



UvA-DARE (Digital Academic Repository)

On the role of information in strategic and individual decision making

Ioannidis, K.

Publication date

2024

Document Version

Final published version

[Link to publication](#)

Citation for published version (APA):

Ioannidis, K. (2024). *On the role of information in strategic and individual decision making*. [Thesis, fully internal, Universiteit van Amsterdam].

General rights

It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations

If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: <https://uba.uva.nl/en/contact>, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.



tinbergen
institute

GRADUATE PROGRAM |



On the role of information in strategic and individual decision making

Konstantinos Ioannidis

On the role of information in strategic and individual decision making

ISBN: 978 90 361 0754 9

Cover design: Crasborn Graphic Designers bno, Valkenburg a.d. Geul

This book is no. **846** of the Tinbergen Institute Research Series, established through cooperation between Rozenberg Publishers and the Tinbergen Institute. A list of books which already appeared in the series can be found in the back.

On the role of information in strategic and individual decision making

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam op
gezag van de Rector Magnificus

prof. dr. ir. P.P.C.C. Verbeek

ten overstaan van een door het College voor Promoties ingestelde commissie,
in het openbaar te verdedigen in de Aula der Universiteit van Amsterdam
op vrijdag 17 mei 2024, te 14:00 uur door

Konstantinos Ioannidis

geboren te Kozani

Promotiecommissie

Promotores

prof. dr. T.J.S. Offerman

Universiteit van Amsterdam

prof. dr. R. Sloof

Universiteit van Amsterdam

Overige Leden

prof. dr. H.J. Holm

Lund University

dr. S. Dominguez Martinez

Universiteit van Amsterdam

prof. dr. F.A.A.M. van Winden

Universiteit van Amsterdam

prof. dr. J. Hinlopen

Universiteit van Amsterdam

prof. dr. A.J.H.C. Schram

Universiteit van Amsterdam

Faculteit Economie en Bedrijfskunde

Acknowledgement of financial support

The research for this doctoral thesis received financial assistance from the Research Priority Area Behavioral Economics of the University of Amsterdam, and from the Amsterdam Center for Behavioural Change of the University of Amsterdam.

Acknowledgements

When I decided back in 2015 to apply to Tinbergen Institute, I knew little about economics. I was terrified to start a PhD program in a new field, but I was fascinated by everything I was about to learn. There were times when reaching the end of this long (honestly really long) journey felt impossible. I am grateful to the University of Amsterdam, where most of my PhD was developed, to the University of Birmingham where I spend a year and developed the last chapter of this thesis, and to the University of Cambridge and the Leverhulme International Professorship in Neuroeconomics where the thesis was completed.

First and foremost, I want to thank my supervisors Theo and Randolph. It is hard to put in words how much I appreciate both of you. You taught me so much about how to think of interesting research questions, how to even approach writing a theory paper, how to sharpen my experimental designs, but what I am most grateful for is how you helped me grow as a person and as a researcher. In one of our first meetings you said that in the first chapter you will guide me a lot, in the second chapter you will guide me a little bit, and on the next chapters you hope to proudly watch me succeed on my own. When I shared with you the first draft of my job market paper (chapter 2 of this thesis), you replied that you were (positively) surprised. This was the moment that any doubt on whether I could reach the end of the journey disappeared; so I hope indeed I made you proud.

The first two years at Tinbergen Institute were intense. I first want to thank the cohort before us for helping us navigate our first year (special note to Huyen, Jenny, and Vadim). Stephan, you deserve a special mention as you were always there as a friend, an office-mate, a table football teammate, and a headbanger at metal festivals! Looking back to those years, assignments and exams have faded, and I am left with fond memories of pub quizzes, chess games, dinners, and parties. You guys got me a bike when I was too stubborn to get one in my first months in Amsterdam, a Nintendo switch so I can remain a kid at heart, and even sent me to Wacken! You are all unbelievable! Thank you for all we have enjoyed together and I hope to see you again soon. Asli and Johan (my paranymphs) we started this journey together and we were there for each other all the way, I could not have done this without you!

I also want to thank the CREED family. I always felt I could knock on any-

one's door and expect to be welcomed. Attending lunch seminars and observing the constructive feedback on experimental designs (either of my own or of colleagues) taught me so much. To overcome the fear of missing anyone out, I would like to thank every CREEDer that made the past few years so memorable. Special thanks to the postdocs at the time, Ivan (my *Una Faccia*, *Una Razza* teammate when we were crowned CREED table football champions in the Christmas party of 2017) and Margarita (you are always welcome to knock my door and interrupt me, I am a quick flight away!).

As I mentioned earlier, this journey was a long one, and extended beyond my time in Amsterdam. Moving to a new country, taking up new roles, learning new skills, while working on my PhD on evenings and weekends was quite intense. I cannot thank Hamideh and Moumita enough for being the sweetest friends and making my time in Birmingham so much more pleasant. A special massive thank you goes to Peter Bossaerts who easily topped any gift I have ever received by making me an offer to join Cambridge the night before my wedding!

The desire to aim for a PhD started when I was very young. A massive thank you goes to my parents, Apostolos and Efthimia, and my sister Kiki, who supported me in every possible way all those years until I reached where I am today. I think of you all every single day of living abroad chasing my dreams. Roula, you always dreamt of studying Mathematics, but eventually studied Economics, and you said that your dream of combining the two was fulfilled through me; I wish you were here to share this moment.

Finally, I want to thank my biggest fan, my most patient supporter, my rival from team Mystic, and the love of my life, Tina. You were with me every step of the way, in good times, in stressful times, even when changing countries, and I could always count on you! You indirectly went through a whole PhD process with me, and still have the bravery and ambition to go for one yourself (good luck!). I am grateful to have met you and I look forward to our adventures together.

*Konstantinos Ioannidis
Cambridge, April 2024*

Contents

Introduction

1	Verifiability approach	1
1.1	Introduction	2
1.2	Lie detection and the Verifiability Approach	6
1.3	Baseline model	8
1.4	Equilibrium analysis	12
1.4.1	Perfect Bayesian equilibria	12
1.4.2	Improvements in the investigation technology and valuable information revelation	21
1.4.2.1	The effect of information on the judge's equilibrium payoffs	21
1.4.2.2	Comparative statics in the verification technology	23
1.5	Model extensions	26
1.5.1	Plea bargaining	26
1.5.2	Right to silence	29
1.6	Conclusion	32
1.7	Appendix to Chapter 1	34
1.7.1	Proof of Proposition 2	34
2	Habitual communication	37
2.1	Introduction	38
2.2	Design & Predictions	43
2.2.1	The sender-receiver game	43
2.2.2	Perfect Bayesian equilibria of the sender-receiver game	44
2.2.3	Treatments	45
2.2.4	Predictions	46
2.2.5	Procedure	47
2.3	Results	49
2.3.1	Manipulation check: Differences in behaviour in part one	50

2.3.2	Treatment effects: Comparing communication after aligned vs conflict	51
2.3.3	Overcommunication and undercommunication	53
2.3.4	Habit formation and inattention at the individual level	54
2.4	Concluding remarks	58
2.5	Appendix to Chapter 2	60
2.5.1	Additional results and robustness checks	60
2.5.1.1	All Bayesian equilibria of the game	60
2.5.1.2	Replicating past cheap-talk experimental findings	62
2.5.1.3	Econometric tests for treatment effects	63
2.5.1.4	Econometric tests for overcommunication and undercommunication	64
2.5.1.5	Full classification of behavioural strategies	65
2.5.1.6	Robustness of habitual classification with respect to threshold	68
2.5.2	Payoff tables for different values of bias parameter	68
2.5.3	Experimental instructions	69
3	Anchoring and markets	77
3.1	Introduction	78
3.2	Experimental design and implementation	79
3.3	Results	83
3.3.1	Anchoring manipulation	83
3.3.2	Market effect on valuations	85
3.4	Concluding discussion	88
3.5	Appendix to Chapter 3	93
3.5.1	Experimental instructions	93
4	Whistleblowing under competition	103
4.1	Introduction	104
4.2	Experimental design and predictions	106
4.2.1	The baseline whistleblowing game	106
4.2.2	Treatments	107
4.2.3	Predictions	108
4.2.4	Post-experiment survey	112
4.2.5	Implementation	113
4.3	Results	114
4.3.1	The effect of competition on whistleblowing and lawbreaking	114
4.3.2	The effect of beliefs and morality judgements on behaviour	115
4.4	Concluding discussion	118

4.5	Appendix to Chapter 4	119
4.5.1	Additional treatments and exploratory results	119
4.5.1.1	Behaviour between participant pools	119
4.5.1.2	Robustness treatments	120
4.5.2	Instructions and decision screens	123

Summaries

Bibliography

Introduction

This thesis contains four chapters reporting on four different research projects. The chapters can be read independently of each other. A common theme of all the chapters of this thesis is an interest in understanding how information affects the way people make decisions. The first two chapters study a strategic setting between a sender who has *information* that a receiver would like to find out. The first chapter provides a game theoretic analysis whereas the second chapter uses an experimental approach. The third chapter studies whether irrelevant *information* presented to people affects how they value goods, and whether market interactions reduce that effect. The last chapter studies the decision to reveal *information* about corporate fraud via whistleblowing.

Chapter 1 is motivated by recent lie detection methods which are admissible as scientific evidence in courts. Among such methods, a very promising one is the Verifiability Approach. The method is based on the premise that truth-tellers provide many precise details that can be verified, whereas liars will provide many imprecise details to avoid being exposed. We develop a game-theoretic model between a suspect who wants to be acquitted, and a judge who wants to reach a correct verdict. We model the verifiability approach as a costly signal that the judge can obtain about the veracity of the statement of the suspect; a signal whose accuracy depends on the precision of the statement. We further assume that producing a precise false statement is cognitively costly, and that if the statement is falsified, an additional penalty is imposed on the suspect.

We provide an equilibrium analysis that allows us to pin down the conditions under which the interaction provides valuable information to the judge about the likelihood that the suspect is innocent or guilty. If providing a precise false statement is feasible, then the best the judge can achieve is a partially-separating equilibrium where the judge investigates precise statements often enough so that guilty suspects are deterred away from always producing them. The condition under which this equilibrium exists becomes easier to meet if the investigation becomes more accurate or if the penalty after a statement that was found to be false increases. After satisfying the condition, further increasing this penalty does not result in more valuable information being revealed.

Chapter 2 is closely related to the strategic communication between a better-informed sender and a lesser-informed receiver. A common finding in the literature is that senders reveal more information and receivers trust information more, compared to a benchmark where both senders and receivers are self-interested money-maximisers. We conjecture that this phenomenon may be attributed to habitual communication. If the majority of human communication happens between senders and receivers with common interests, then senders may form the habit of telling the truth, and receivers may form the habit of believing what they hear. Similarly, if the majority of human communication happens between senders and receivers with conflicting interests, then senders may form the habit of lying, and receivers may form the habit of distrusting what they hear.

We provide a two-stage experiment in which participants play a sender-receiver game in which the sender is informed about the state of the world, and can send a message to the receiver, who does not know the state of the world but always wants to correctly infer it. We vary whether in the first stage the senders are incentivised to truthfully reveal the state of the world, or to convince the receiver that the state of the world is much higher than it actually is. In the second stage, participants interact in an environment where the sender is incentivised to convince the receiver that the state of the world is mildly higher than it actually is. Our primary interest is to test whether habits from the first stage affect communication in the unfamiliar environment in the second stage.

We find that habits do affect strategic communication. If in stage one senders and receivers interact in the common-interest environment, we find that they overcommunicate in stage two; a finding in line with previous experimental literature. If in stage one senders and receivers interact in the conflicting-interest environment, we find that they undercommunicate in stage two; a novel finding. We also vary how frequently participants interact in the unfamiliar environment, and find that habits affect communication only when the unfamiliar environment is infrequent. We interpret this as evidence that habits shape our attention rather than our preferences.

Chapter 3 combines both individual-decision and strategic-decision settings. We study anchoring, which is a cognitive bias positing that people incorporate irrelevant pieces of information when making valuations about goods. In an experiment, we first elicit participants' valuation of a good after first asking them whether they value it more or less than a randomly determined price. The random price is determined by die rolls and aims to provide the irrelevant information. Next, participants interact in a market, and we are interested in whether exposure to markets mitigates the anchoring effects on valuations. We find that our anchoring manipulation failed as the initial valuation was unaffected by the random price. We provide a concise meta-analysis suggesting that anchoring is less likely to emerge if the anchor is trans-

parently uninformative.

Chapter 4 is related to whistleblowing on corporate fraud. While the literature on cartel formation has provided evidence that leniency programs increase the propensity of firms to report a cartel, the experimental evidence on the effect of analogous interventions, such as monetary rewards or protection from retaliation, on whistleblowing by employees of a firm has primarily come from settings where firms operate independently. We conjecture that competition for market revenue provides strategic and non-strategic motivations against whistleblowing. We use an experiment with two treatments, with and without competition, and find an insignificant reduction of whistleblowing under competition. We also find evidence that behaviour correlates with beliefs, but it does not correlate with morality judgements.

Chapter 1

Verifiability Approach

1.1 Introduction

After decades of research on lying detection, psychologists have recently made a breakthrough in revealing who is lying. The early literature focused on the idea that liars can be identified by facial microexpressions of emotions and other unintentional behaviours. In two meta-analyses, DePaulo et al. (2003) and Bond Jr and DePaulo (2006) showed that nonverbal cues of lying are weak and unreliable. A typical finding is that approximately 54% of examiners' judgements are correct, only slightly better than chance (50%). One important reason why non-verbal cues are unreliable is that liars try to mimic the expressions of truth tellers when they become aware of which cues are used by investigators. For example, Ekman et al. (1988) have shown that truth tellers often smile when they express genuine positive feelings and that liars mimic them by also smiling. The challenge that examiners then face is that they have to distinguish between fake and genuine smiles.

The breakthrough involves recent methods of lie detection which focus on the content of what is being said. In the Verifiability Approach (VA), the examiner judges a statement based on the presence and frequency of verifiable details. VA exploits a dilemma that liars face. Liars have an incentive to include verifiable details in their statement, because detailed accounts are more likely to be believed (Bell and Loftus, 1989). At the same time, presenting specific details is risky because it makes it easier for the examiner to check a statement (Nahari et al., 2014a). Truth-tellers typically do not have this dilemma and can reveal as many verifiable details as possible. The relative frequency of verifiable details in a statement may then become an informative signal of its truth. Using VA, examiners' judgements are correct in approximately 70% of the cases (Vrij, 2018).¹ Moreover, in contrast to the nonverbal cues, the accuracy of VA is enhanced when interviewees are made aware of it. Doing so results in truth tellers adding more verifiable details to their statement than liars do (Nahari et al., 2014a; Harvey et al., 2017).

In this chapter we provide a game-theoretic foundation for the strategic effect that underlies VA and explore the potential interaction among its main drivers. Our analysis takes into account the cognitive costs of fabricating precise but false statements, the higher reliability of verifying detailed (rather than vague) statements, as well as the potential use of penalties for 'obstruction of the investigation process' that VA may allow for. The main focus is on how these different elements jointly affect the strategic trade-off liars face and contribute to precise statements becoming an informative signal per se (even without being actually verified).

Our model considers a speaker who wants to convince an investigator that he

¹Vrij (2018) provides an elaborate discussion of the state-of-the-art methods in lying detection. Besides VA, he discusses six prominent methods; see the next section for a brief overview. Among all these methods, VA stands out because of its success and the ease with which it is implemented.

is innocent and an investigator who pursues the truth. Applications of this type of strategic interaction abound. A mother may want to find out if her son is using drugs; a parole officer is interested to know if an offender lives up to the agreement made; an airport officer wants to find out if a passenger is carrying dangerous items; an insurance company wants to find out whether a claim was rightly made; an employer interviews an applicant (and potentially verifies references) to learn whether he has been thorough and truthful in drafting his CV; a judge questions a suspect to assess whether he is guilty. Throughout the chapter we use labels that correspond to the judge-suspect example for ease of illustration. A suspect is privately informed about whether he is guilty or innocent. The judge has already collected some evidence that furnishes a prior belief about whether the suspect is guilty. The suspect is asked to make a statement about what happened. He either makes a precise statement that includes verifiable and distinctive details, or a vague statement. Providing a false precise statement is assumed to be cognitively costly. After listening to the suspect, the judge can decide to reach a verdict immediately or to check the statement at some cost. Checking a precise statement gives a more reliable signal than checking a vague statement does. If the judge convicts the suspect after his statement was checked and falsified, an additional obstruction of justice penalty is imposed on the suspect (Decker, 2004). The suspect always wants to be acquitted whereas the judge wants to reach a correct verdict. Moreover, she (weakly) prefers to wrongly acquit a guilty suspect over wrongly convicting an innocent one.

We derive all perfect Bayesian equilibria of the game and identify the conditions under which either full or partial information revelation occurs in equilibrium (and when this information is truly beneficial). A separating equilibrium exists only when the cognitive costs of lying are prohibitively high, such that guilty types always refrain from making a precise statement. The research on VA has identified interviewing techniques that can increase the cognitive load of lying.² Moreover, experimental evidence shows that deception can sometimes be (probabilistically) detected even without possibilities for ex-post verification and consequences for lying (Jupe et al., 2017). Nevertheless, full separation may be hard to achieve in actual practice given the high stakes liars typically have in hiding the truth. In that case, lower (but positive) cognitive lying costs may still enable a partial pooling equilibrium in which the guilty suspect mixes between being precise and remaining vague (and an innocent suspect always makes a precise statement). However, in order for this information to be truly valuable to the judge and increase her expected payoffs, the possibility of actual verification then plays a key role. In particular, valuable information trans-

²For example, asking surprise questions or requesting a narrative in reverse chronological order have been shown to be successful in Vrij et al. (2007) and Sorochinski et al. (2014). Other methods such as the Sheffield Lie Test exploit the fact that lying takes time and that response times can be used to distinguish truth tellers from liars (Suchotzki et al., 2017).

mission then necessarily requires that the guilty type is deterred away from always lying if the judge would *always* verify a precise statement. If this 'deterrence-by-verification condition' is not met, equilibria may still exist in which information is revealed via either the statement or the investigation, but the judge never gains in terms of expected payoffs relative to reaching a verdict immediately based on the prior belief.

