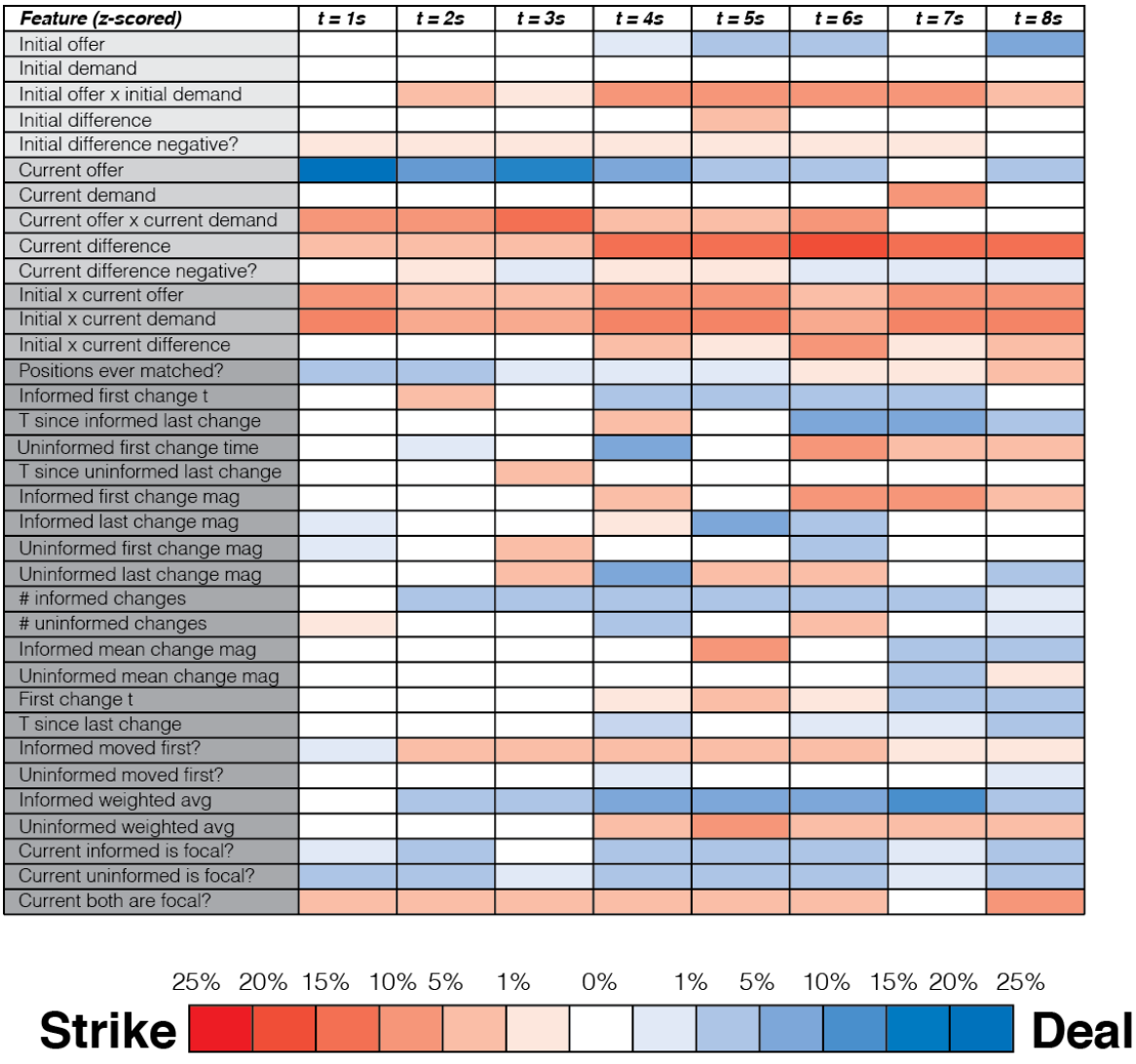
Figure, 4.10 summarizes all of the process features used to predict bargaining outcomes. We provide further details of calculating some of the features below.

Initial difference negative? A binary indicator that equals one if the initial offer of the informed player is greater than the initial uninformed player's demand and zero otherwise.

Positions ever matched? A binary indicator that equals one if the players' bargaining positions had previously matched and they later changed their minds.

Informed/ uninformed first change T. The first time in the game in which the informed / uninformed player has updated his or her initial bargaining position.

Figure 4.10: bargaining process features used for outcome prediction (deal=1) and their estimated marginal effects.



T since informed/ uninformed last change T. The time since the last time in which the informed / uninformed player has updated his or her bargaining position.

Informed/ uninformed first/last change mag. The magnitude of the last informed / uninformed position change.

 # informed/ uninformed changes. The number of times that the informed / uninformed player has changed his or her bargaining position since the start of the game.

Informed/ mean change mag. The mean magnitude of change in the informed / uninformed player when he or she changed bargaining positions.

first change T. The first time in the game in which either player has updated his or her initial bargaining position.

T since last change. The time since the last time in which wither player has updated his or her bargaining position.

Informed / uninformed moved first? A binary indicator that equals one if the informed / uninformed player was the first to change his or her bargaining position in the game.

Informed / uninformed weighted avg. A weighted sum of the informed/ uninformed bargaining positions across time.

$$\sum_{t=0}^{T} w_t x_t, \tag{4.35}$$

such that t denotes time (between 0 and the current time T, sampled in a $0.1sec$ resolution) and x_t is bargaining position in time t. The weight w_t equals

$$w_t = \frac{t^2}{\sum_{q=0}^{T} q^2}. \tag{4.36}$$

This results a linear combination where later bargaining positions are weighted more heavily than earlier ones.

Current informed / uninformed / both are focal? A binary indicator that equals one if the informed / uninformed / both players bargaining positions match the half of either possible pie size (i.e., 0.4, 0.6, 1, 1.4, 1.6, 2, 2.4, 2.6, 3).

4.D instructions

This is an experiment about bargaining. You will play 120 rounds of a bargaining game.

In the game, one participant (the informed player) is told the total amount of money (pie size) in each round. This amount will be $\$1, 2, 3, 4, 5$, or 6, chosen randomly in each trial. The amount will appear on the top left corner of the screen.

The other player is not informed of the pie size.

During each round, participants bargain over the uninformed player's payoff.

The roles are randomly selected and fixed for the duration of the experiment. Before each round, informed and uninformed players are randomly matched.

Participants negotiate by clicking on a scale from \$0 to 6 (see Figure 1). Amounts on the scale represent the uninformed player's payoff.

During the first 2 seconds, participants select their initial offers. Note that the initial location of the cursors is random. In the following 10 seconds, the participants bargain, using the mouse to select payoffs for the uninformed player. Clicking the mouse on a different part of the scale moves the cursor.

A deal occurs when the cursors are in the same place for 1.5 seconds. When both cursors are in the same place on the scale, a green rectangle will appear (see Figure 2).

If a deal is made, the informed player's payoff is equal to the pie size minus the negotiated uninformed player's payoff. If the agreement exceeds the total amount of money, the payoff will be negative.

If no deal has been made after 10 seconds of bargaining, both participants get \$0.

Following each trial, the uninformed player will be shown of the pie size.

The game has total 120 trials.

Before the experiment begins there will be 15 training trials, to allow you to practice.

At the end of the game, you will receive payment based on randomly selected 10% of your trials.

You will receive a $5 participation fee in addition to whatever you earn from playing the game.

Quiz

Total amount is $3. Cursors were matched in $1. How much money does the informed participant get? How much does the uninformed participant get?

Total amount is $2. Cursors were matched in $4.1. How much money does the informed participant get? How much does the uninformed participant get?

One second before the end of the trial, both participants have agreed on payoff of $2 and the green rectangle appears. What is going to happen when the trial ends?

Both participants have agreed on payoff of $2 and the green rectangle appears. After one second, the uninformed player changed his offer to $2.5. What is going to happen?

BIBLIOGRAPHY

Ames, Daniel R and Malia F Mason (2015). "Tandem anchoring: Informational and politeness effects of range offers in social exchange." In: *Journal of personality and social psychology* 108.2, p. 254.

Andreoni, James and B Douglas Bernheim (2007). *Social image and the 50-50 norm*. Tech. rep. mimeo.