Larger cognitive lying costs, a higher reliability of verification and a higher obstruction penalty all contribute to meeting the deterrence-by-verification condition. If indeed met, a partially pooling equilibrium exists in which the guilty suspect mixes between being precise and remaining vague and the judge only now and then verifies a precise statement (with a vague statement leading to immediate conviction). The judge then effectively has two sources of information complementing each other: the strategic behaviour of the suspect (i.e. the statement *per se*) and the outcome of the occasional verification. Within this equilibrium, increasing either the cognitive lying costs or the obstruction penalty further does not increase the provision of valuable information. What the judge can learn from the suspect's statement *per se* remains unaffected, because guilty suspects keep on lying with the same frequency. The amount of valuable information obtained via verification is actually reduced, because higher lying costs or a higher obstruction penalty induces the judge to investigate less. Improvements in the verification technology that make it more reliable continue to have a beneficial impact, however, because a higher accuracy does induce the guilty type to lie less often. Interestingly, although all else equal the improved accuracy would by itself have led to more valuable information obtained via verification as well, in equilibrium it actually leads to less. The main drivers here are that precise statements are made less often by the guilty type and (therefore) also verified less often by the judge. Hence, if the verification technology becomes more accurate, the additional benefits that come with it are purely due to the deterrence effect of the potential verification. Finally, a decrease in the investigation costs has similar effects on the amount and source of valuable information obtained in the partial pooling equilibrium as an improved reliability has; driven by the deterrence effect again more information is obtained from the strategic behaviour of the suspect itself and less from the actual verification of messages.

The overall upshot of our analysis is that especially improvements in the accuracy of the verification technology are beneficial. Even if the guilty type is willing to always lie, such improvements make actual verification that may occur in a pooling (on precise) equilibrium more cost-effective. And as soon as the threshold that deters the guilty type from always lying is met, such improvements enlarge the deterrence effect. As a result, the guilty type reveals more information via the statement *per se* and the actual verification process itself actually yields less valuable information.

Higher lying costs or a higher obstruction penalty play a supporting role in meeting the relevant deterrence threshold. But once this threshold is met, they do not facilitate further valuable information transmission.

We extend our analysis in two ways. First, we allow the suspect to confess to receive a penalty reduction. In that case the guilty type no longer provides a vague statement in equilibrium and mixes between mimicking the innocent type by providing a precise statement and confessing instead. The equilibrium analysis is essentially equivalent as for the baseline model, with the single difference that the lying costs should now be enlarged with the opportunity costs of not taking the opportunity to confess. A penalty reduction after confession thus complements the cognitive lying costs and the obstruction penalty in facilitating more informative equilibria and in fact represent two sides of the same coin.³ That is, to reduce mimicking of the innocent type one can either make it more costly via the lying costs or the obstruction penalty, or less attractive via the penalty reduction. Second, we also consider the case in which the suspect has a ‘right to silence’. In that case silence cannot be held against the suspect, effectively restricting the judge’s choice of action in case the suspect refrains from making a precise statement. Such a right to silence may alter, but does not eliminate, strategic information revelation by the suspect and thus neither its complementary role to direct verification of statements.⁴ Moreover, for an intermediate range of prior beliefs the lying costs and the obstruction penalty no longer play a supportive role, reinforcing that especially a higher reliability is advantageous.

The remainder of this chapter is organised as follows. Section 1.2 briefly discusses various lie detection methods that have received attention in the psychology literature and the verifiability approach in particular. It also discusses how we account for the key features of this approach in our theoretical analysis. Section 1.3 presents the setup of our baseline model. In Section 1.4 we derive the set of perfect Bayesian equilibria. Additionally, we discuss how the amount of (valuable) information revelation and the effective reliance on the different information sources varies with the characteristics of the verification technology. In Section 1.5 we consider two extensions of the baseline setup: the possibility of plea bargaining and accounting for a right to silence. Here we also discuss the connection with earlier game-theoretic analyses of the latter two aspects within the law and economics literature. Section 1.6 summarises the chapter and concludes.

³As such, our chapter relates to earlier game-theoretic analyses of plea bargaining; see Grossman and Katz (1983), Reinganum (1988), Baker and Mezzetti (2001), Bjerck (2007), Kim (2010) and Tsur (2017). When discussing plea bargaining in Subsection 1.5.1, we make this connection (as well as the differences) with our model more precise.

⁴The right to silence has been analysed from a game-theoretic perspective by Seidmann and Stein (2000), Seidmann (2005), Mialon (2008) and Leshem (2010). In Subsection 1.5.2 we discuss the insights from these studies within the context of our model.

1.2 Lie detection and the Verifiability Approach

The origins of deception detection research can be traced back to Zuckerman et al. (1981) who categorised emotion, arousal, control and cognitive processing as four different cues to deception. Various methods were developed over the years which were based on the first three of these cues. The methods focused on non-verbal behaviour, compared levels of arousal between liars and truth-tellers and did not intervene in the information gathering process. In a meta-analysis, DePaulo et al. (2003) showed that those methods were not reliable as the observed behaviours showed no direct links to deception. According to Vrij (2019), to overcome those issues, modern research in deception detection has made three major shifts. Modern methods (1) focus on the content of a statement, (2) take into account the cognitive process behind lying and (3) have developed interview protocols to optimise the information gathering process. Some of these are already admissible as evidence in courts in countries like the United States, Germany and the Netherlands (Vrij, 2008).

Vrij (2018) provides an elaborate discussion of the state-of-the-art methods in deception detection. He compares the seven most prominent methods in terms of how ready they are to be applied in judicial systems. The list of methods includes Criteria Based Content Analysis (CBCA), Reality Monitoring (RM), Scientific Content Analysis (SCAN), Cognitive Credibility Assessment (CCA), Strategic Use of Evidence (SUE), the Verifiability Approach (VA) and Assessment Criteria Indicative of Deception (ACID). He does so on the basis of 14 criteria, which can be grouped into two sets: academic, such as whether the method has been tested and whether it has been subjected to peer review, as well as procedural, such as whether it is easy to use and whether it provides an information gathering protocol. Five of those criteria, also known as the Daubert standard, are the minimal requirements for scientific evidence to be admissible in US courts (*Daubert v. Merrell Dow Pharmaceuticals, Inc.*, 1993).⁵ Of the seven methods, only three abide by the Daubert standard, namely RM, ACID and VA. However, ACID is not easy to incorporate in interviews and RM does not provide a within-subject measure of truthfulness. Hence, our chapter models VA as the investigation mechanism available to the judge.

As the name suggests, VA is based on the verifiability of details. A detail is considered verifiable if it describes an activity experienced with an identifiable person or witnessed by an identifiable person or recorded through technology (Nahari et al., 2014a). Based on the finding that lying is cognitively more demanding (Vrij et al., 2017), there exist interviewing techniques which aim to magnify the cognitive task

⁵The full list of criteria for admissibility of scientific evidence in US courts is: i) Has the technique been tested in actual field conditions (and not just in a laboratory)?; ii) Has the technique been subject to peer review and publication?; iii) What is the known or potential rate of error?; iv) Do standards exist for the control of the technique's operation?; and v) Has the technique been generally accepted within the relevant scientific community?

for liars. On the one hand, the interviewer asks the interviewee to include as many details as possible. On the other hand, the interviewee would like to avoid mentioning details that can easily be checked by the interviewer. Balancing those orthogonal incentives, one would expect a liar to provide many non-verifiable details in a statement. The ratio of verifiable over non-verifiable details is a within-subject measure of the probability that a statement is true or fabricated. Additional benefits of VA are the fact that it is robust to countermeasures (Nahari et al., 2014b) and that VA scoring could be computer-automatised as suggested by Kleinberg et al. (2016).

To capture the essential element of VA within a simple model of strategic information transmission, we extend an otherwise standard sender-receiver game with an (bilaterally) endogenous verification technology. On the disclosing side, the sender can choose between various statements that differ directly in their costs and indirectly in their degree of verifiability. More precise statements in principle allow for a more reliable investigation than less precise – i.e. more ‘vague’ – statements do. At the same time, coming up with a precise false statement is cognitively costly. On the receiving side, the receiver subsequently decides whether to indeed investigate the actual statement made (at same costs) or not. Lying – i.e. fabricating a precise but false statement as to mislead the receiver – is clearly possible. However, with cognitive costs and with potential verification, the sender’s statement is not pure cheap talk. From a modelling perspective our chapter thus fits within the broader theoretical literature on strategic communication with either intrinsically costly (cf. Kartik (2009)), or detectable deceit (cf. Holm (2010); Balbuzanov (2019); Dziuda and Salas (2018); Ispano and Vida (2021)). Our setup differs, among other things, in verification being costly and at the receiver’s discretion.

Our chapter is motivated by the above discussed findings from psychology that lying is cognitively more demanding. We model this psychological component both explicitly as a direct lying cost, as well as via the implied reliability of the investigation technology and its relationship to the sender’s statement and underlying type. Glazer and Rubinstein (2012) provide a model of strategic persuasion in which a speaker is boundedly rational in the sense that she uses the truth as an anchor for cheating. In particular, when fabricating a false set of answers to a given questionnaire, the speaker starts from the truth and tries to modify her answers to satisfy the listener’s preset acceptance conditions. Modifications are limited to adapting the consequence of originally violated ‘if-then’ conditions.⁶ As a result, truth-tellers al-

⁶Glazer and Rubinstein (2012) provide experimental evidence that supports these two key ingredients (truth anchor and the specific type of modifications considered) of bounded rationality. In Glazer and Rubinstein (2014) they study a related setup in which the speaker does not know the exact set of acceptance conditions, but can make inferences about these from the observed earlier acceptance decisions of the listener. Here bounded rationality is captured by limitations on the complexity of the regularities that the speaker can detect in the acceptance data. By making the questionnaire sufficiently complex, the listener can almost completely eliminate successful cheating by liars. A different micro-foundation for using complex interview protocols rooted in bounded rationality is provided by

ways get their way, whereas liars – given their truth anchor and modification limits – may not be able to satisfy the listener’s ‘codex’. Similar to other recent articles on detectable deceit (Dziuda and Salas (2018); Balbuzanov (2019)), we simply incorporate this element directly by allowing for (endogenous) variation in lie detection probabilities. The main focus is then on the implications for the amount of strategic information revelation that results.

Key within VA is the *verifiability of distinctive details*. We label statements that contain many such verifiable details as ‘precise’ and statements with none or only a very few of these as ‘vague’. This labelling does not necessarily coincide with using precise or vague language though.⁷ In linguistics a term is considered vague if it exhibits borderline cases. For instance, there are no clear cut bounds on the number of grains that define a ‘heap’ of sand (O’Connor, 2014). Within our setting, clear statements that might nevertheless be hard to verify (like ‘I was home alone sleeping in my bed’) are considered vague. Similarly so are essentially empty statements that are (almost) tautologically true, like ‘I was on planet earth’. (Note that such statements are also not cognitively demanding to fabricate.) Key difference between a precise and a vague statement in our setup is the larger extent to which the former provides a convincing alibi when verified as well as reason for serious suspicion if falsified, i.e. its distinctiveness.

1.3 Baseline model

Although the strategic interaction that we model arguably matches various real life applications (cf. the Introduction), for concreteness we describe it in terms of the interaction between a suspect (speaker) and a judge (investigator). Assume a crime has been committed and a suspect (he) is being questioned. The judge (she) can use the statement of the suspect to update her beliefs on his innocence. She can do so immediately or after conducting a costly investigation that can, with some commonly known error probability, verify or falsify the statement. The suspect wants to be acquitted and the judge wants to reach a correct verdict, viz. to acquit innocent suspects and to convict guilty ones. Additionally, the judge prefers to acquit a guilty suspect over convicting an innocent one.

We note up front that our conceptualisation of the interaction between a suspect

Jehiel (2021). He analyses multi-round cheap talk communication assuming liars have more limited memory than truth-tellers have. The liar’s fear of issuing inconsistent statements over time can then be exploited to facilitate information revelation.

⁷In cheap talk experiments, messages such as “The true value of a variable belongs to set S ” are labelled precise if S is a singleton and vague otherwise. Using this definition, vague language has been shown to increase efficiency in experiments involving public good games with hidden value (Serra-Garcia et al., 2011) and coordination games with multiple equilibria (Agranov and Schotter, 2012). Without the possibility to verify messages before taking an action, i.e. when messages are pure cheap talk, both type of messages would be considered vague in our setup.

and a judge is based on a number of simplifications. In real life, a suspect can get arrested by the police, provide a statement and a prosecutor may decide whether to file charges and bring the case to court or not. If she does so, the suspect becomes a defendant and may provide additional testimony during the trial. All evidence is examined by a judge and/or a jury and once a verdict is reached, the judge imposes a penalty or not. In our reduced form model we have condensed the timing, the actors and the type of information provided. We use ‘judge’ as label for a representative of the judicial system with the understanding that in practice some actions described in the model might be taken by prosecutors or the jury.⁸ Essentially, our model assumes that at some point during the entire judicial process, the suspect will be asked to provide some information. The untruthfulness of this information is assumed to have consequences for the sentence the suspect may be facing, if he gets convicted.

Our model corresponds to a sender-receiver game, where the sender is the suspect and the receiver is the judge. The suspect knows his own type (T), that is whether he is innocent ($T = I$) or guilty ($T = G$). The type of the suspect is unknown to the judge, but she holds a commonly known prior belief of $b = Pr(T = I)$ that the suspect is innocent. These prior beliefs can be interpreted as the evidence collected by the judge before questioning the suspect, so that in principle she can convict (or acquit) without requesting a statement.

The suspect can choose between two actions. He can choose to answer all the questions,⁹ which results in a precise statement ($S = P$), or he can choose to refrain from providing clear and distinctive answers, which results in a vague statement ($S = V$). Providing a precise, but false statement is cognitively costly. After seeing the statement, the judge must reach a verdict to acquit (A) or convict (C) the suspect. This decision can be taken either before or after having investigated (I) the statement made.

The investigation mechanism works as follows. If the judge decides to investigate statement $S \in \{V, P\}$, the investigation mechanism provides an outcome that has a probability of r_S of being correct (which means verified for the statement of the innocent type and falsified for the statement of the guilty type) and a probability of $1 - r_S$ of being wrong (which means falsified for the statement of the innocent type and verified for the statement of the guilty type). Parameters r_V and r_P thus reflect the reliability of investigating the various statements. We assume that the investigation mechanism has at least some informational value, in the sense that it

⁸Assuming a unitary actor for the judicial system is an arguably reasonable simplification to the extent that the various actors within the judiciary share the same preferences and information. We briefly return to this in Subsection 1.5.1 where we discuss the possibility of plea bargaining and relate our strategic setup to existing models of plea bargaining in the literature.

⁹An implicit assumption in the model is that when answering questions, an innocent suspect tells the truth whereas a guilty suspect lies. Allowing both of them to choose whether to answer truthfully or not is a possible extension for future research.

gets the judge closer to the truth. This assumption translates to both probabilities r_V and r_P being larger than $\frac{1}{2}$.¹⁰ Aligned with the psychology literature on content-based deception detection methods, we also assume that the differences in content between the statement of the innocent and the guilty type will be more pronounced in a more detailed statement (Harvey et al., 2017). With a precise statement, the judge then gets better clues exactly what to look for, allowing her to steer her investigation in a more promising direction. As a result, investigating a precise statement is more likely to produce a correct outcome than investigating a vague one, i.e. we assume that investigation probabilities satisfy $\frac{1}{2} < r_V < r_P < 1$.

Preferences of the two suspect types depend on both the statement they provide and on the decision of the judge. To capture that fabricating a detailed lie might be cognitively costly to the suspect, we assume that the guilty type suffers a direct lying cost equal to $\lambda_P \geq 0$ when providing a precise but false statement. Making a vague statement does not entail any cognitive costs, however, and neither does telling the truth in full detail via a precise statement for the innocent type. The choices of the judge affect the payoffs of both suspect types in the following way. Both suspect types get a payoff of 1 if they get acquitted. If they get convicted, they receive a lower payoff which depends on the amount of evidence that resulted in their conviction. If they get convicted on the basis of prior evidence, which happens when the judge does not investigate the statement or when investigation verifies the statement and provides no additional evidence against them, they receive a payoff of 0 (so the imposed sentence leads to a payoff reduction of 1). If they get convicted after their statement S was investigated and falsified, they receive an additional obstruction penalty π_S , for which we assume that $0 \leq \pi_V \leq \pi_P$.¹¹ We also assume that this obstruction penalty is only applied when a suspect is eventually convicted.¹²

The preferences of the judge are modelled in the following way. The judge gets 1 for reaching a correct verdict, that is to acquit an innocent suspect and to convict a guilty suspect. In case the judge makes a mistake, she receives a lower payoff that depends on the type of mistake made. We normalise the payoff of acquitting a guilty suspect to 0 and set the payoff of convicting an innocent suspect to $-\alpha$. The assumption that $\alpha \geq 0$ captures the notion that the judge (weakly) prefers to let go

¹⁰The distinctiveness of verifiable statement S can be inversely captured by the odds ratio $\frac{1-r_S}{r_S}$ of verification being unreliable. This ratio ranges from 0 for $r_S = 1$ (maximal distinctiveness), to 1 for $r_S = \frac{1}{2}$.