Ausubel, Lawrence M, Peter Cramton, and Raymond J Deneckere (2002). "Bargaining with incomplete information". In: *Handbook of game theory with economic applications* 3, pp. 1897–1945.

Ausubel, Lawrence M and Raymond J Deneckere (1993). "Efficient sequential bargaining". In: *The Review of Economic Studies* 60.2, pp. 435–461.

Bajari, Patrick et al. (2015). "Machine learning methods for demand estimation". In: *The American Economic Review* 105.5, pp. 481–485.

Bardolet, David, Craig R Fox, and Dan Lovallo (2011). "Corporate capital allocation: A behavioral perspective". In: *Strategic Management Journal* 32.13, pp. 1465–1483.

Bardsley, Nicholas et al. (2010). "Explaining Focal Points: Cognitive Hierarchy Theory versus Team Reasoning*". In: *The Economic Journal* 120.543, pp. 40–79.

Behrman, Jere R and Mark R Rosenzweig (2004). "Parental allocations to children: New evidence on bequest differences among siblings". In: *Review of Economics and Statistics* 86.2, pp. 637–640.

Belloni, Alexandre, Daniel Chen, et al. (2012). "Sparse models and methods for optimal instruments with an application to eminent domain". In: *Econometrica* 80.6, pp. 2369–2429.

Belloni, Alexandre and Victor Chernozhukov (2009). "Least squares after model selection in high-dimensional sparse models". In:

Binmore, Ken and Larry Samuelson (2006). "The evolution of focal points". In: *Games and Economic Behavior* 55.1, pp. 21–42.

Blanco, Mariana, Dirk Engelmann, and Hans Theo Normann (2011). "A within-subject analysis of other-regarding preferences". In: *Games and Economic Behavior* 72.2, pp. 321–338.

Bolton, Gary E and Axel Ockenfels (2006). "Inequality aversion, efficiency, and maximin preferences in simple distribution experiments: comment". In: *The American economic review*, pp. 1906–1911.

Bradley, Andrew P (1997). "The use of the area under the ROC curve in the evaluation of machine learning algorithms". In: *Pattern recognition* 30.7, pp. 1145–1159.

Buelens, Marc and Dirk Van Poucke (2004). "Determinants of a negotiator's initial opening offer". In: *Journal of Business and Psychology* 19.1, pp. 23–35.

Cai, Hongbin and Joseph Tao-Yi Wang (2006). "Overcommunication in strategic information transmission games". In: *Games and Economic Behavior* 56.1, pp. 7–36.

Camerer, Colin F et al. (1993). "Cognition and framing in sequential bargaining for gains and losses". In: *Frontiers of game theory* 104, pp. 27–47.

Charness, Gary and Matthew Rabin (2002). "Understanding social preferences with simple tests". In: *The Quarterly Journal of Economics* 117.3, pp. 817–869.

Chmura, Thorsten et al. (2005). "Testing (beliefs about) social preferences: Evidence from an experimental coordination game". In: *Economics Letters* 88.2, pp. 214–220.

Cramton, Peter (1984). "Bargaining with incomplete information: An infinite-horizon model with two-sided uncertainty". In: *The Review of Economic Studies* 51.4, pp. 579–593.

Crawford, Vincent P (2003). "Lying for strategic advantage: Rational and boundedly rational misrepresentation of intentions". In: *American Economic Review*, pp. 133–149.

Croson, Rachel, Terry Boles, and J Keith Murnighan (2003). "Cheap talk in bargaining experiments: lying and threats in ultimatum games". In: *Journal of Economic Behavior & Organization* 51.2, pp. 143–159.

Daniel, Terry E, Darryl A Seale, and Amnon Rapoport (1998). "Strategic play and adaptive learning in the sealed-bid bargaining mechanism". In: *Journal of Mathematical Psychology* 42.2, pp. 133–166.

De Bruyn, Arnaud and Gary E Bolton (2008). "Estimating the influence of fairness on bargaining behavior". In: *Management Science* 54.10, pp. 1774–1791.

Einav, Liran and Jonathan Levin (2014). "Economics in the age of big data". In: *Science* 346.6210, p. 1243089.

El Harbi, Sana et al. (2015). "Efficiency, equality, positionality: What do people maximize? Experimental vs. hypothetical evidence from Tunisia". In: *Journal of Economic Psychology* 47, pp. 77–84.