¹¹This penalty can be interpreted in various ways. If the lying was under oath, then the defendant may be charged with perjury (US Sentencing Commission, 2018, §2J1.3.). If the lying significantly impeded official investigation, then the defendant may be charged with obstruction of justice (US Sentencing Commission, 2018, §3C1.1.). The sentencing guidelines also recommend a reduction of penalty if a defendant provided substantial assistance in the investigation, for example by giving truthful, complete and reliable testimony (see §5K1.1.). In this case, the penalty can be interpreted as the difference between the full and the reduced sentence.

¹²A prosecutor will very often drop a criminal charge if it is determined that the evidence against the accused is not strong enough, see Cohen (1992).

of a guilty suspect over sending an innocent suspect to jail. Higher values of α result in a tighter threshold on the judge's belief for her to prefer conviction over acquittal; it thereby essentially quantifies exactly what is meant by "beyond any reasonable doubt" and sets the standard of proof. In particular, with these payoffs the (updated) belief that the suspect is innocent should exceed the tipping point of $\frac{1}{2+\alpha}$ for the judge to acquit.¹³

Besides obtaining the above payoffs the judge has to pay a positive cost $c > 0$ when she investigates statement S . These costs not only reflect that investigating the truthfulness of statements is costly in terms of the resources needed (time and detectives), but may also capture other, more indirect types of costs. For instance, in criminal cases of high importance that receive widespread public attention, society often really disapproves of cases that last for years, so our cost parameter could also be seen as pressure to reach a verdict faster. Note that our assumptions regarding the judge's payoffs arguably make these largely aligned with what society would seem to require. Her expected payoffs could thus potentially serve as a first approximation to a more encompassing welfare analysis.

To simplify the exposition we finally assume that the innocent type always provides a precise statement. Telling the truth – and in full detail if asked to do so – comes as a default to innocent people who have no incentive to lie (Verschuere and Shalvi, 2014), and innocent people even waive their right to remain silent due to their belief that their truth will shine (Kassin and Norwick, 2004). In an earlier version of this chapter we did not make this simplifying assumption and analysed the model assuming that the innocent type is also a strategic agent who endogenously chooses between a vague and a precise statement as well. The single notable difference is that in that case additional equilibria may exist alongside the other equilibria in which both suspect types always provide a vague statement. These pooling equilibria generally do not survive standard equilibrium refinements based on payoff dominance or on restricting out-of-equilibrium beliefs, like e.g. the divinity concept of Banks and Sobel (1987). Our simplifying assumption essentially solves the multiplicity of equilibria issue in a simpler way without losing much nuance.

Figure 1.1 provides a succinct summary of the order of moves in the strategic interaction between the suspect and the judge and Table 1.1 summarises the payoffs of all agents.

¹³Given beliefs b , the judge's expected payoffs from acquit equal $1 \cdot b + 0 \cdot (1 - b) = b$ and thus increase with b , while the expected payoffs from convict equal $1 \cdot (1 - b) - \alpha \cdot b = 1 - (1 + \alpha)b$ and decrease with b . At $b = \frac{1}{2+\alpha}$ these expected payoffs coincide.

Figure 1.1: Timeline of the game

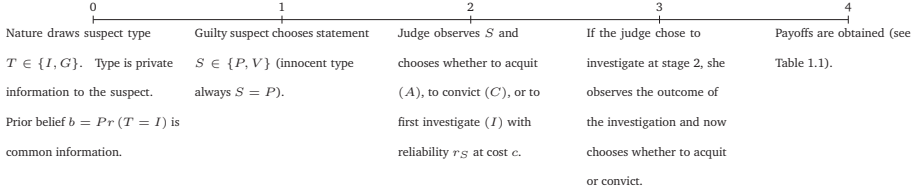


Table 1.1: Payoffs of suspect and judge for all type-action combinations

	Convict		Acquit	
	Suspect	Judge	Suspect	Judge
<i>Innocent type:</i>				
Precise w/out verification	0	$-\alpha$	1	1
Precise with verification	$-\pi_P$	$-\alpha - c$	1	$1 - c$
<i>Guilty type:</i>				
Vague w/out verification	0	1	1	0
Vague with verification	$-\pi_V$	$1 - c$	1	$-c$
Precise w/out verification	$-\lambda_P$	1	$1 - \lambda_P$	0
Precise with verification	$-\lambda_P - \pi_P$	$1 - c$	$1 - \lambda_P$	$-c$

Note: By assumption $\alpha \geq 0$, $c > 0$, $0 \leq \pi_V \leq \pi_P$ and $\lambda_P \geq 0$.

1.4 Equilibrium analysis

1.4.1 Perfect Bayesian equilibria

Besides her prior belief, the judge in principle has two information sources available: investigation (at cost c) of the actual statement made by the subject and the potentially different strategies the two types of suspects employ in making statements. In this section we explore the extent to which these different information sources are actually drawn upon in equilibrium and how they interact, by providing an encompassing (perfect Bayesian) equilibrium analysis.

For the judge the main goal of the entire process is to get a better idea of whether the suspect is guilty or not. Given the assumptions made, a vague statement can only be coming from a guilty suspect and, consequently, leads to immediate conviction without further costly investigation. Starting from a prior belief b that the suspect is innocent, after seeing a precise statement the judge updates her initial belief based on the strategic behaviour of the suspect. Let p denote the probability that the guilty type gives a precise statement. Using Bayes' rule, a rational judge then updates her

belief that the suspect is innocent to:

$$b^P \equiv Pr(T = I | S = P) = \frac{b}{b + (1 - b)p} \quad (1.1)$$

Note that $b \leq b^P \leq 1$. The more the guilty suspect lies, i.e. the higher p , the closer the posterior belief is to the prior. Likewise, the less the guilty suspect lies, the closer the posterior gets to 1.

Having seen a precise statement, the judge convicts, investigates, and acquits with respective probabilities q_C , q_I and q_A . (As noted, after a vague statement the judge convicts for sure given the assumptions made.) In case the judge investigates, she obtains additional information that allows her to update her beliefs another time, based on the outcome of the investigation. From the given reliability of the investigation process and again Bayes' rule, we immediately obtain that these beliefs equal:

$$b^{P+} \equiv Pr(T = I | S = P \text{ and verified}) = \frac{b^P r_P}{b^P r_P + (1 - b^P)(1 - r_P)} \quad (1.2)$$

$$b^{P-} \equiv Pr(T = I | S = P \text{ and falsified}) = \frac{b^P(1 - r_P)}{b^P(1 - r_P) + (1 - b^P)r_P} \quad (1.3)$$

From these expressions, together with $r_P > \frac{1}{2}$, it follows that $b^{P-} \leq b^P \leq b^{P+}$. Falsification of the statement made by the suspect thus lowers the judge's belief that he is innocent, while a verified statement increases this belief.

Because investigating a precise statement is costly to her, the judge is willing to do so only if it yields her valuable information. That is, the information received should be *influential*; the judge's optimal decision whether to acquit or convict should (strictly) vary with the outcome of the investigation process.¹⁴ Otherwise the judge could better immediately opt for the decision she would in the end take anyway and avoid costly investigation altogether. Recall from the previous section that the tipping point (in terms of beliefs) for the judge to prefer acquit over convict equals $\frac{1}{2+\alpha}$. Influential information thus requires that updated beliefs satisfy $b^{P-} < \frac{1}{2+\alpha} < b^{P+}$, such that the judge acquits when the suspect's precise statement is verified and convicts when the precise statement is falsified.¹⁵ Lemma 1 details this requirement in terms of the posterior belief b^P .

Lemma 1. *Investigating a precise statement is influential iff: $\frac{1-r_P}{\alpha r_P + 1} < b^P < \frac{r_P}{\alpha(1-r_P) + 1}$.*

¹⁴Note that the notion of the investigation being *influential* is stronger than that it being *informative*. The latter holds as long as the outcome of the investigation is more likely to be aligned with the truth, which is guaranteed by our assumption that $\frac{1}{2} < r_V < r_P$. Sobel (2020) provides an insightful discussion of the differences between the definitions of informative and influential.

¹⁵When either $b^{P-} = \frac{1}{2+\alpha} < b^{P+}$ or $b^{P-} < \frac{1}{2+\alpha} = b^{P+}$, the judge would be indifferent between acquit and convict after either falsification or verification, respectively. In both cases (which cannot happen simultaneously), the judge essentially always weakly prefers either acquit or convict, irrespective of the outcome of the investigation; she thus would not be willing to invest $c > 0$ in it. That is why we require the optimal outcome to *strictly* vary with the outcome of the investigation.

In that case the judge would acquit if a precise statement were to be verified and convict if a precise statement were to be falsified.

Proof. Investigating a precise statement is influential as long as $b^{P-} < \frac{1}{2+\alpha} < b^{P+}$. Using expressions (1.2) and (1.3) for b^{P+} and b^{P-} above and rewriting immediately gives the result. \square

Intuitively, investigation can be influential only if, after having just heard a precise statement and correctly inferring the suspect's strategic behaviour (in particular, probability p with which a guilty suspect makes such a statement), the judge is still insufficiently confident about the suspect's type. That is, she is neither sufficiently convinced that the suspect is guilty (b^P is not very low), nor sufficiently convinced that the suspect is innocent (b^P is neither very high).

Obtaining influential information is a necessary requirement for the judge to investigate, yet it is not a sufficient. The expected benefits from the influential information received should also outweigh the costs of investigation c . Lemma 2 precisely characterises this requirement and pins down the judge's optimal choice for any posterior belief $b^P \in [0, 1]$ she might have.

Lemma 2. *Define lower and upper belief thresholds as $\underline{b}(r_P, c; \alpha) \equiv \min \left\{ \frac{(1-r_P)+c}{\alpha r_P+1}, \frac{1}{2+\alpha} \right\}$ and $\bar{b}(r_P, c; \alpha) \equiv \max \left\{ \frac{r_P-c}{\alpha(1-r_P)+1}, \frac{1}{2+\alpha} \right\}$. Moreover, let $\hat{c}(r_P; \alpha) \equiv \frac{1+\alpha}{2+\alpha} \cdot (2r_P - 1)$ respectively. After a precise statement and based on updated belief b^P , the judge's optimal choice of action equals:*

- (a) convict if $b^P < \underline{b}(r_P, c; \alpha)$;
- (b) investigate if $b^P \in (\underline{b}(r_P, c; \alpha), \bar{b}(r_P, c; \alpha))$;
- (c) acquit if $b^P > \bar{b}(r_P, c; \alpha)$.

The interval $(\underline{b}(r_P, c; \alpha), \bar{b}(r_P, c; \alpha))$ is non-empty and equals $\left(\frac{(1-r_P)+c}{\alpha r_P+1}, \frac{r_P-c}{\alpha(1-r_P)+1} \right)$ iff $c < \hat{c}(r_P; \alpha)$. In that case the judge is indifferent between convict and investigate if $b^P = \underline{b}(r_P, c; \alpha)$, and indifferent between investigate and acquit if $b^P = \bar{b}(r_P, c; \alpha)$. If $c > \hat{c}(r_P; \alpha)$ and thus $\underline{b}(r_P, c; \alpha) = \bar{b}(r_P, c; \alpha) = \frac{1}{2+\alpha}$, the judge is indifferent between convict and acquit when $b^P = \frac{1}{2+\alpha}$.

Proof. For updated beliefs b^P that the suspect is innocent, immediate acquittal after a precise statement yields the judge b^P in expected payoffs while immediate conviction yields her $1 - (1 + \alpha)b^P$ in expectation. Acquittal thus dominates conviction iff $b^P > \frac{1}{2+\alpha}$. Given that an investigation is costly ($c > 0$), the judge is only willing to do so if it is influential (cf. Lemma 1); it then leads to an expected payoff of $r_P - b^P(1 - r_P)\alpha - c$. This exceeds the payoff of convicting if $b^P > \frac{(1-r_P)+c}{\alpha r_P+1}$ and the one of acquitting if $b^P < \frac{r_P-c}{\alpha(1-r_P)+1}$. For these thresholds it holds that $\frac{(1-r_P)+c}{\alpha r_P+1} \leq \frac{1}{2+\alpha}$ and $\frac{r_P-c}{\alpha(1-r_P)+1} \geq \frac{1}{2+\alpha}$

iff $c \leq \hat{c}(r_P; \alpha)$. Hence, if $c < \hat{c}(r_P; \alpha)$, the interval $(\underline{b}(r_P, c; \alpha), \bar{b}(r_P, c; \alpha))$ is non-empty and in this range the judge prefers investigation. \square

The belief interval where costly investigation pays off collapses when the verification is completely inaccurate. Put differently, the break-even cost threshold equals $\hat{c}(r_P; \alpha) = 0$ for $r_P = \frac{1}{2}$. Recall from the Introduction that non-verbal deception detection methods are almost indistinguishable from chance as their accuracy is close to 50%. Thus, relying on such methods, while costly, does not facilitate information revelation. Verbal deception detection methods can achieve higher accuracy which benefits the judge. Intuitively, the range of beliefs $b^P \in (\underline{b}(r_P, c; \alpha), \bar{b}(r_P, c; \alpha))$ for which investigation pays off widens if the verification process becomes more reliable, i.e. when r_P increases, and when investigation becomes cheaper (lower c). If non-empty, the interval always contains the tipping point $\frac{1}{2+\alpha}$ between convicting and acquitting. The further away beliefs b^P are from this point of indifference, the more confident the judge is to solely act on the basis of the existing evidence – i.e., the prior belief and the statements per se – and to skip costly investigation altogether.

Turning to the guilty type of suspect, in equilibrium he chooses a best response to the judge's anticipated behaviour. Our next lemma characterises his optimal choice of statement when he anticipates that the judge responds with (q_A, q_I, q_C) to a precise statement.

Lemma 3. Define $\hat{\lambda}(r_P, \pi_P) \equiv 1 - r_P(1 + \pi_P)$. If the judge chooses (q_A, q_I, q_C) in response to a precise statement, the guilty type's optimal choice of statement equals:

- (a) a precise one $S = P$ if $\lambda_P < q_A + q_I \cdot \hat{\lambda}(r_P, \pi_P)$;
- (b) a vague one $S = V$ if $\lambda_P > q_A + q_I \cdot \hat{\lambda}(r_P, \pi_P)$.

The guilty type is indifferent between a precise and vague statement iff $\lambda_P = q_A + q_I \cdot \hat{\lambda}(r_P, \pi_P)$.

Proof. With the judge's response (q_A, q_I, q_C) , the expected payoffs from choosing a precise statement equal $q_A \cdot 1 + q_I \cdot ((1 - r_P) \cdot 1 - r_P \cdot \pi_P) - \lambda_P = q_A + q_I \cdot \hat{\lambda}(r_P, \pi_P) - \lambda_P$. Choosing a vague statement leads to immediate conviction and thus payoffs equal to 0. Comparing these payoffs gives the result. \square

Providing a vague statement leads to immediate conviction and a payoff of 0. The guilty type is then only willing to make a precise statement if the cognitive costs of doing so are not prohibitively large compared to the expected benefits of a potentially more favourable decision (than conviction) such a statement might bring. The relevant threshold for λ_P thus depends on the judge's response to a precise statement. If the judge would always acquit ($q_A = 1$), a precise statement would yield the guilty type a payoff of 1. The expected benefits relative to the benchmark of

conviction (yielding 0) then equal 1. If the judge would always investigate ($q_I = 1$) after a precise statement, it would yield the guilty type an expected payoff equal to $(1 - r_P) - r_P\pi_P$. This expression follows because with probability $(1 - r_P)$ the guilty type gets away with his lie and is acquitted, while with the remaining probability r_P he is caught lying and, besides conviction, is imposed obstruction penalty π_P . The overall expected benefits from a precise statement then equal $\hat{\lambda}(r_P, \pi_P)$. Note that $\hat{\lambda}(r_P, \pi_P)$ falls short of $\frac{1}{2}$ given $r_P > \frac{1}{2}$ and decreases with both r_P and π_P (and becomes negative for π_P large). As Lemma 3 illustrates, for a general anticipated response from the judge the cost-benefit analysis for the guilty type compares the direct lying costs λ_P with the appropriately weighted average of the two relevant thresholds 1 and $\hat{\lambda}(r_P, \pi_P)$.

Based on the best responses in Lemma 2 and Lemma 3, *mutual* best responses – and thereby equilibrium outcomes – can now be intuitively understood. First observe from Lemma 3 that if $\lambda_P > 1$, the guilty type will choose a vague statement for sure. Put differently, if the cognitive costs of fabricating a false precise statement are prohibitively high, the guilty type necessarily chooses to willingly expose himself by making a vague statement. A precise statement then provides conclusive evidence that the suspect is innocent, inducing the judge to acquit for sure after such a statement. We thus immediately obtain a unique separating equilibrium in this case. In this separating equilibrium the strategic behaviour of the two suspect types is fully revealing and the judge always reaches the correct verdict, without the need to ever verify the statements made.

Arguably, in criminal cases the conditions for full separation are often not met. A guilty suspect may either be cognitively able to produce a detailed (but false) statement, or can afford the legal expertise to help him produce one. In those instances where $\lambda_P < 1$ and the guilty type in principle would be willing to provide a precise statement, completely revealing equilibria do not exist. In that case the evidence of the case as captured by the prior belief b determines the extent to which he actually will do so in equilibrium. Given that a precise statement always induces the judge to update her belief upwards (i.e. $b^P \geq b$ by equation (1.1)), making such a statement can always ensure acquittal if the prior belief would already do so. From Lemma 2 it follows that this happens when $b > \bar{b}(r_P, c; \alpha)$. In that case the guilty suspect can safely lie and completely get away with it. This yields a pooling equilibrium in which no additional information at all is obtained and the judge reaches a verdict purely based on her prior belief.

For completeness we formally describe these two – arguably unrealistic – ‘corner’ equilibria in the following proposition. Here we omit the choice of the innocent type as we have assumed he always provides a precise statement. We also omit the choice of the judge after a vague statement as we established earlier that a vague statement

leads to immediate conviction. Therefore, the equilibria are described with the probability that the guilty suspect provides a precise statement (p), the posterior belief of the judge after she observes such a precise statement (b^P), and the subsequent decision of the judge right after updating her beliefs (q_A, q_I, q_C).