Embrey, Matthew, Kyle B Hyndman, and Arno Riedl (2014). "Bargaining with a residual claimant: An experimental study". In: *Available at SSRN 2502891*.

Engelmann, Dirk and Martin Strobel (2004). "Inequality aversion, efficiency, and maximin preferences in simple distribution experiments". In: *American economic review*, pp. 857–869.

– (2006). "Inequality aversion, efficiency, and maximin preferences in simple distribution experiments: Reply". In: *The American economic review*, pp. 1918–1923.

Fehr, Ernst, Michael Naef, and Klaus M Schmidt (2006). "Inequality aversion, efficiency, and maximin preferences in simple distribution experiments: Comment". In: *The American economic review*, pp. 1912–1917.

Fehr, Ernst and Klaus M Schmidt (1999). "A theory of fairness, competition, and cooperation". In: *The quarterly journal of economics* 114.3, pp. 817–868.

Fischbacher, Urs (2007). "z-Tree: Zurich toolbox for ready-made economic experiments". In: *Experimental economics* 10.2, pp. 171–178.

Forsythe, Robert, John Kennan, and Barry Sopher (1991). "An experimental analysis of strikes in bargaining games with one-sided private information". In: *American Economic Review* 81.1, pp. 253–78.

Friedman, Daniel and Ryan Oprea (2012). "A continuous dilemma". In: *The American Economic Review*, pp. 337–363.

Fudenberg, Drew, David Levine, and Jean Tirole (1985). "Infinite-horizon models of bargaining with one-sided incomplete information". In: *Game-theoretic models of bargaining*. Ed. by Alvin E. Roth. Cambridge: Cambridge University Press. Chap. 5, pp. 73–98.

Gachter, Simon and Arno Riedl (2006). "Dividing Justly in Bargaining Problems With Claims: Normative Judgments and Actual Negotiations". In: *Social Choice and Welfare* 27.3, pp. 571–594.

Gächter, Simon and Arno Riedl (2005). "Moral property rights in bargaining with infeasible claims". In: *Management Science* 51.2, pp. 249–263.

Galeotti, Fabio, Maria Montero, and Anders Poulsen (2015). "Efficiency Versus Equality in Bargaining". In: *Available at SSRN 2688599*.

Galinsky, Adam D and Thomas Mussweiler (2001). "First offers as anchors: the role of perspective-taking and negotiator focus." In: *Journal of personality and social psychology* 81.4, p. 657.

Grossman, Sanford J and Motty Perry (1986). "Sequential bargaining under asymmetric information". In: *Journal of Economic Theory* 39.1, pp. 120–154.

Gul, Faruk and Hugo Sonnenschein (1988). "On delay in bargaining with one-sided uncertainty". In: *Econometrica: Journal of the Econometric Society*, pp. 601–611.

Güth, Werner, Steffen Huck, and Peter Ockenfels (1996). "Two-level ultimatum bargaining with incomplete information: An experimental study". In: *The Economic Journal*, pp. 593–604.

Güth, Werner and Eric Van Damme (1998). "Information, strategic behavior, and fairness in ultimatum bargaining: An experimental study". In: *Journal of Mathematical Psychology* 42.2, pp. 227–247.

Haller, Hans and Steinar Holden (1990). "A letter to the editor on wage bargaining". In: *Journal of Economic Theory* 52.1, pp. 232–236.

Hanley, James A and Barbara J McNeil (1982). "The meaning and use of the area under a receiver operating characteristic (ROC) curve." In: *Radiology* 143.1, pp. 29–36.

Hargreaves Heap, Shaun, David Rojo Arjona, and Robert Sugden (2014). "How Portable Is Level-0 Behavior? A Test of Level-k Theory in Games With Non-Neutral Frames". In: *Econometrica* 82.3, pp. 1133–1151.

Herreiner, Dorothea K and Clemens Puppe (2004). *Equitable Allocations in Experimental Bargaining Games: Inequality A version versus Efficiency*. Tech. rep. Bonn econ discussion papers.

Holmström, Bengt and Roger B Myerson (1983). "Efficient and durable decision rules with incomplete information". In: *Econometrica: Journal of the Econometric Society*, pp. 1799–1819.