Proposition 1. *Consider the case with either $\lambda_P > 1$ or $b > \bar{b}(r_P, c; \alpha)$. Then there exists a unique Perfect Bayesian equilibrium which is either separating (Sep) or pooling (Pool) and characterised as follows.¹⁶*

Sep *Suppose $\lambda_P > 1$. Then the guilty type always gives a vague statement and the judge always acquits after a precise statement. Formally: $p = 0, b^P = 1, q_A = 1$.*

Pool *Suppose $\lambda_P < 1$ and $b > \bar{b}(r_P, c; \alpha)$. Then the guilty type always gives a precise statement and the judge always acquits after a precise statement. Formally: $p = 1, b^P = b, q_A = 1$.*

Proof. If $\lambda_P > 1$, then $p = 0$ from Lemma 3. In turn, $b^P = 1$ from equation (1.1) and thus $q_A = 1$ from Lemma 2. This gives the separating equilibrium Sep. Next, let $\lambda_P < 1$. If $b > \bar{b}(r_P, c; \alpha)$, then $q_A = 1$ necessarily from Lemma 2 and $b^P \geq b$. From $\lambda_P < 1$ and Lemma 3, the guilty type's best response then equals $p = 1$. This yields equilibrium Pool. \square

If neither the cognitive costs nor the prior beliefs are that high and thus the conditions of Proposition 1 are not met, necessarily some but not all information is revealed in equilibrium.¹⁷ The amount of information revelation, as well as the information source effectively drawn upon in equilibrium, then depends on how the prior belief b and the characteristics of the investigation technology as reflected by parameters $(\lambda_P, c, r_P, \pi_P)$ compare to the relevant thresholds $\bar{b}(r_P, c; \alpha)$ and $\hat{c}(r_P; \alpha)$ from Lemma 2, and $\hat{\lambda}(r_P, \pi_P)$ from Lemma 3. For each distinct class of parameter combinations, Proposition 2 characterises the unique informative equilibrium that exists. The numbering of these equilibria reflects their desirability from the perspective of the judge (to which we return in the next subsection).

Proposition 2. *Consider the case with $\lambda_P < 1$ and $b < \bar{b}(r_P, c; \alpha)$. Then in the generically unique perfect Bayesian equilibrium the judge necessarily obtains some information beyond her prior beliefs b . This Informative equilibrium corresponds to one from the list below.*

¹⁶Here and in the sequel we focus on “generic” cases. In non-generic cases multiple equilibria may exist side by side. For instance, in the knife-edge case where $\lambda_P = 1$ (and $b > \bar{b}(r_P, c; \alpha)$) equilibria Sep and Pool co-exist.

¹⁷This follows because no information revelation would require that the guilty type always makes a precise statement (i.e. $p = 1$) such that the statement per se reveals no information and thus $b^P = b$. For $b < \bar{b}(r_P, c; \alpha)$ it then follows from Lemma 2 that the judge either convicts or investigates after a precise statement. The latter is incompatible with the judge not getting additional information beyond her prior. But if the judge would always convict after a precise statement, the guilty type would not be willing to bear the cognitive costs λ_P .

Inf.1 Suppose $c < \hat{c}(r_P; \alpha)$ and $\lambda_P > \hat{\lambda}(r_P, \pi_P)$. Then the guilty type mixes between a vague and a precise statement and a partially pooling equilibrium results. The judge mixes between acquit and investigate after a precise statement. Formally: $p = \frac{b}{1-b} \cdot \frac{(1-r_P)(1+\alpha)+c}{r_P-c}$, $b^P = \bar{b}(r_P, c; \alpha)$, $q_A = \frac{\lambda_P - \hat{\lambda}(r_P, \pi_P)}{1 - \hat{\lambda}(r_P, \pi_P)}$, $q_I = 1 - q_A$.

Inf.2 Suppose $c < \hat{c}(r_P; \alpha)$, $\lambda_P < \hat{\lambda}(r_P, \pi_P)$ and $b > \underline{b}(r_P, c; \alpha)$. Then the guilty type always gives a precise statement and a pooling equilibrium results. The judge always investigates after a precise statement. Formally: $p = 1$, $b^P = b$, $q_I = 1$.

Inf.3 Suppose $c < \hat{c}(r_P; \alpha)$, $\lambda_P < \hat{\lambda}(r_P, \pi_P)$ and $b < \underline{b}(r_P, c; \alpha)$. Then the guilty type mixes between a vague and a precise statement and a partially pooling equilibrium results. The judge mixes between investigate and convict after a precise statement. Formally: $p = \frac{b}{1-b} \cdot \frac{r_P(1+\alpha)-c}{1-r_P+c}$, $b^P = \underline{b}(r_P, c; \alpha)$, $q_I = \frac{\lambda_P}{\hat{\lambda}(r_P, \pi_P)}$, $q_C = 1 - q_I$.

Inf.4 Suppose $c > \hat{c}(r_P; \alpha)$. Then the guilty type mixes between a vague and a precise statement and a partially pooling equilibrium results. The judge mixes between acquit and convict after a precise statement. Formally: $p = \frac{b}{1-b} \cdot (1 + \alpha)$, $b^P = \frac{1}{2+\alpha}$, $q_A = \lambda_P$, $q_C = 1 - q_A$.

Proof. See subsection 1.7.1. □

For prohibitively high investigation costs $c > \hat{c}(r_P; \alpha)$, the interval of posterior beliefs for which the judge prefers to investigate a precise statement is empty; that is, $\underline{b}(r_P, c; \alpha) = \bar{b}(r_P, c; \alpha) = \frac{1}{2+\alpha}$ and $q_I = 0$ (cf. Lemma 2). In that case, when the prior belief favours conviction ($b < \frac{1}{2+\alpha}$), the guilty type necessarily uses a mixed strategy in equilibrium. This follows because always making a precise statement would induce a posterior belief equal to the prior, and thus a payoff $0 - \lambda_P \leq 0$. If instead the guilty type would always make a vague statement, the judge would acquit for sure after a precise statement given that then $b^P = 1$, providing the guilty type a strong incentive to deviate (given $\lambda_P < 1$ in the case considered here). The mixed strategy the guilty type employs in equilibrium makes the judge indifferent between convict and acquit after receiving a precise statement. Vice versa, the judge's equilibrium probability of acquittal equal to λ_P makes the guilty type indifferent between the two statements. This yields equilibrium Inf.4 in which some information is revealed only via the statements per se.

Only when the investigation costs are sufficiently low (i.e. $c < \hat{c}(r_P; \alpha)$), the judge may potentially want to investigate after a precise statement. If she would always do so, then by Lemma 3 the guilty type would be deterred from making a precise statement iff

$$\lambda_P > \hat{\lambda}(r_P, \pi_P) \quad (\equiv 1 - r_P(1 + \pi_P)) \quad (1.4)$$

If condition (1.4) holds, the guilty type is effectively deterred away from *always* lying. In equilibrium he then only does so occasionally (i.e. $0 < p < 1$) as to ensure the judge will mix between acquitting and investigating after a precise statement. This yields Inf.1 in which both information sources are drawn upon.

When condition (1.4) is not met, the guilty type prefers to make a precise statement even if such a statement would always be verified. If the guilty type would indeed always lie, the judge's posterior belief b^P equals prior b and always verifying a precise statement is a best response only in case the prior is intermediate, i.e. if $\underline{b}(r_P, c; \alpha) < b < \bar{b}(r_P, c; \alpha)$. This gives pooling equilibrium Inf.2 in which the judge only obtains additional information via investigation. If instead the prior is low and favours conviction ($b < \underline{b}(r_P, c; \alpha)$), the guilty type necessarily employs a mixed strategy. Loosely put, the best he can do is then to convince the judge to not always convict but occasionally investigate instead. This yields Inf.3.

In the partially pooling equilibria Inf.1 and Inf.3, the judge's two information sources complement each other. In both equilibria, strategic information revelation by the suspect allows the judge to update her belief that he is innocent upwards ($b^P > b$) after having received a precise statement. This induces her to now and then verify such a statement and, if she indeed does so, to acquit if verified and convict if falsified. The two equilibria differ in what happens if the judge does not investigate though: acquittal in case of Inf.1 and conviction in case of Inf.3. Therefore, while in Inf.3 a *verified* precise statement is necessary to get acquitted, in Inf.1 an *unchecked* precise statement already suffices. The comparative statics of how the equilibrium behaviour of the guilty type (i.e. p) and the judge (i.e. q_I) varies with the characteristics of the verification technology as reflected by λ_P , r_P , π_P and c , is also opposite in the two equilibria (cf. the next subsection).

Condition (1.4) intuitively captures the deterrence effect of the potential verification of precise statements and the verification technology more broadly. Investigation becomes a stronger threat the more reliable it is (higher r_p) and the higher the obstruction penalty π_P becomes. This deterrence effect comes on top of the cognitive costs λ_P of formulating a precise (but false) statement, effectively creating an interdependence between the two. The higher these cognitive costs, the lower r_P and π_P can be for condition (1.4) still to be met. Note, for instance, that even in the absence of an obstruction penalty ($\pi_P = 0$), the condition still holds as long as the direct lying costs are high enough: $\lambda_P > 1 - r_P (= \hat{\lambda}(r_P, 0))$. The cognitive costs thus play a supporting role for information revelation even when full separation cannot be achieved (i.e. when $\lambda_P < 1$). Also note from condition (1.4) that a high reliability r_P is by itself not a sufficient deterrent for the guilty type to refrain from always making a precise statement. It should be complemented with either sufficient cognitive costs of lying ($\lambda_P > 0$) or a sufficiently high obstruction penalty ($\pi_P > 0$) for the guilty

type to be discouraged to always mimic the innocent type.

An illustrative summary of the conditions under which each equilibrium arises is provided in tree form in Figure 1.2. The tree splits into separate branches with respect to how the values of the lying cost λ_P , the prior belief b and the investigation cost c compare to the relevant thresholds. For generic parameter values, there is a unique equilibrium outcome; the labels in the boxes just refer to the equilibria listed in Proposition 1 and Proposition 2. The tree is augmented with two additional columns: the information source drawn upon along the equilibrium path (statements per se or investigation) and the expected equilibrium payoffs of the judge. The latter are explained in the next subsection, where we explore in detail how the amount of (valuable) information revelation and the effective reliance on different information sources varies with the characteristics of the verification technology, as captured by parameters λ_P, r_P, π_P and c .

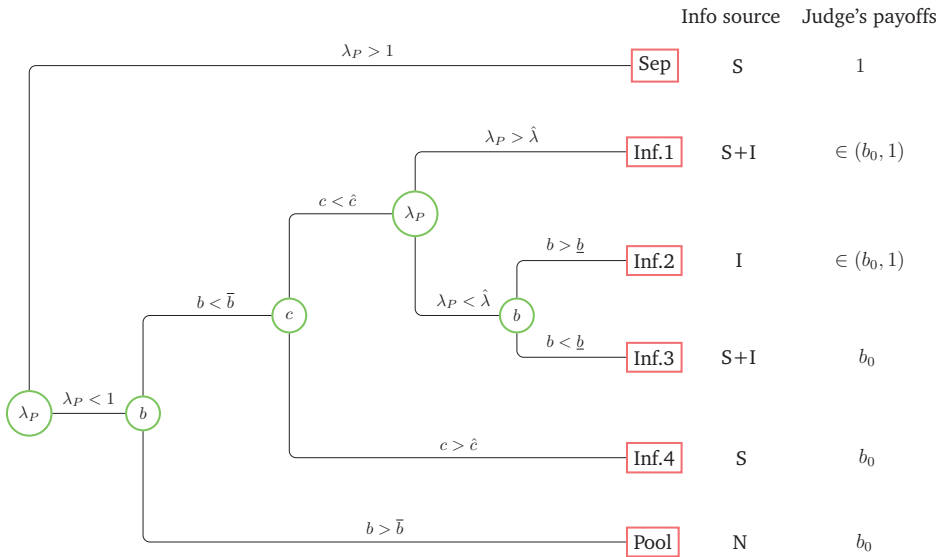


Figure 1.2: All equilibria with conditions for existence and payoffs to the judge

Info source: S=statement, I=investigation, S+I=statement+investigation, N=no information

Abbreviations: $\bar{b} = \bar{b}(r_P, c; \alpha)$, $\underline{b} = \underline{b}(r_P, c; \alpha)$, $\hat{c} = \hat{c}(r_P; \alpha)$, $\hat{\lambda} = \hat{\lambda}(r_P, \pi_P)$

1.4.2 Improvements in the investigation technology and valuable information revelation

1.4.2.1 The effect of information on the judge's equilibrium payoffs

The judge does not want to obtain just any information per se, but rather influential information that is instrumental to her decision. The effective value of such information can be inferred from how her payoffs are affected. In the absence of any additional information beyond her prior belief, the best the judge can achieve in terms of expected payoffs is

$$b_0 \equiv \max\{1 - b(1 + \alpha), b\},$$

i.e. the best from either convicting or acquitting for sure.¹⁸ Relative to this, getting additional information will always make her weakly better off. If the judge would always take the right decision (without bearing further investigation costs) she would get her maximum payoffs equal to 1. Proposition 3 ranks (for a given level of b) the payoffs of the judge in the various equilibria by comparing these with the lower bound b_0 and the upper bound of 1.

Proposition 3. *For the judge's equilibrium payoffs it holds that:*

- (a) *In Sep the judge's expected payoffs equal the upper bound of 1;*
- (b) *In Inf.1 and Inf.2 the judge's expected payoffs are strictly in between b_0 and 1. Holding prior belief b constant, the judge earns strictly more in Inf.1 than in Inf.2;*
- (c) *In Inf.3, Inf.4 and Pool the judge's expected payoffs equal the lower bound b_0 .*¹⁹

Proof. The equilibrium payoffs in Sep and thus part (a) follow immediately. In Pool the judge always acquits and obtains b in expected payoffs. This equals b_0 for the range $b > \bar{b}(r_P, c; \alpha) \geq \frac{1}{2+\alpha}$ where Pool exists. In equilibria Inf.3 and Inf.4 the judge chooses $q_C > 0$. Conviction is thus always a best response (for the given equilibrium behaviour p of the guilty type) and equilibrium payoffs for the judge coincide with those of always choosing conviction for sure, i.e. $1 - b(1 + \alpha)$. This corresponds to b_0 under the conditions of existence for these equilibria, which require $b < \frac{1}{2+\alpha}$. This yields part (c).

¹⁸The analysis presented in this subsection applies for any $\alpha \geq 0$, thus also for $\alpha = 0$. This effectively implies that the exact same conclusions are obtained if we just focus on the probability of taking the correct decision instead, rather than on the judge's payoff function (which weighs taking the correct decision differently in different eventualities).

¹⁹Note that, although they all reach lower bound b_0 , equilibrium Pool on the one hand and Inf.3 and Inf.4 on the other hand cannot all occur for a given level of b ; Pool requires $b > \frac{1}{2+\alpha}$ and thus $b_0 = b$, while Inf.3 and Inf.4 require $b < \frac{1}{2+\alpha}$ and thus $b_0 = 1 - b(1 + \alpha)$.

The equilibrium payoffs in Inf.2 equal $r_P - b(1 - r_P)\alpha - c$. From Lemma 2 it immediately follows that these strictly exceed b_0 on the range $\underline{b}(r_P, c; \alpha) < b < \bar{b}(r_P, c; \alpha)$ where this equilibrium exists. With \bar{b} as a shorthand for $\bar{b}(r_P, c; \alpha)$, these payoffs can be rewritten as $r_P - b(1 - r_P)\alpha - c = r_P - c - b[(1 - r_P)\alpha + 1] + b = \left(\frac{\bar{b}-b}{\bar{b}}\right) \cdot [r_P - c] + b$. Given $q_A > 0$ in Inf.1 and thus acquit being a best response (taking the equilibrium p as given), the judge's equilibrium payoffs there coincide with always acquitting after a precise statement. In that case the judge only arrives at the wrong verdict if the suspect is indeed guilty and makes a precise statement, which happens with probability $(1 - b)p$. Hence the judge's expected payoffs in Inf.1 equal $1 - (1 - b)p$, with $p = \frac{b}{1-b} \cdot \frac{(1-r_P)(1+\alpha)+c}{r_P-c} = \left(\frac{b}{1-b}\right) \cdot \left(\frac{1-\bar{b}}{\bar{b}}\right)$ from Proposition 2. Rewriting gives expected payoffs of $\left(\frac{\bar{b}-b}{\bar{b}}\right) + b$ in Inf.1. From $r_P - c < 1$ it follows that these strictly exceed the payoffs in Inf.2 derived above. This gives part (b). \square

As Proposition 3 reveals the judge's payoffs are equal to lower bound b_0 in Inf.3, Inf.4 and Pool. This is immediate in the pooling equilibrium Pool in which she does not get any additional information from either the statements per se or the verification thereof. Yet, perhaps somewhat surprisingly, strategic information revelation and potential verification do not guarantee higher payoffs to the judge, as Inf.3 and Inf.4 exemplify. In equilibrium Inf.4 the guilty suspect mixes between a vague and a precise statement. This information is influential because it affects the choice the judge makes, but effectively immaterial as her expected payoffs do not improve. A similar observation holds with respect to Inf.3. Here the guilty suspect again mixes between a vague and a precise statement and the judge occasionally investigates the latter. Despite both the statements per se and the investigation revealing influential (i.e. decision relevant) information, the judge again gains nothing in expected payoffs terms (as the benefits of a better verdict cancel out against the investigation costs c borne).

The judge does strictly improve upon deciding on her prior belief in the remaining equilibria. In separating equilibrium Sep she does so to the fullest extent possible and obtains her maximum payoff equal to 1. The incremental value $1 - b_0$ of the information received can be solely attributed to the statements per se. In Inf.1 and Inf.2 the judge also strictly benefits from the additional information obtained, albeit to a smaller extent. Holding prior belief b constant, the judge's expected payoffs are higher in Inf.1 than in Inf.2 (and, similarly so, higher in Inf.1 than in Inf.3 and Inf.4 for a given b). The intuition here is that in both equilibria it is a best response for the judge to investigate a precise statement, making that the judge is equally well off if such a statement is indeed received. Yet only in Inf.1 the guilty type now and then sends a vague statement ($0 < p < 1$ vs. $p = 1$ in Inf.2) and the judge does strictly better in those instances. Overall, in Inf.1 the judge thus obtains valuable information via both the statements per se as well as from (occasional) investigation,

while in Inf.2 the judge only obtains valuable information via investigation.

1.4.2.2 Comparative statics in the verification technology

For generic parameter values there is a unique equilibrium outcome. Taking the prior level of evidence (as captured by b) and thus the extent of the investigation problem as given, the judge may benefit from shifts in the parameters that characterise the verification technology. These may either induce a shift towards a ‘better’ equilibrium as ranked in Proposition 3, or improve the judge’s expected payoffs within a given equilibrium. Proposition 4 formally characterises both these extensive and intensive margin (comparative statics) effects.

Proposition 4. *Shifts in the parameters $(\lambda_P, r_P, \pi_P, c)$ of the verification technology may have both intensive margin (within equilibrium) and extensive margin (shift to a different equilibrium) effects on the judge’s equilibrium payoffs.*

(a) *Shifts in λ_P and π_P have extensive margin effects only. An increase in λ_P makes a beneficial shift towards either Sep or Inf.1 more likely, while an increase in π_P makes a beneficial shift towards Inf.1 more likely;*

(b) *Shifts in r_P and c have both extensive and intensive margin effects:*

(Ext) *both an increase in r_P and a decrease in c make a beneficial shift towards either Inf.1 or Inf.2 more likely;*

(Int) *the judge’s payoffs within Inf.1 and Inf.2 are increasing in r_P and decreasing in c .*

Proof. From Proposition 3 the judge’s equilibrium payoffs in Int.3, Int.4, and Pool equal b_0 and those in Sep equal 1. These payoffs are independent of $(\lambda_P, r_P, \pi_P, c)$. Hence intensive margin effects only concern Inf.1 and Inf.2. The judge expected equilibrium payoffs in Inf.2 equal $r_P - b(1 - r_P)\alpha - c$ and thus increase with r_P , decrease with c and are independent of λ_P and π_P . From the proof of Proposition 3, the equilibrium payoffs in Inf.1 equal $1 - (1 - b)p = 1 - b \cdot \frac{(1 - r_P)(1 + \alpha) + c}{r_P - c}$. Also these increase with r_P and decrease with c and are independent of λ_P and π_P . This yields the claims about intensive margin effects in part (a) and (b.Int).

The extensive margin effects in both part (a) and (b.Ext) follow from the payoff ranking of equilibria in Proposition 3, together with the conditions for existence in Proposition 1 and Proposition 2, and the comparative statics of the relevant thresholds in $(\lambda_P, r_P, \pi_P, c)$. In particular, $\underline{b}(r_P, c; \alpha)$ decreases with r_P and increases with c , $\bar{b}(r_P, c; \alpha)$ increases with r_P and decreases with c , $\hat{c}(r_P; \alpha)$ increases with r_P and $\hat{\lambda}(r_P, \pi_P)$ decreases with both r_P and π_P . \square

Clearly, the judge would prefer a verification technology that allows for the highest ranked equilibrium as identified in Proposition 3. She thus would prefer the cognitive lying costs λ_P to be prohibitively high for the guilty type, such that Sep materialises. Otherwise it would be best for her to have low verification costs c and ‘sorting’ condition (1.4) to be satisfied, as to enable Inf.1. In meeting condition (1.4) a high λ_P is again conducive, but also a sufficiently harsh obstruction penalty π_P helps (because threshold $\hat{\lambda}(r_P, \pi_P)$ decreases with π_P). Beyond their extensive margin effects of enabling Inf.1, however, an increase in either λ_P or π_P provides no additional benefits. The main intuition here is that the guilty type’s statement strategy within a given equilibrium (as reflected by p) does not vary with these parameters and hence neither do the judge’s equilibrium payoffs.

In contrast, variations in r_P and c do have both extensive, as well as intensive margin – i.e. within equilibrium – effects. The extensive margin effects follow from how compliance with the relevant thresholds $\underline{b}(r_P, c; \alpha)$, $\bar{b}(r_P, c; \alpha)$, $\hat{c}(r_P; \alpha)$, and $\hat{\lambda}(r_P, \pi_P)$ is affected. Both an increase in r_P and a decrease in c facilitate a beneficial shift from a lower ranked equilibrium towards either Inf.2 or Inf.1 (including for r_P a potential shift from Inf.2 to Inf.1). The intensive margin effects derive from two different causes. In equilibrium Inf.2 they follow from how changes in r_P and c affect the cost-effectiveness of the actual verification process itself. To illustrate, the judge’s expected payoffs in Inf.2 can be decomposed as:²⁰

$$r_P - b(1 - r_P)\alpha - c = \underbrace{b_0}_{\text{prior}} + \underbrace{0}_{\text{statements per se}} + \underbrace{[r_P - b(1 - r_P)\alpha - b_0]}_{\text{verification}} - c$$

Since both suspect types always make a precise statement in Inf.2, observing such a statement per se does not provide any information and, thus, has zero incremental value. Relative to deciding without first verifying, which would yield b_0 , verification of the precise statement received has two opposing effects. On the one hand, it improves decision making if it corrects a would-be wrong verdict based on the prior belief alone. On the other hand, it worsens decision making in those instances where it wrongly overturns a would-be correct verdict based on b alone. The overall net informational effect – reflected within square brackets – is positive and outweighs the costs of verification c . This informational value of verification increases with r_P and is independent of (λ_P , π_P and) c .

In contrast, in equilibrium Inf.1 the benefits from improvements in the investigation technology via r_P and c effectively follow entirely from their spill-over effects on

²⁰To intuitively understand the expected payoffs on the l.h.s., note that in Inf.2 the judge always verifies and thus always bears cost c . She arrives at a correct verdict and thus a payoff of 1 with probability r_P . With the remaining probability $(1 - r_P)$ she takes the wrong decision, with negative payoffs $-\alpha$ (only) if she wrongly convicts an innocent suspect (whose frequency of occurrence in the population is b).

the strategic behaviour of the guilty type, because such improvements induce him to mimic the innocent type less often. To illustrate, the effective reliance on the different information sources can again be inferred from the decomposition of the judge's equilibrium payoffs:²¹

$$1 - (1 - b)p = \underbrace{b_0}_{\text{prior}} + \underbrace{(1 - \sigma_P) + b - b_0}_{\text{statements per se}} + \underbrace{\sigma_P q_I [r_P - b^P(1 - r_P)\alpha - b^P]}_{\text{verification}} - \sigma_P q_I c,$$

where $\sigma_P \equiv b + (1 - b)p$ denotes the overall probability that a precise statement is made in equilibrium. The final two terms cancel out, reflecting that in Inf.1 the judge is indifferent between verifying a precise statement and immediately acquitting for sure. Based on just the statements per se, the judge would convict after a vague statement and acquit after a precise one, yielding $(1 - \sigma_P) + b$ in expected payoffs. The incremental value $(1 - \sigma_P) + b - b_0$ increases with r_P and decreases with c (and is independent of λ_P and π_P). This reflects the indirect, 'deterrence' effect of potential verification. If verification becomes either more reliable or less costly, it deters the guilty type from mimicking the innocent type often, i.e. it lowers p and thus σ_P . This in turn makes a precise statement per se more informative. The direct informational value of now and then checking on a precise statement made equals $\sigma_P q_I [r_P - b^P(1 - r_P)\alpha - b^P]$. This informational value *decreases* with r_P and *increases* with c .²²

Intuitively, if the verification technology becomes more reliable (higher r_P), overall more valuable information is obtained in Inf.1. But perhaps somewhat counter-intuitively, as the above decomposition reveals this beneficial impact is completely driven by the incremental benefits from the statements per se. *Less* valuable information is actually obtained from verification the higher r_P is. The driving force here is that actual verification occurs less often if r_P increases.²³ This reflects the general

²¹As explained in the proof of Proposition 3, given the judge's indifference in Inf.1 between convict and acquit after a precise statement, her equilibrium payoffs coincide with those of always acquitting after such a statement (keeping the equilibrium p fixed). In that case the judge only arrives at a wrong verdict if the suspect is guilty and makes a precise statement, which happens with probability $(1 - b)p$. In all other instances she takes the right decision, yielding 1. This explains her expected payoffs $1 - (1 - b)p$ in Inf.1.

²²Increases in cognitive lying costs λ_P or obstruction penalty π_P have no intensive margin effects in Inf.1 as they bring no overall net benefits to the judge (cf. Proposition 4). Such marginal changes do not impact the amount of (valuable) information the statements per se reveal in Inf.1, i.e. do not strengthen the deterrence effect. This follows because in Inf.1 an increase in either λ_P or π_P reduces the frequency of actual verification q_I . The latter also implies that actually *less* (valuable) information is obtained from occasional verification. This is counterbalanced by incurring the costs of verification equally less often.

²³The informational value of actually verifying a precise statement received equals the term within square brackets $[r_P - b^P(1 - r_P)\alpha - b^P]$. The judge's indifference in Inf.1 between whether or not to verify a precise statement implies that this term equals c and thus is independent of r_P . Comparative statics of the direct informational value of occasional verification w.r.t. r_P thus solely follow from how $\sigma_P q_I$ is affected; this term is strictly decreasing in r_P .

intuition that a more effective stick works as a stronger deterrent and thus in the end needs to be used less often. Similarly so, a decrease in c causes that also less valuable information is obtained from verification, both because precise statements are made less often by the guilty type and – as a result – their actual verification then yields less additional information.

In summary, higher cognitive costs of lying and a higher obstruction penalty are beneficial to the judge to the extent that these enable more informative equilibria in which also the suspect's statements per se provide valuable information. The latter requires that the guilty type is effectively deterred away from always lying. Once the relevant threshold for this is met, however, increasing the lying costs or the obstruction penalty further does not increase the provision of valuable information. Improvements in the verification technology that make it more reliable or less costly do have an impact beyond meeting the relevant threshold, however. Even if the guilty type is willing to always lie, such improvements make actual verification more cost-effective (cf. Inf.2). And as soon as the threshold that deters the guilty type from always lying is met (cf. condition (1.4)), such improvements enlarge the deterrence effect. This creates a positive spill-over effect because the guilty type then reveals more information via the statement per se and the actual verification process itself actually yields less valuable information (cf. Inf.1).

1.5 Model extensions

In this section we discuss two extensions that add additional realism to the model: (i) incorporating the possibility of plea bargaining and (ii) accounting for a right to silence. The overall conclusion that follows from the discussion is that these extensions leave the main insights obtained from our basic setup largely unaffected.

1.5.1 Plea bargaining

In practice, a very high percentage of cases – up to 95%, see US Bureau of Justice Statistics (2003) – never reach the courtroom and is settled through some sort of plea bargaining. In this case, the prosecutor offers a penalty reduction in exchange for the suspect pleading guilty. In the literature, plea bargaining has been studied as having (among other things) an informational role in the screening of suspect types.

To incorporate this realistic element in our setup, we allow a third option to the suspect: besides making either a vague or a precise statement and the case going to court, he can also choose to confess and immediately receive a payoff of m , with $0 < m < 1$. Confession then yields strictly more than providing a vague statement

(inducing immediate conviction) does.²⁴ A direct implication of this added choice option is that in equilibrium the guilty type no longer provides a vague statement and effectively chooses between confessing and providing a precise statement only.²⁵

The single substantive difference with the equilibrium analysis in section 1.4 is that the relevant benchmark for λ_P in Lemma 3 has to be adapted from $q_A + q_I \cdot \hat{\lambda}(r_P, \pi_P)$ originally to $q_A + q_I \cdot \hat{\lambda}(r_P, \pi_P) - m$ now. This accounts for the fact that the relative benefits of a potentially more favourable decision after a precise statement are now – compared to the new and better alternative of confession – an amount m lower than before. Put differently, the original benchmark $q_A + q_I \cdot \hat{\lambda}(r_P, \pi_P)$ now applies to $\lambda_P + m$ rather than just λ_P before. Acknowledging this, we immediately obtain the following corollary from our main analysis.

Corollary 1. *Suppose that, besides making a statement $S \in \{P, V\}$, the guilty type can also Confess and immediately receive payoff m , with $0 < m < 1$. The guilty type then never chooses a vague statement. Propositions 1 through 4 immediately apply when we replace λ_P by $\lambda_P + m$ and let $1 - p$ now reflect the probability with which the guilty type confesses.*

In the presence of plea bargaining, full separation (Sep) is achieved if $\lambda_P > 1 - m$. Similarly, the condition for potential verification of precise statements to have a sufficiently strong deterrent effect now becomes:

$$\lambda_P > \hat{\lambda}(r_P, \pi_P) - m \tag{1.5}$$

Compared to condition (1.4), the opportunity costs m of lying are subtracted from the r.h.s., to account for the fact that the benefits of a plea bargain (yielding m) are foregone if the guilty type decides to give a precise statement instead. With plea bargaining the judge has an additional tool in trying to induce guilty suspects to come forward. The penalty reduction m complements cognitive lying costs λ_P and the obstruction penalty π_P in facilitating more informative equilibria. For the guilty type to refrain from always mimicking the innocent type, one can either make mimicking less attractive (i.e. a higher λ_P or π_P), or otherwise make the alternative

²⁴The imposed sentence after confession thus leads to a lower payoff reduction than the imposed sentence after conviction without confession does: $1 - m < 1$. Although in practice the prosecutor may have some discretion in the size of the penalty reduction offered, this discretion may be considerably restricted by binding guidelines, see e.g. the 2017 “Reduction in sentence for a guilty plea: Definitive guideline” from the sentencing council in the UK (UK Sentencing Council, 2017). Existing game theoretical models of plea bargaining typically allow the prosecutor to endogenously choose the penalty reduction; qualitatively this leads to the same conclusions with respect to amount of information revelation in equilibrium, see the discussion below.

²⁵We maintain the assumption that innocent suspects always provide precise statements. Dropping this assumption leads to the existence of additional pooling equilibria where both types choose to confess. Analogously to the main model, such pooling equilibria do not survive standard equilibrium refinements (Banks and Sobel, 1987). The maintained assumption essentially solves the multiplicity of equilibria issue in a simpler way.

of not mimicking more attractive (which is essentially what the plea bargain does). Beyond meeting the threshold for enabling information revelation via the statements per se, an increase in m has no beneficial impact though.

In an early game theoretic analysis of plea bargaining, Grossman and Katz (1983) showed that – if a prosecutor could *commit* to proceed to court if the plea offer is rejected – the plea offer can be used as a screening device to fully separate the guilty types from the innocent ones. A similar observation was made by Reinganum (1988) when extending the framework of Grossman and Katz (1983) by assuming that the prosecutor has private information regarding the strength of the case. Baker and Mezzetti (2001) have challenged this equilibrium separation possibility, as the underlying commitment on which it is based “...is inherently non-credible because any defendant that the prosecutor knows for sure is innocent will never stand trial” Baker and Mezzetti (2001, p. 151). Models that drop this possibility to fully commit to go to trial all find that plea bargaining is (at most) essentially semi-separating, with the plea offer accepted by the guilty type with some probability but rejected by the innocent type for sure.²⁶ In this equilibrium, the prosecutor still proceeds to trial with probability one if the plea offer is rejected (although this is now based on an equilibrium best response rather than on an ex ante commitment as in the earlier articles). The partially pooling equilibria in our setup are qualitatively similar in terms of the guilty type using a mixed strategy, but differ in the judge/prosecutor doing so as well. This causes that if either λ_P , π_P or m increases, only the judge adapts her behaviour, while leaving the behaviour of the suspect unaffected. Such changes thus do not have positive intensive margin (within equilibrium) effects (cf. Subsection 1.4.2).²⁷

²⁶See Baker and Mezzetti (2001); Bjerk (2007); Kim (2010); Tsur (2017). A remaining criticism of some of these models is that the behaviour of the judge/jury is assumed to be purely exogenous and does not react to (the information revealed by) the behaviour of the prosecutor and the suspect. This arguably provides another unrealistic commitment possibility, viz. to a mechanical conviction rule. Bjerk (2007) and Tsur (2017) endogenise the behaviour of the judge/jury and obtain the same type of semi-separating equilibrium (though a multiplicity of these may exist). Note that our simplified setup with a unitary judiciary actor essentially corresponds to the case where different representatives of the judiciary share the same information and beliefs, and endogenously act on these; the probability of conviction is thus entirely the result of equilibrium strategies.

²⁷Although the obstruction penalty and the penalty reduction play a similar deterrence role in incentivising the guilty type to sometimes either implicitly (via a vague statement) or explicitly confess, their payoff implications for the suspect are quite different. He is clearly better off with higher penalty reductions than with higher obstruction penalties. From a broader social welfare perspective, however, society might dislike penalty reductions as they allow offenders to largely ‘get away with it’ and rather prefer penalties for obstruction (Fagan, 1981; Cohen and Doob, 1989; Herzog, 2003; Johnson, 2019). A reduced form way to incorporate such broader considerations in our model would be to let the judge’s expected payoffs depend on m and π_P as well.

1.5.2 Right to silence

In our baseline model, the judge can use both the strategic behaviour of the suspect as well as the outcome of the potential investigation to update her belief about the suspect's innocence and act accordingly without any restrictions. In particular, the suspect's choice of making a vague statement can be fully held against him and lead to immediate conviction. Traditional common law systems, however, typically give the suspect the 'right to remain silent'; if a suspect refuses to answer any questions, the verdict must solely be based on other evidence and the suspect's silence cannot be considered evidence of his guilt (*Miranda v. Arizona*, 1966). Effectively, this right thus works as a commitment to ignore some of the suspect's strategic information revelation.