Isoni, Andrea et al. (2013a). "Efficiency, Equality and Labelling: An Experimental Investigation of Focal Points in Explicit Bargaining". In:

– (2013b). "Focal points in tacit bargaining problems: Experimental evidence". In: *European Economic Review* 59, pp. 167–188.

Jacquemet, Nicolas and Adam Zylbersztejn (2014). "What drives failure to maximize payoffs in the lab? A test of the inequality aversion hypothesis". In: *Review of Economic Design* 18.4, pp. 243–264.

Janssen, Maarten CW (2001). "Rationalizing focal points". In: *Theory and Decision* 50.2, pp. 119–148.

– (2006). "On the strategic use of focal points in bargaining situations". In: *Journal of Economic Psychology* 27.5, pp. 622–634.

Johnson, Eric J et al. (2002). "Detecting failures of backward induction: Monitoring information search in sequential bargaining". In: *Journal of Economic Theory* 104.1, pp. 16–47.

Kagel, John H, Chung Kim, and Donald Moser (1996). "Fairness in ultimatum games with asymmetric information and asymmetric payoffs". In: *Games and Economic Behavior* 13.1, pp. 100–110.

Kagel, John H and Katherine Willey Wolfe (2001). "Tests of fairness models based on equity considerations in a three-person ultimatum game". In: *Experimental Economics* 4.3, pp. 203–219.

Kalai, Ehud and Meir Smorodinsky (1975). "Other solutions to Nash's bargaining problem". In: *Econometrica: Journal of the Econometric Society*, pp. 513–518.

Kennan, John (1986). "The Economics of Strikes". In: *Handbook of Labor Economics, Vol. 2*. Chapter 19. Elsevier, pp. 1091–1137.

Kennan, John and Robert Wilson (1990). "Can strategic bargaining models explain collective bargaining data?" In: *The American Economic Review*, pp. 405–409.

Kennan, John and Robert B Wilson (1993). "Bargaining with private information". In: *Journal of Economic Literature* 31, pp. 45–45.

Krajbich, Ian et al. (2009). "Using neural measures of economic value to solve the public goods free-rider problem". In: *Science* 326.5952, pp. 596–599.

Kriss, Peter H, Rosemarie Nagel, and Roberto A Weber (2013). "Implicit vs. explicit deception in ultimatum games with incomplete information". In: *Journal of Economic Behavior & Organization* 93, pp. 337–346.

Kristensen, Henrik and Tommy Gärling (1997). "The effects of anchor points and reference points on negotiation process and outcome". In: *Organizational Behavior and Human Decision Processes* 71.1, pp. 85–94.

Kritikos, Alexander and Friedel Bolle (2001). "Distributional concerns: equity-or efficiency-oriented?" In: *Economics Letters* 73.3, pp. 333–338.

López-Pérez, Raúl, Ágnes Pintér, and Hubert János Kiss (2013). *Does Payoff Equity Facilitate Coordination? A test of Schelling's Conjecture*. Tech. rep. IEHAS Discussion Papers.

Mason, Malia F et al. (2013). "Precise offers are potent anchors: Conciliatory counteroffers and attributions of knowledge in negotiations". In: *Journal of Experimental Social Psychology* 49.4, pp. 759–763.

Menchik, Paul L (1980). "Primogeniture, equal sharing, and the US distribution of wealth". In: *The Quarterly Journal of Economics*, pp. 299–316.

Mitzkewitz, Michael and Rosemarie Nagel (1993). "Experimental results on ultimatum games with incomplete information". In: *International Journal of Game Theory* 22.2, pp. 171–198.

Mullainathan, Sendhil (2014). "Machine Learning Applications for Economics". In: *The Royal Economic Society annual conference, Hahn lecture*.

Myerson, Roger B (1979). "Incentive compatibility and the bargaining problem". In: *Econometrica* 47.1, pp. 61–73.

– (1984). "Two-person bargaining problems with incomplete information". In: *Econometrica: Journal of the Econometric Society*, pp. 461–487.

Nash Jr, John F (1950). "The bargaining problem". In: *Econometrica: Journal of the Econometric Society*, pp. 155–162.

Nydegger, Rudy V and Guillermo Owen (1974). "Two-person bargaining: An experimental test of the Nash axioms". In: *International Journal of game theory* 3.4, pp. 239–249.

Ochs, Jack and Alvin E Roth (1989). "An experimental study of sequential bargaining". In: *American Economic Review* 79.3, pp. 355–384.

Pruitt, Dean G (2013). *Negotiation behavior*. Academic Press.

Radner, Roy and Andrew Schotter (1989). "The sealed-bid mechanism: An experimental study". In: *Journal of Economic Theory* 48.1, pp. 179–220.

Rapoport, Amnon, Terry E Daniel, and Darryl A Seale (1998). "Reinforcement-based adaptive learning in asymmetric two-person bargaining with incomplete information". In: *Experimental Economics* 1.3, pp. 221–253.

Rapoport, Amnon, Ido Erev, and Rami Zwick (1995). "An experimental study of buyer-seller negotiation with one-sided incomplete information and time discounting". In: *Management Science* 41.3, pp. 377–394.

Rapoport, Amnon and Mark A Fuller (1995). "Bidding strategies in a bilateral monopoly with two-sided incomplete information". In: *Journal of Mathematical Psychology* 39.2, pp. 179–196.

Roth, Alvin E (1985). "Toward a focal point theory of bargaining". In: *Game-theoretic models of bargaining*, pp. 259–268.

Roth, Alvin E and Michael W Malouf (1979). "Game-theoretic models and the role of information in bargaining." In: *Psychological review* 86.6, p. 574.

Roth, Alvin E, Michael WK Malouf, and J Keith Murnighan (1981). "Sociological versus strategic factors in bargaining". In: *Journal of Economic Behavior & Organization* 2.2, pp. 153–177.

Roth, Alvin E and J Keith Murnighan (1982). "The role of information in bargaining: An experimental study". In: *Econometrica: Journal of the Econometric Society*, pp. 1123–1142.

Roth, Alvin E, J Keith Murnighan, and Françoise Schoumaker (1988). "The deadline effect in bargaining: Some experimental evidence". In: *The American Economic Review*, pp. 806–823.

Rubinstein, Ariel (1982). "Perfect equilibrium in a bargaining model". In: *Econometrica: Journal of the Econometric Society*, pp. 97–109.

– (1985). "A bargaining model with incomplete information about time preferences". In: *Econometrica: Journal of the Econometric Society*, pp. 1151–1172.

Schelling, Thomas C (1960). "The strategy of conflict". In: *Cambridge, Mass*.

Smith, Alec et al. (2014). "Neural Activity Reveals Preferences without Choices". In: *American Economic Journal: Microeconomics* 6.2, pp. 1–36.

Srivastava, Joydeep (2001). "The role of inferences in sequential bargaining with one-sided incomplete information: Some experimental evidence". In: *Organizational behavior and human decision processes* 85.1, pp. 166–187.

Ståhl, Ingolf (1972). "Bargaining theory". In:

Stone, Mervyn (1974). "Cross-validatory choice and assessment of statistical predictions". In: *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 111–147.

Thompson, Leigh L, Jiunwen Wang, and Brian C Gunia (2010). "Negotiation". In: *Annual review of psychology* 61, pp. 491–515.

Tibshirani, Robert (1996). "Regression shrinkage and selection via the lasso". In: *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288.

Valley, Kathleen et al. (2002). "How communication improves efficiency in bargaining games". In: *Games and Economic Behavior* 38.1, pp. 127–155.

Van Poucke, Dirk and Marc Buelens (2002). "Predicting the outcome of a two-party price negotiation: Contribution of reservation price, aspiration price and opening offer". In: *Journal of Economic Psychology* 23.1, pp. 67–76.

Varian, Hal R (2014). "Big data: New tricks for econometrics". In: *The Journal of Economic Perspectives* 28.2, pp. 3–27.

Wang, Joseph Tao-yi, Michael Spezio, and Colin F Camerer (2010). "Pinocchio's pupil: using eyetracking and pupil dilation to understand truth telling and deception in sender-receiver games". In: *The American Economic Review* 100.3, pp. 984–1007.

White, Sally Blount and Margaret A Neale (1994). "The role of negotiator aspirations and settlement expectancies in bargaining outcomes". In: *Organizational Behavior and Human Decision Processes* 57.2, pp. 303–317.