To analyse the implications of a right to silence for our analysis, we again introduce a third option to the suspect: besides making a precise ($S = P$) or a vague ($S = V$) statement as in the baseline model, he can now also remain silent ($S = \phi$). Since no viable leads are obtained at all, a silent statement is even more difficult to investigate than a vague one, so we assume $\frac{1}{2} \leq r_\phi < r_V$.²⁸ Moreover, by remaining silent the suspect is not obstructing justice in any way (except perhaps in highly unusual circumstances), implying that $\pi_\phi = 0 \leq \pi_V$. We continue to assume that the innocent suspect always makes a precise statement. Therefore, observing $S = \phi$ is a clear indication of being guilty and the introduction of this additional option to the suspect per se has no impact on equilibrium outcomes if no further assumptions are made. Within our setup prior belief b can be interpreted as the evidence collected by the judge before any statement is received. If silence cannot be held against the subject, this then constitutes all the evidence there is. We thus incorporate a right to silence in the following way.

Assumption 1.1. *RTS Under a right to silence the judge's choice of action after a silent statement $S = \phi$ should be guided by a restricted posterior belief $b^\phi = b$, rather than by a Bayesian posterior belief $b^\phi = 0$ that applies in the absence of such a right.*

Similar to analysis in the previous subsection, the single substantive difference with the baseline model is that the benchmark payoffs to which the relative benefits of making a precise statement have to be compared are now (potentially) different. Note that with a RTS, making a vague statement is (weakly) dominated by remaining silent for the guilty type. The relevant benchmark payoffs are thus given by the judge's choice of action after $S = \phi$. For this we can immediately apply Lemma 2 when we replace r_P with r_ϕ and b^P with b . Hence, if $b < \underline{b}(r_\phi, c; \alpha)$, the judge convicts for sure after remaining silent and the equilibrium analysis coincides with the one of

²⁸Clearly, with a silent statement there is nothing to verify. Investigation of a silent statement thus should be interpreted as additional independent investigation by the judge not inspired by the empty statement made.

the baseline model. A RTS is then inconsequential. In case $b \in (\underline{b}(r_\phi, c; \alpha), \bar{b}(r_\phi, c; \alpha))$, a RTS effectively forces the judge to investigate in case of silence. By the equivalent of Lemma 1, the guilty type is then acquitted with probability $1 - r_\phi$ when keeping silent. This gives the guilty type an expected payoff of $1 - r_\phi$ after $S = \phi$, rather than 0. These opportunity costs $1 - r_\phi$ from making a precise statement now come on top of the direct lying costs λ_P , but apart from that the analysis is as before. Finally, if $b > \bar{b}(r_\phi, c; \alpha)$, a RTS ensures that the suspect is always acquitted after silence. As this also happens after a precise statement (made by the innocent type), the equilibrium outcome in terms of the judge's verdict is then the same as in equilibrium Pool in Proposition 1. From these observations we immediately obtain the following corollary.²⁹

Corollary 2. *Suppose that, besides making a statement $S \in \{P, V\}$, the guilty type can also remain silent, i.e. $S = \phi$, with $\frac{1}{2} \leq r_\phi < r_V < r_P$ and $\pi_\phi = 0 \leq \pi_V \leq \pi_P$.*

- (a) *Without a RTS, Propositions 1 through 4 immediately apply when we let $1 - p$ now reflect the probability with which the guilty type either chooses $S = V$ or $S = \phi$ (both leading to immediate conviction);*
- (b) *With a RTS, the guilty type never chooses a vague statement. Letting $1 - p$ now reflect the probability with which the guilty type chooses $S = \phi$, it then holds that:*
 - (b.1) *if $b < \underline{b}(r_\phi, c; \alpha)$, then Propositions 1 through 4 continue to apply (for b in this range) and we either have Sep, Inf.1, Inf.2, Inf.3 or Inf.4;*
 - (b.2) *if $\underline{b}(r_\phi, c; \alpha) < b < \bar{b}(r_\phi, c; \alpha)$, then Propositions 1 through 4 continue to apply (for b in this range) when we replace λ_P with $\lambda_P + (1 - r_\phi)$ and we either have Sep or Inf.1. The judge now always investigates in case of silence;*
 - (b.3) *if $b > \bar{b}(r_\phi, c; \alpha)$, then the guilty suspect always remains silent and is always acquitted (outcome equivalent to equilibrium Pool).*

In case (b.2) of Corollary 2, potential verification of precise statements is a sufficiently powerful deterrent if:

$$\lambda_P > \hat{\lambda}(r_P, \pi_P) - (1 - r_\phi) \quad (= r_\phi - r_P(1 + \pi_P)) \quad (1.6)$$

From $r_\phi < r_P$ the r.h.s. is negative. The condition is thus always satisfied, irrespective of λ_P and π_P . Therefore, only equilibria that are equivalents of Sep and Inf.1 remain to exist (and Inf.2 and Inf.3 disappear), which correspond to these after replacing

²⁹For listing the different equilibria that exist in the various subcases of part (b) we have used that $(\underline{b}(r_\phi, c; \alpha), \bar{b}(r_\phi, c; \alpha)) \subset (\underline{b}(r_P, c; \alpha), \bar{b}(r_P, c; \alpha))$ given that \underline{b} decreases with r , \bar{b} increases with r and $r_\phi < r_P$. This also implies that $(\underline{b}(r_\phi, c; \alpha), \bar{b}(r_\phi, c; \alpha))$ is empty if $(\underline{b}(r_P, c; \alpha), \bar{b}(r_P, c; \alpha))$ is, i.e. for $c > \hat{c}(r_P; \alpha)$. Moreover, for case (b.2) we have used that sorting condition (1.6) discussed below always holds.

λ_P with $\lambda_P + (1 - r_\phi)$. The probability p with which the guilty type makes a precise statement stays exactly the same as without a RTS, but the judge now always investigates after a silent statement and immediately acquits after a precise statement with increased probability $q_A = \frac{\lambda_P + (1 - r_\phi) - \hat{\lambda}(r_P, \pi_P)}{1 - \hat{\lambda}(r_P, \pi_P)}$ in Inf.1.

The above shifts in equilibria are in line with the effects of a right to silence identified by the game theoretic analyses of Seidmann (2005) and Leshem (2010) (see also Seidmann and Stein (2000); Mialon (2008)). In particular, the innocent type benefits from such a right in two ways. A first, direct benefit is that it provides “innocent suspects, who are otherwise compelled to speak, with the alternative of silence” (Leshem, 2010, page 400). In our simplified setup this effect is reflected by the non-existence of the informative equilibria in case $b > \bar{b}(r_\phi, c; \alpha)$; the judge is then compelled to acquit in the absence of further information and only an outcome equivalent to Pool remains. In general, exercising the right to silence provides the innocent type a safe alternative to making a precise statement, as with the latter he runs the potential risk of his statement being wrongly falsified. A second, indirect benefit is that innocent types who choose to make a precise statement are less likely to be wrongfully convicted. This effect is exemplified by the increased probability of immediate acquittal q_A in Inf.1 above.

More generally, Corollary 2 reveals that if $b < \underline{b}(r_\phi, c; \alpha)$ a RTS is immaterial for the equilibrium payoffs of both subject types and the judge. For a sufficiently high prior belief $b > \bar{b}(r_\phi, c; \alpha)$ a RTS either increases the equilibrium payoffs of the innocent and the guilty type or leaves these unaffected. The opposite holds for the judge; she then either earns the same or loses (cf. Proposition 3). In the intermediate range where $\underline{b}(r_\phi, c; \alpha) < b < \bar{b}(r_\phi, c; \alpha)$ introducing a RTS again always (weakly) benefits the innocent and the guilty type. But, interestingly, the effect for the judge then can go either way. The typical case remains that the judge loses.³⁰ Yet the opposite may happen (only) when introducing a RTS induces a shift from Inf.1 to Sep. The induced change in the guilty type’s behaviour then makes that the judge can convict him with probability r_ϕ under a RTS at investigation costs c to her, compared to probability $1 - p$ before (where p is given in Proposition 2 for Inf.1). Depending

³⁰In case (b.2) of Corollary 2 we either have Sep, Inf.1 or Inf.2 in the absence of a RTS (note that $c < \hat{c}(r_P; \alpha)$ for the belief range to be non-empty). Now if Inf.2 applies, then introducing a RTS necessarily leads to a shift to Inf.1. This follows because $\lambda_P < \hat{\lambda}(r_P, \pi_P) \implies \lambda_P + 1 - r_\phi < 1$, given that $\hat{\lambda}(r_P, \pi_P) < \frac{1}{2}$ and $r_\phi \geq \frac{1}{2}$. This shift benefits both the innocent (as q_A increases) and the guilty type (now convicted with smaller probability $p \cdot r_P + (1 - p) \cdot r_\phi < r_P$ and bearing λ_P less often), but harms the judge. She gets $(1 - b) \cdot (1 - p) \cdot (r_P - r_\phi)$ less in Inf.1 with a RTS as compared to Inf.2 without. (As $q_I > 0$ in Inf.1 and thus investigation a best response, the judge’s equilibrium payoff can be calculated as if $q_I = 1$. In that case only the outcome for a guilty type is different from Inf.2 in the instances that he now remains silent in Inf.1.) In case Sep applies without a RTS, it continues to apply with a RTS. This leaves the innocent type unaffected, benefits the guilty type (given now investigation after silence) but harms the judge (as the guilty type is now sometimes acquitted). The same – guilty wins, judge loses, innocent unaffected – holds if Inf.1 applies both without and with a RTS. This leaves the case where Inf.1 applies without a RTS and Sep with a RTS, which is discussed in the main text.

on parameter values, we either have $r_\phi - c < 1 - p$ or $r_\phi - c > 1 - p$.³¹ In the latter case the judge is strictly better off in Sep under a RTS than in Inf.1 without a RTS.³² Unlike Seidmann (2005) and Leshem (2010), therefore, we do obtain instances in which the judge explicitly benefits from an ex ante commitment to block adverse (but correct!) inferences from silence.

Most important for our purposes, however, is that the qualitative features of the informative equilibria are robust to introducing a right to silence. Although such a right diminishes the role for strategic information revelation when the judge is initially inclined to acquit, this role essentially remains the same when this is not the case. Strategic information revelation via the statements per se thus continues to play an important role in affecting the judge choice of action and remains complementary to the judge now and then checking on messages. Moreover, a right to silence reinforces the attractiveness of improvements in reliability, as neither the lying costs nor the obstruction penalty have a supportive role when the judge is a priori insufficiently confident about what the appropriate verdict would be. The verifiability approach to lie detection thus continues to have a strong bite even in the presence of a right to silence.

1.6 Conclusion

In this chapter, we analyse the strategic interaction between a speaker who wants to convince an investigator of his innocence and an investigator who wants to know the truth, i.e. whether the speaker is guilty or innocent. In our model, the investigator can check the specific details in the statement of the speaker at some cost. This yields informative, but imperfect evidence. The more detailed the speaker's statement is, the more reliable the examination of this statement becomes. This encourages innocent speakers to be forthcoming in providing many verifiable details in their statement, while guilty types would prefer to remain vague, also because fabricating a precise but false statement is cognitively costly to them. If, on the basis of an investigation, the investigator concludes that the speaker is lying, an additional obstruction penalty is imposed on the speaker.

We show that complete separation is possible only if lying is prohibitively costly to a guilty speaker. Full information revelation then takes place via the statements per se. With lower cognitive costs of fabricating a precise but false statement, the speaker's statement is partially revealing at best and provides valuable information

³¹To illustrate, consider the following numerical example. Let $b = \frac{1}{3}$, $\alpha = 1$, $\lambda_P = \frac{3}{5}$, $c = \frac{1}{20}$, $\pi_P = 0$ and $r_\phi = \frac{11}{20}$. Then $\underline{b}(r_\phi, c; \alpha) \approx 0,323$ and $\bar{b}(r_\phi, c; \alpha) \approx 0,345$ and thus case (b.2) indeed applies. For these parameters $r_\phi - c = \frac{1}{2}$. Now for $r_P > \frac{7}{10}$ it holds that $1 - p > \frac{1}{2}$ and for $r_P < \frac{7}{10}$ that $1 - p < \frac{1}{2}$.

³²Because $q_A > 0$ in Inf.1, the judge payoffs can be calculated as if $q_A = 1$ and only the outcome for the guilty type is different in Inf.1 and Sep.

only if its potential verification is a sufficiently strong deterrent. The latter requires that the joint effect of the lying costs, the reliability of verification, and the obstruction penalty is strong enough to potentially tip the balance in the trade-off for a guilty speaker. If this is indeed the case, a partially pooling equilibrium exists in which the guilty type mixes between making a vague and making a precise statement (with the innocent type making a precise statement for sure). Precise statements are now and then investigated by the investigator to verify their veracity. In this equilibrium verification and strategic information revelation by the speaker thus go hand in hand.

Our analysis allows us to understand the behavioural patterns observed for lie detection methods. It explains the shortcomings of the early approaches that were based on a speaker's micro-expression of emotional cues that do not convey sufficient reliable information. In particular, our model explains why no beneficial information will be revealed in equilibrium when the observer's investigation is not sufficiently reliable and mimicking an innocent type is cognitively not prohibitively costly. For recent advances with the verifiability approach, the picture is more promising. By judging the frequency of precise verifiable details in a speaker's statement, more reliable information is acquired. In such settings our analysis suggests that a partial pooling equilibrium is most plausible. This equilibrium agrees with empirical observations, in which innocent types furnish their statements with precise, verifiable details, whereas guilty types face a difficult trade-off that they solve by sometimes imitating the innocent types and by remaining vague at other times.

Our analysis also offers some insights that go beyond what has been observed in the recent psychological literature on lie detection. The overall amount of information provision in the partially pooling equilibrium is especially facilitated by an improved reliability of the verification technology. This renders verification more informative per se and (thus) makes the investigator more willing to investigate. Realising this, the guilty type reduces the likelihood with which he makes a precise statement, in turn providing the investigator actually less incentives to investigate. The overall net effect is that, when reliability improves, more can be learned from the strategic behaviour of the speaker and actually less is learned via actual verification. In contrast, not much is accomplished by enhancing the obstruction penalty further. Once the deterrence-by-verification condition for the existence of the partial pooling equilibrium is met, such an increase has no further impact on the usefulness of a lie detection method. An increase in the obstruction penalty then leads the investigator to investigate less, but leaves the amount of strategic information revelation unaffected. The investigator – and thus also “truth” – is better served by an improved reliability of the verification technology. A similar remark applies to the cognitive lying costs. These only have extensive margin effects in enabling more informative equilibria, but leave the amount of *valuable* information transmission within a given

equilibrium unaffected.

In our approach, the quality of the verification technology is exogenous to the model. In practice, the relevant actors can make decisions that affect the quality of the investigation technology. The legal system may benefit from novel scientific insights and investments therein, such as in the area of the development of DNA identification or in the area of verbal detection methods. A judge can also order earlier searches of a suspect's house before any evidence is destroyed. Alternatively, a mother suspecting her son is using drugs can search his phone before asking him. After such actions, any statement made by the son or the suspect can be verified or falsified more accurately. However, such endogenous improvements in accuracy do not come as a free lunch. Searching her son's phone may destroy trust in the relationship between the mother and the son; sweeping a suspect's house before pressing charges may violate their right against unreasonable searches and may be deemed as inadmissible evidence in court. To mitigate such adverse effects, a mother can only check the whereabouts of her son after hearing his explanations, and a judge can increase the number of witnesses to examine. Allowing the relevant actor to endogenously decide the scope of investigation and taking the adverse effects of the increase in accuracy into account goes beyond the scope of our chapter, but constitutes a fruitful direction for future research.

1.7 Appendix to Chapter 1

1.7.1 Proof of Proposition 2

Proof. Let $\lambda_P < 1$ and $b < \bar{b}(r_P, c; \alpha)$. Observe first that then $p = 0$ cannot happen in equilibrium; this would induce $q_A = 1$ by Lemma 2, in turn providing the guilty type an incentive to deviate to $p = 1$ per Lemma 3. Hence necessarily either $p > 0$ or $p = 1$.

We next consider the various mutually exclusive parameter ranges in turn. First consider the case $c > \hat{c}(r_P; \alpha)$. From Lemma 2 then $q_I = 0$ necessarily and thus $q_A = 1 - q_C$. Now suppose $p = 1$. Then we would have $b^P = b < \bar{b}(r_P, c; \alpha) = \frac{1}{2+\alpha}$ and thus $q_A = 0$ as well. But for $q_C = 1$ the guilty type would want to choose $p = 0$ by Lemma 3, contradicting $p = 1$.³³ Hence $0 < p < 1$ necessarily. The required indifference of the guilty type between making a vague and a precise statement then

³³In the non-generic case $\lambda_P = 0$ the guilty type is willing to choose $p > 0$ even when $q_C = 1$. Then multiple equilibria exist, which are all payoff and outcome equivalent to Inf.4 (with $q_A = 0$) as derived here.

implies for q_A that:

$$0 = q_A - \lambda_P \implies q_A = \lambda_P$$

In turn, $0 < q_A < 1$ requires that the judge is indifferent between acquit and convict after $S = P$. From Lemma 2 and equation (1.1) we then obtain that:

$$b^P = \frac{1}{2 + \alpha} \implies p = \frac{b}{1 - b} \cdot (1 + \alpha)$$

This yields equilibrium Inf.4.