White, Sally Blount, Kathleen L Valley, et al. (1994). "Alternative models of price behavior in dyadic negotiations: Market prices, reservation prices, and negotiator aspirations". In: *Organizational Behavior and Human Decision Processes* 57.3, pp. 430–447.

Young, H Peyton and Mary A Burke (2001). "Competition and custom in economic contracts: a case study of Illinois agriculture". In: *American Economic Review*, pp. 559–573.

Yukl, Gary A (1974). "Effects of situational variables and opponent concessions on a bargainer's perception, aspirations, and concessions." In: *Journal of Personality and Social Psychology* 29.2, p. 227.

Chapter 5

CONCLUSION

Over the past few decades, research in the field of Behavioral Economics has led to extensive documentation of systematic biases from the "rational" benchmark in human decision-making. Many biases persist in seemingly efficient market environments, and endure even when incentives are high and learning is beneficial. These findings have challenged many core assumptions underlying traditional economic models, and their impact has been far reaching.

An important catalyst of Behavioral Economics research was the development of experimental methodologies for quantifying and describing human behavior in the laboratory. In this sense, behavioral economics' role in advancing the understanding of how humans make decisions is akin to the role that Psychophysics had in the progress of vision research. Much of what we know about vision today had emerged from extensive documentation of perceptual biases in carefully controlled laboratory settings, along with the development of metrics and experimental methodologies in the field of Psychophysics.

Behavioral Economics offers a descriptive level of analysis, which is an essential step towards understanding the phenomena it studies. For the case of vision research, much of the impetus that followed required two additional levels of analysis, originally suggested by David Marr (Marr and Poggio, 1976). In Marr's framework, Psychophysics served as the "computational" level of analysis, mapping the correspondence between a visual environment and a psychological representation. Psychophysics was combined with an "algorithmic" level analysis, aiming to reverse-engineer the process that generates the mapping in a formal mathematical fashion, and an "implementation" level analysis designed to test the feasibility of the algorithmic hypotheses using biological and neural data.

We are now in the point in history when our understanding of human decision making can benefit from combining the three levels of analysis together in a similar fashion. Technological and methodological developments allow measuring, manipulating and modeling human behavioral and biological variables, enriching the descriptive mapping between stimulus and response. Non-choice measurements, such as brain activity, eye tracking, skin conductance, response times, facial expres-

sions, hormonal levels, and more, allow generalizing Marr's framework to studying the mechanisms underlying economics choice in a tri-level analysis:

Computational: finding the correspondence between stimulus and actions, as traditionally studied by experimental and behavioral economists.

Algorithmic: reverse-engineering the cognitive processes that might generate the observed computational correspondence, taking into account biological aspects (e.g., computational constraints, processing speed) and the type of problems that the human brain has honed to solve over the coarse of evolution.

Implementation: testing the feasibility of the algorithmic hypothesis, using biological and neural models and data.

The current dissertation demonstrates the potential contribution of this tri-level framework. The first chapter illuminates a well-documented decision bias, extrapolative belief formation. In contrast to the traditional economic approach of deriving predictions that are based on ad-hoc axiomatic statements (Rabin, 2000), I propose a neurally plausible algorithm (Yu and Cohen, 2009), which theoretical predictions closely match the behavioral data. The two other chapters further demonstrate how biological variables (hormonal levels) and non-choice measurements (bargaining process data), that are traditionally ignored by economists, can be used for predicting meaningful economic outcomes. I am hopeful that this work will set the grounds for further investigations at the algorithmic level, that will shed further light on how these "hidden" variables influence welfare loss due to inaccurate responses (e.g., incorrect CRT answers) and costly disagreements in bargaining.

Pasadena, California, May 2016.

BIBLIOGRAPHY

Marr, David and Tomaso Poggio (1976). “From understanding computation to understanding neural circuitry”. In:

Rabin, Matthew et al. (2000). *Inference by believers in the law of small numbers*. Institute of Business and Economic Research.

Yu, Angela J and Jonathan D Cohen (2009). “Sequential effects: superstition or rational behavior?” In: *Advances in neural information processing systems*, pp. 1873–1880.

INDEX