From now on assume $c < \hat{c}(r_P; \alpha)$. In that case $\underline{b}(r_P, c; \alpha) < \frac{1}{2 + \alpha} < \bar{b}(r_P, c; \alpha)$. From Lemma 2 this implies that the judge may potentially mix in equilibrium between two options (from convict, investigate and acquit) at most, because b^P cannot meet more than one of these different thresholds at the same time.³⁴

Consider the case where $\lambda_P > \hat{\lambda}(r_P, \pi_P)$. Suppose $p = 1$. Then we would have $b^P = b < \bar{b}(r_P, c; \alpha)$ and thus $q_A = 0$ from Lemma 2. But then the guilty type would want to choose $p = 0$ per Lemma 3, a contradiction. Hence $0 < p < 1$ necessarily. The required indifference of the guilty type then implies by the same lemma:

$$0 = q_A + (1 - q_A) \cdot \hat{\lambda}(r_P, \pi_P) - \lambda_P \implies q_A = \frac{\lambda_P - \hat{\lambda}(r_P, \pi_P)}{1 - \hat{\lambda}(r_P, \pi_P)}$$

In turn, $0 < q_A < 1$ requires that the judge is indifferent between acquit and investigate after $S = P$. From Lemma 2 and equation (1.1) we then obtain that:

$$b^P = \bar{b}(r_P, c; \alpha) \implies p = \frac{b}{1 - b} \cdot \frac{(1 - r_P)(1 + \alpha) + c}{r_P - c}$$

This yields equilibrium Inf.1.

Finally, consider the case where $\lambda_P < \hat{\lambda}(r_P, \pi_P)$ (besides $c < \hat{c}(r_P; \alpha)$). First assume $b > \underline{b}(r_P, c; \alpha)$. From $b^P \geq b$ by equation (1.1) it then follows that $q_C = 0$ by Lemma 2. In turn, by Lemma 3 we obtain that $p = 1$ necessarily. Hence $b^P = b$ and $q_I = 1$ by Lemma 2. This yields equilibrium Inf.2.

Next assume $b < \underline{b}(r_P, c; \alpha)$. Suppose $p = 1$. Then we would have $b^P = b < \underline{b}(r_P, c; \alpha)$ and thus $q_C = 1$ by Lemma 2. But then the guilty type would want to choose $p = 0$ per Lemma 3, a contradiction. Hence $0 < p < 1$ necessarily. The

³⁴Mixing between all three options convict, investigate and acquit would require both $b^P = \frac{1}{2 + \alpha}$ to make the judge indifferent between convict and acquit, as well as $c = \hat{c}(r_P; \alpha)$ to ensure indifference with investigate. This thus can happen in non-generic knife-edge cases only.

required indifference of the guilty type then implies:

$$0 = q_I \cdot \hat{\lambda}(r_P, \pi_P) - \lambda_P \implies q_I = \frac{\lambda_P}{\hat{\lambda}(r_P, \pi_P)}$$

In turn, $0 < q_I < 1$ requires that the judge is indifferent between investigate and convict after $S = P$. From Lemma 2 and equation (1.1) we then obtain that:

$$b^P = \underline{b}(r_P, c; \alpha) \implies p = \frac{b}{1-b} \cdot \frac{r_P(1+\alpha) - c}{1-r_P+c}$$

This yields equilibrium Inf.3. □

Chapter 2

Habitual communication

This chapter is based on Ioannidis (2022).

2.1 Introduction

People communicate more honestly than predicted by economic models of self-interested agents maximising their monetary utility, both in individual (Gibson et al., 2013; Abeler et al., 2014, 2019) and in strategic settings (Gneezy, 2005; Leib, 2021). The propensity to communicate honestly has been shown to vary both between individuals (Sánchez-Pagés and Vorsatz, 2007; Hurkens and Kartik, 2009; Serota et al., 2010) and between groups such as country (Dieckmann et al., 2016; Cohn et al., 2019), occupation (Cohn et al., 2014, 2015) and religiosity (Arbel et al., 2014). Primarily interacting in common-interest settings may facilitate the formation of habits of truth-telling and believing messages. Primarily interacting in conflicting-interest settings may facilitate the formation of habits of lying and distrusting messages. If communication is affected by habits, then excessive honesty may be derived from familiarity with common-interest settings. This chapter provides empirical evidence for this line of reasoning.

We focus on strategic communication in the form of strategic information transmission between two asymmetrically informed agents where (i) preferences are misaligned and (ii) messages do not directly affect monetary payoffs. Many economically relevant interactions are characterised by such information asymmetry. A suspect knows if he is guilty or not, whereas a judge does not. A seller knows the true quality of his product, whereas a buyer may not. In such situations, the informed agent may send a message to the uninformed one. How informative will this communication be? In a seminal article, Crawford and Sobel (1982) analysed such cheap talk games and showed that communication becomes less informative when the preferences of the sender and the receiver diverge.

Modern psychology and neuroscience define habits as cue-response associations acquired through repeated interactions in a stable context (Wood and Runger, 2016; Mazar and Wood, 2018). Habitual behaviour is fast, subconscious, and, even though initially driven by goal pursuit, eventually follows automatically from the cues without goal dependence.¹ In the framework of dual process theory of reasoning (Kahneman, 2011), habits shape the default automatic response (System 1) and are only sometimes overridden by deliberate thinking with sufficient motivation (System 2). Empirical evidence document that a large part of everyday activities are habitual. Diary studies asking subjects to report their activities every hour have found that about 43% of human activities are repeated almost every day, in the same way, at the same time, without conscious deliberation (Wood et al., 2002; Lally et al., 2010). The

¹One of the first definition of habits dates all the way back to Aristotle in *Nicomachean Ethics*, where he defines them as dispositions, acquired through repetition, to perform certain types of action. We refer to Fleetwood (2019) for a comprehensive discussion of the different definitions of habits across economics, psychology and sociology.

prevalence of habits implies that, for many activities, the answer to why people act the way they do is simply because they are used to it.

The primary goal of the current chapter is to investigate whether and how habitual behaviour affects strategic communication. Specifically, we are interested to experimentally test two hypotheses: (i) whether familiarity with common-interest environments leads to more informative communication in unfamiliar environments compared to familiarity with conflicting-interests environments, and (ii) whether familiarity with common-interest environments leads to overcommunication whereas familiarity with conflicting-interest environments leads to undercommunication in unfamiliar environments.

The second goal of the chapter is to contrast two behavioural mechanisms that can explain habit reliance. The first mechanism is preference formation (Stigler and Becker, 1977; Becker and Murphy, 1988). In our setup, preference formation would imply that senders (receivers) familiar with a common-interest environment may develop a taste for truth-telling (following messages), whereas senders (receivers) who typically communicate in a conflicting-interests environment may develop a taste for lying (going against messages). In other words, exposure to a common-interest environment increases lying aversion whereas exposure to a conflicting-interests environment decreases it. Consequently, when interacting in a new environment, more lying averse agents will communicate more informatively than less lying averse agents. The second mechanism is inattention (Anderson, 2016; Jiang and Sisk, 2019). Agents may insufficiently adapt their strategy either because they failed to notice the change in the environment or because the consequences of sticking to their strategy are moderate. Thus, inattention would predict that the likelihood of changing communication strategy depends on the expected costs and benefits of doing so as well as the salience of the change in the environment.²

We use a controlled laboratory experiment to address our research questions. Our subjects play multiple rounds of a cheap talk sender-receiver game. In each round, the payoff-relevant state of the world is randomly drawn. The sender observes the true state whereas the receiver does not. The sender sends a message about the state to the receiver who then chooses an action determining the payoffs of both players. Our treatments vary the preference alignment between the two players. We use a 2×2 between-subjects treatment design. Subjects play overall 60 rounds of the sender-receiver game with either (fully) conflicting, partially aligned or (fully) aligned interests. The 60 rounds are divided in two parts of 30 rounds each. Treatments vary in (i) whether sender and receiver start with having conflicting or aligned interests in all 30 rounds of part one and (ii) whether they subsequently move on to having partially aligned interests throughout all the remaining 30 rounds or only oc-

²Byrne et al. (2021) use a field experiment on shower water usage to contrast consumption habits with attention habits and find evidence supporting the attention mechanism.

asionally so (randomly in 10 out of 30 rounds). Our primary data are the choices of subjects, i.e. sender messages and receiver actions. Additionally, we record decision times, we measure cognitive ability (via the CRT), and we elicit risk attitudes and trust attitudes.

Part one facilitates the formation of different communication habits. We use 30 rounds as habit formation requires long repetition in a stable environment (Wood and Runger, 2016). We are interested in the effect of the (potentially) formed habits on the behaviour in the unfamiliar environment with partially aligned preferences. We hypothesize that communication will be more informative for subjects who started with the aligned environment than for subjects who started with the conflicting environment. We measure the informativeness of the communication by the correlation between states and actions. Part two varies how often subjects interact in the unfamiliar environment. If the preference formation mechanism dominates, we would expect to see a treatment effect (higher correlations after aligned than after conflicting environment) irrespective of how often the new environment occurs. If inattention dominates, we would expect to see a treatment effect when the new environment occurs rarely, but not when it occurs frequently. By varying the frequency of the environment with partially aligned interests, we experimentally manipulate the salience of the change in preference alignment. This variation allows us to compare the strength of those two mechanisms. Additional measures such as decision times and CRT scores also shed light on the mechanisms.

Our main finding is that (on the aggregate level) communication under partially aligned interests is more informative for subjects who started with common-interests in part one, but only if they face the new environment rarely. This effect persists over time. When the new environment occurs frequently, subjects quickly adapt their behaviour and we find no difference in the informativeness of communication. Thus, our evidence is consistent with inattention rather than preference formation. Additionally, the actual correlations between action and state provide a point estimate of the informativeness of communication. We find that, compared to the most informative equilibrium, subjects who started with the common-interest environment overcommunicate in the partial aligned case whereas subjects who started with the conflicting-interest environment undercommunicate.

To better understand behaviour at the individual level, we classify subjects as habitual if their choices satisfy two conditions: (i) they use a stable strategy for the majority of decisions in part one, and (ii) they use the same strategy when interacting in the new environment. This classification reveals interesting patterns. First, more subjects are classified as habitual if they started with common-interest environment, which suggests that full alignment of preferences provided a simpler environment than fully conflicting and stronger habits were formed. Second, habitual subjects

make decisions faster and have lower CRT scores, further suggesting that inattention increases the likelihood on relying on habit as a heuristic. Third, habitual subjects earn slightly less than non-habitual subjects, suggesting that reliance on habits was moderately costly.³

Our chapter speaks to various strands of research. First, it is part of the economic literature on habit formation.⁴ Most of the studies focus on consumption habits, and, more specifically, on the effect of past consumption on future consumption (see Havranek et al. (2017) for a literature review and meta analysis of relevant studies). Empirical evidence also documents saving habits (De Mel et al., 2013; Schaner, 2018), exercising habits (Charness and Gneezy, 2009; Acland and Levy, 2015; Royer et al., 2015) and voting habits (Gerber et al., 2003; Meredith et al., 2009; Coppock and Green, 2016; Fujiwara et al., 2016).⁵ Closest to our design is Peysakhovich and Rand (2016). Motivated to explain the heterogeneity of prosocial preferences, they also use a two stage experiment. In stage one, they experimentally create norms of cooperation and defection by letting subjects play repeated prisoner dilemma games with either high or low continuation probabilities. In stage two, they elicit choices in a range of prosocial one-shot games like trust game, ultimatum game and dictator game. They find that subjects from the cooperative environment exhibit higher levels of prosociality. Our design is parallel to theirs in the setting of strategic information transmission, but allows for testing long term persistence since in our experiment the new environment occurs more than once. Our contribution to this literature is providing experimental evidence of habit formation in strategic communication.

A closely related research question is explored in Belot and van de Ven (2019). They expose subjects to either low and high incentives to lie in a sender–receiver game and reverse the incentives halfway through the experiment. They find no evidence of persistency of either honest or dishonest communication. Their design is similar to one of our treatments, namely the one where the shift to the new environment is permanent. A notable difference is that in their experiment subjects played 12-14 rounds in total whereas in ours they played 60 rounds. As also mentioned in their discussion, habit formation takes time and their shorter experiment may not

³This pattern is consistent with the general principles of rational inattention models. Such models essentially assume a trade-off between the cognitive cost of adjusting strategy and the cost of sticking to the same strategy (Sims, 2003; Caplin, 2016).

⁴There is also economic theory literature on habit formation, primarily aimed at relaxing the assumption of time-separable preferences. See for example Rozen (2010) and Chetty and Szeidl (2016).

⁵Related are also articles which study history-dependence and behavioural spillovers. Romero (2015) compares coordination in a minimum-effort game and finds higher effort levels when the cost parameter increased to a given value compared to when it decreased to the same value. Buser and Dreber (2016) find that subjects participating in a tournament paying scheme contribute less in a subsequent public good game than subjects paid on a fixed piece rate. Herz and Taubinsky (2018) show that subjects familiar with higher prices judge high prices as more fair compared to subjects familiar with lower prices.

have been able to facilitate it. However, we also find no effect of past experience when the change of environment is permanent, despite utilising a longer experiment where habits could be (and actually are) formed. Hence, their results are strengthened in light of our results.

Second, our results speak to the literature documenting communication differences between individuals (Sánchez-Pagés and Vorsatz, 2007; Hurkens and Kartik, 2009; Serota et al., 2010) as well as between groups such as countries (Holm and Kawagoe, 2010; Robert and Arnab, 2013; Pascual-Ezama et al., 2015; Hugh-Jones, 2016), occupation (Cohn et al., 2014, 2015) and religiosity (Arbel et al., 2014).⁶ With a randomised controlled experiment, we present evidence for a causal link between past environment and communication in a new environment. Thus, we document that habits can solidify communication differences and can (partially) explain the stickiness of these differences in atypical situations. Our chapter, therefore, complements those studies and provides a habit formation interpretation of how such differences may have emerged.

Third, our chapter belongs in the line of experimental cheap-talk games. Starting from Dickhaut et al. (1995), a long list of experiments have investigated the comparative statics of Crawford and Sobel (1982). A common finding is overcommunication; subjects typically communicate more information than the most informative equilibrium predicted by theory (Cai and Wang, 2006; Sánchez-Pagés and Vorsatz, 2007; Wang et al., 2010; De Haan et al., 2015).⁷ Our design allows us to test the conjecture that overcommunication is observed because subjects are used to common-interest environments outside of the lab. Such environments facilitate the formation of habits of honest informative communication. When participating in an experiment, subjects may carry this disposition towards honest communication with them. By varying their past experience, we observe both overcommunication and undercommunication, which is consistent with the conjecture.

A notable exception to the common overcommunication finding is Cabrales et al. (2020), who have also documented undercommunication in a cheap talk experiment. In their experiment, they introduce a market for information and vary whether the traded information is verifiable or not. They find that when information is unverifiable -as is the case in our experiment-, the level of market activity is much lower than equilibrium predictions whereas when information is verifiable, the level of market activity is similar to equilibrium. Our experiments differ substantially. In their experiment, information acquisition is costly and an auction mechanism determines whether information is sold or not. In our experiment, information is freely

⁶There is also mixed evidence for gender differences in lying aversion. Rosenbaum et al. (2014) presents a comprehensive literature review of experiments on honesty and discusses heterogeneity across various dimensions.

⁷A comprehensive literature review of experimental cheap talk games can be found in Blume et al. (2020).

observable by the sender and they always send a free message to the receiver, thus, eliminating any direct cost of information for both agents.

The results of our study arguably have some broader implications. We find that heterogeneity in communication can be (partially) attributed to familiarity with communicating in common-interest or conflicting-interest environments. Peysakhovich and Rand (2016) show that heterogeneity in cooperation can be attributed to familiarity with interacting in more cooperative or less cooperative environments. We view those findings as evidence that habits affect behaviour in a wide range of situations. Thus, we need to take into account how familiar agents are with a given situation when studying human behaviour. This is particularly important when the degree of familiarity is low and decisions may be influenced by sufficiently similar everyday activities where habits are formed. To enhance our understanding of habits, it is fruitful to study habit formation both empirically and experimentally in different domains. At the same time, incorporating habitual behaviour in theoretical models will also help us better predict behaviour, especially when deriving predictions for relatively rarely occurring situations.⁸

The remaining of the chapter is organized as follows. section 2.2 provides a detailed presentation of the sender-receiver game and equilibrium predictions, the experimental design, and the predictions. All results are presented in section 2.3. We end the chapter with section 2.4 which interprets the results, positions the contributions and suggests areas for future research.

2.2 Design & Predictions

2.2.1 The sender-receiver game

The experiment considers a discrete cheap talk game with five possible states. In the beginning of each round, the state of the world (s) is uniformly drawn from the set $S = \{1, 2, 3, 4, 5\}$. The prior distribution is commonly known. The sender privately observes the draw and has to send a message (m) to the receiver. The possible messages are of the form “*The state is m* ”, where $m \in M = \{1, 2, 3, 4, 5\}$. The receiver is uninformed about the true state of the world. After observing the sender’s message, the receiver chooses an action (a) from the set $A = \{1, 2, 3, 4, 5\}$. The action determines the payoffs of both players and ends the round.

The payoffs depend only on the state and the action (and not on the message),

⁸Theoretical models in this direction are Samuelson (2001) and Jehiel (2005).

and are given below.⁹

$$U^S(a, s, b) = 110 - 20|s + b - a|^{1.4} \text{ and } U^R(a, s) = 110 - 20|s - a|^{1.4}$$

From the (induced) utility functions, it is clear that the receiver optimally wants to match the true state ($a = s$) whereas the sender wants the receiver to choose an action higher than the state ($a = s + b$). Thus, the parameter b naturally captures the alignment of interests between the sender and the receiver; the larger b , the larger the sender's bias is.

2.2.2 Perfect Bayesian equilibria of the sender-receiver game

Crawford and Sobel (1982) analyzed such games and showed that all equilibria are partition equilibria. In such an equilibrium, the sender partitions the state space and randomly selects one message from each element of the partition. The larger the bias parameter, the more coarse the partition is. In other words, less information is revealed by the sender and less faith is placed in the message by the receiver when their preferences are less aligned. Typically there exist multiple equilibria for each value of b . Crawford and Sobel (1982) showed that the most informative equilibrium is Pareto superior to all other equilibria.¹⁰

In our treatments, we use three bias values. The fully aligned environment corresponds to $b = 0.2$, the partially aligned corresponds to $b = 1$, and the fully conflicting corresponds to $b = 2$. Table 2.1 lists all the perfect Bayesian equilibria for the values of b used in our treatments.¹¹ The equilibria are Pareto ranked with the last equilibrium for each bias value being the most informative as well as the most profitable.

Each row of the table represents one equilibrium. The *Messages* column describes the sender's partition of the state space. The *Actions* column describes the receiver's partition of the message space. For example, the second row when $b = 1$ is to be read as follows. The sender partitions the state space into two elements, $\{1\}$ and $\{2, 3, 4, 5\}$. If the state is 1, the sender sends the message "The state is 1". If the state is either 2,3,4 or 5, the senders randomly sends a message between "The state is 2", "The state is 3", "The state is 4" and "The state is 5". In this equilibrium, the message "The state is 1" is followed by the receiver who chooses action 1. Any other message is interpreted as carrying the information that the true state is equally likely

⁹The payoff functions are taken from Cai and Wang (2006) and Wang et al. (2010). The value of 1.4 in the exponent is used to enhance the magnitude of payoff differences across receiver actions. Cai and Wang (2006) used various values as a robustness check with similar results. You can see the table version of the payoffs in subsection 2.5.2.

¹⁰While some equilibrium selection criteria are too strict and eliminate all equilibria in Crawford-Sobel like games (Matthews et al., 1991; Farrell, 1993), criteria that do select an equilibrium, typically select the most informative one (Chen et al., 2008; de Groot Ruiz et al., 2015).

¹¹In subsection 2.5.1.1 of the Appendix, we present the full list of equilibria for all positive values of b .

$b = 0.2$	Messages	Actions	Corr(S,A)
1	{1, 2, 3, 4, 5}	{3}	0.00
2	{1, 2}, {3, 4, 5}	{1, 2}, {4}	0.84
3	{1, 2, 3}, {4, 5}	{2}, {4, 5}	0.84
4	{1}, {2, 3}, {4, 5}	{1}, {2, 3}, {4, 5}	0.90
5	{1, 2}, {3}, {4, 5}	{1, 2}, {3}, {4, 5}	0.90
6	{1, 2}, {3, 4}, {5}	{1, 2}, {3, 4}, {5}	0.90
7	{1}, {2}, {3}, {4, 5}	{1}, {2}, {3}, {4, 5}	0.95
8	{1}, {2}, {3, 4}, {5}	{1}, {2}, {3, 4}, {5}	0.95
9	{1}, {2, 3}, {4}, {5}	{1}, {2, 3}, {4}, {5}	0.95
10	{1, 2}, {3}, {4}, {5}	{1, 2}, {3}, {4}, {5}	0.95
11	{1}, {2}, {3}, {4}, {5}	{1}, {2}, {3}, {4}, {5}	1.00
$b = 1.0$	Messages	Actions	Corr(S,A)
1	{1, 2, 3, 4, 5}	{3}	0.00
2	{1}, {2, 3, 4, 5}	{1}, {3, 4}	0.65
$b = 2.0$	Messages	Actions	Corr(S,A)
1	{1, 2, 3, 4, 5}	{3}	0.00

Note: Rows describe the equilibria. Corr(S,A) is correlation between state and action.

Table 2.1: Perfect Bayesian Nash equilibria for values of b used in the experiment

to be anywhere between 2 and 5. In that case, the best response of the receiver is to choose action 3 or action 4 with equal probabilities.

The table is augmented with the correlation between state and action in each equilibrium. We use the correlation as our measure of the informativeness of communication (henceforth just correlation). The correlation ranges from 0 for uninformative communication to 1 for fully informative communication.¹²

2.2.3 Treatments

The subjects play 60 rounds of a sender-receiver game. The rounds are split in part one (rounds 1-30) and part two (rounds 31-60). We use a 2×2 between subjects design varying the value of the bias parameter in the two parts. It is important to emphasize that the subjects are only aware that they will play 60 rounds, but not that there are two parts.

Part one is either *Aligned* or *Conflict*. In *Aligned*, the subjects play 30 rounds with a fixed bias parameter of $b = 0.2$. In *Conflict*, the subjects play 30 rounds with a fixed bias parameter of $b = 2$. Part two is either *Rare* or *Frequent*. In *Frequent*, the subjects play all rounds with a bias parameter of $b = 1$. In *Rare*, the subjects play in random order 10 rounds with $b = 1$ and 20 rounds with the same bias parameter as in part one ($b = 0.2$ if they started with *Aligned* and $b = 2$ if they started with *Conflict*).

¹²The choice of correlation between states and actions as a measure of informativeness is motivated by previous experimental literature to facilitate comparisons (Cai and Wang, 2006; Kawagoe and Takizawa, 2009; Wang et al., 2010).

The random draws of rounds with $b = 1$ had been done beforehand and was kept constant across (the Rare) sessions. Overall, our design has four treatments, namely *Aligned-Rare*, *Aligned-Frequent*, *Conflict-Rare*, *Conflict-Frequent*. They are visualised in Figure 2.1.

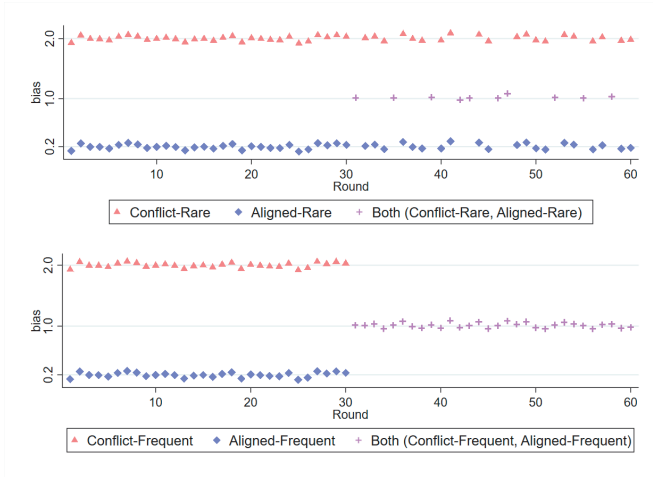


Figure 2.1: Bias per round for each treatment

As can be seen from the figure, in each round the payoffs were slightly perturbed with a small noise. The noise was chosen to be small so that the overall incentive structure was not affected. This was to avoid experimental demand effect when the underlying bias changed. Without the noise, the bias would change after being the same for 30 rounds. This could alert subjects and they would arguably think they are supposed to make different choices. With the small noise, their utility functions slightly change in every round (and more sharply change when the underlying bias also changes).

2.2.4 Predictions

We are interested in the behaviour of subjects in the unfamiliar new environment where $b = 1$. More specifically, we want to test whether starting in a common-interest environment results in more informative communication in the new environment compared to starting in a conflicting-interests environment. To test whether interacting in a new environment less frequently results in more habitual behaviour, we compare correlations two times: (i) between Aligned-Rare and Conflict-Rare, and (ii) between Aligned-Frequent and Conflict-Frequent. To test for the persistence of the effect, we compare the correlations in the early (first 5) and in late rounds (next 5) of $b = 1$.

We use two-sided tests for all our hypotheses. We present our predictions for the direction of the effects.

Prediction 2.1 (Habitual communication when new environment occurs rarely).

- (a) *Correlation in Aligned-Rare will be higher than in Conflict-Rare in early rounds.*
- (b) *Correlation in Aligned-Rare will be higher than in Conflict-Rare in late rounds.*

Prediction 2.2 (Habitual communication when new environment occurs frequently).

- (a) *Correlation in Aligned-Frequent will be higher than in Conflict-Frequent in early rounds.*
- (b) *Correlation in Aligned-Frequent will be higher than in Conflict-Frequent in late rounds.*

A secondary set of predictions is related to the absolute levels of correlation in each treatment. We predict that the correlation after aligned (conflicting) environment will be higher (lower) than in the (most informative) equilibrium (see Table 2.1).

Prediction 2.3 (Overcommunication and undercommunication).

- (a) *Correlation in Aligned-Rare will be higher than 0.650.*
- (b) *Correlation in Aligned-Frequent will be higher than 0.650.*
- (c) *Correlation in Conflict-Rare will be lower than 0.650.*
- (d) *Correlation in Conflict-Frequent will be lower than 0.650.*

2.2.5 Procedure

The computerised laboratory experiment was conducted in October and November of 2020. All subjects were recruited from the subject pool of the CREED laboratory of the University of Amsterdam. The experiment was programmed in oTree (Chen et al., 2016) and preregistered (Ioannidis, 2020). Each treatment arm used 64 subjects, resulting in 256 subjects in total. Subjects were on average 22 years old (mean = 22.37, SD = 4.29, min = 18, max = 60), primarily Economics students (64%), and evenly balanced across genders (52% female, 48% male). Each subject participated only once. They earned on average €27 (mean = 27.33, SD = 6.12, min = 6.95, max = 35.5) in approximately two hours.

Given that the experiment was run online, connectivity issues could temporarily prevent subjects from accessing the experiment. To avoid delaying the whole session,

a maximum of 180 seconds was allowed per decision. The timer was initially hidden from the subjects and only appeared when they had 30 seconds left.¹³ If a subject failed to make a choice within 180 seconds, they were flagged as inactive. This automatically resulted in 0 points for them in that round. Their partner received 100 points and was informed that their partner was inactive in that round. To ensure the session proceeded without further delays, the maximum time available was reduced by 30 seconds for every round a subject was inactive. Thus, if a subject was inactive for more than five consecutive rounds, they would be removed from the rest of the experiment.¹⁴

16 subjects participated in each session and were randomised into matching groups of eight.¹⁵ Each matching group was randomly assigned to a treatment. Within a matching group, the subjects were randomly assigned a role (i.e. sender or receiver) and kept it throughout the experiment.¹⁶ They were informed that the main experiment will last 60 rounds and that their cumulative earnings from all rounds will be converted to euros at a rate of 200 points per euro. After reading the rules of the sender-receiver game, they had to correctly answer a series of understanding questions.

In the main experiment, they played 60 rounds of the sender-receiver game. They were randomly rematched within their matching group in every round to avoid reputation effects. Eight independent sequences of true states were drawn before the experiment and used for each matching group respectively. The same sequences were used for all treatments to minimise the difference in the variation of true states across all treatments.¹⁷

The payoffs for both players were shown in a table whenever they made their choices; both when the senders were choosing a message and when the receivers were choosing an action. At the end of each round, both players received complete feedback about the true state, the message sent, the action chosen and the realised payoffs of both players. The feedback screen also included the payoff table, allowing the subjects to reflect on their choices.

The experiment ended with three post-experiment questionnaires measuring risk attitudes, cognitive ability, and trust attitudes, as well as a survey of standard demo-

¹³This message was shown in 32 out of 15360 decision screens and in 236 out of 15360 feedback screens.

¹⁴No subject was removed due to technical issues. In total four senders and ten receivers (not paired with each other) were inactive for one round, and two receivers were inactive for two rounds. Thus, we later remove 18 observations from our analysis.

¹⁵Due to attendance issues, two sessions had only 14 subjects and two sessions had 18 subjects. Thus, two matching groups have less subjects (six) and two matching groups have more subjects (ten).

¹⁶To avoid framing, in the experiment players were referred as player A (sender) and player B (receiver).

¹⁷To ensure that Aligned-Rare and Conflict-Rare treatments are as comparable as possible, we fixed the rounds in which $b = 1$ across all matching groups. The rounds in which $b = 1$ were 31, 35, 39, 42, 43, 46, 47, 52, 55 and 58.

graphics (age, gender, field of study).

The first questionnaire measured risk attitudes using the lottery method of Eckel and Grossman (2002).¹⁸ The subjects had to choose from a series of lotteries whose expected payoff increases with variance. Their choice was incentivised, and realised by the computer. Given the informational asymmetry of the interaction, controlling for risk is necessary as, for example, risk averse receivers may choose the ex-ante optimal action ($a = 3$).

The second questionnaire measured cognitive ability using the Cognitive Reflection Test (CRT) of Frederick (2005). CRT consists of questions with intuitive, but wrong, answers and measures the tendency to override intuition and deliberately reflect on the correct answer. It has been shown to correlate with the tendency to rely on heuristics (Welsh et al., 2013) and can predict rational thinking in a range of tasks (Toplak et al., 2014). To avoid subjects being familiar with the questions from participation in previous experiments, we used a modified set of questions.¹⁹ Measuring cognitive ability is interesting as subjects with lower CRT may over-rely on habits, thus adapting their behaviour less in the rounds where they play the unfamiliar game ($b = 1$). The CRT was also incentivised.

The third questionnaire measured general trust attitudes towards strangers. We used two questions adapted from the World Values Survey (Glaeser et al., 2000), namely: (i) “When we communicate with strangers, we tell them the truth.”, and (ii) “When we communicate with strangers, they tell me the truth”. We used a five-point Likert scale from -2 (strongly disagree) to +2 (strongly agree). Their attitudes were elicited to serve as a proxy for their baseline tendency towards honest communication. All else being equal, subjects who are more trusting towards strangers outside the lab may have a higher chance of sending a truthful message as senders or following a message as receivers.

Finally, decision times were recorded throughout the whole experiment.

2.3 Results

All reported tests are two-sided. All analyses (unless noted otherwise) are done on a matching group level aggregated over rounds to ensure all comparisons use fully

¹⁸We chose this method for the simplicity of implementation. See Charness et al. (2013) for a discussion of different risk elicitation methods.

¹⁹The modified version consists of the following questions: (i) “The ages of Mark and Adam add up to 28 years total. Mark is 20 years older than Adam. How many years old is Adam?”, (ii) “If it takes 10 seconds for 10 printers to print out 10 pages of paper, how many seconds will it take 50 printers to print out 50 pages of paper?”, and (iii) “On a loaf of bread, there is a patch of mould. Every day, the patch doubles in size. If it takes 12 days for the patch to cover the entire loaf of bread, how many days would it take for the patch to cover half of the loaf of bread?” (Shenhav et al., 2012; Peysakhovich and Rand, 2016).

independent observations. Results based on alternative specifications are included in the appendix as robustness checks.

2.3.1 Manipulation check: Differences in behaviour in part one

This subsection documents the successful manipulation in part one of the experiment. Two pieces of evidence are presented to support this claim, namely correlations and decision times. For this subsection, which is based on data from part one only, we merge Aligned-Rare and Aligned-Conflict treatments, and Conflict-Rare with Conflict-Frequent treatments and refer to these as Aligned and Conflict environments.

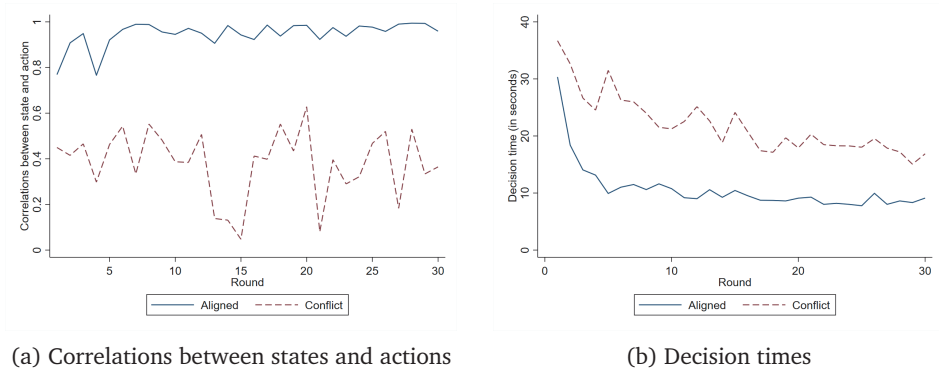


Figure 2.2: Correlations and decision times in part one between Aligned and Conflict

The choices of subjects, as expected, differ dramatically between environments. This is in line with the different incentive structure of Aligned versus Conflict. Figure 2.2a shows the correlations between states and actions over the first 30 rounds. The average correlation in the Aligned environment is higher (mean = 0.953, $N = 16$) than the correlation in the Conflict environment (mean = 0.387, $N = 16$) and the difference is highly significant (Wilcoxon ranksum test, $z = -4.753$, $p < 0.001$, $N = 32$).²⁰ Thus, subjects communicated more informatively in the Aligned environment compared to the Conflict environment.

Decision times differ between environments and decrease over the rounds. This is evident from Figure 2.2b. The average decision time in the Aligned environment was 10.73 seconds and in the Conflict environment 22.04 seconds. This difference is highly significant (Wilcoxon ranksum test, $z = 4.711$, $p < 0.001$, $N = 32$). This observation is also confirmed in a regression of decision time on round and on environment, with errors clustered at the matching group level. The slope of the en-

²⁰In subsection 2.5.1.2, we present tests based also on correlations between states and messages and between messages and actions. The same pattern is observed. We also compare our results with previous experimental results and show that past findings replicate.

vironment is significantly negative ($b = -11.3$, $SE = 1.49$, $CI = [-14.34, -8.27]$, $t = -7.59$, $p < 0.001$, $N = 960$), indicating that subjects decided faster in the Aligned environment than in the Conflict environment. Additionally, the slope of round is also significantly negative ($b = -0.41$, $SE = 0.50$, $CI = [-0.51, -0.31]$, $t = -8.14$, $p < 0.001$, $N = 960$), indicating that, within each environment, subjects decided faster over time.²¹

The difference in decision times reflects the difference in the complexity of the environment. In Aligned, subjects coordinated on the fully revealing equilibrium very fast and their choices were almost automatic. In Conflict, the decision is more complicated due to the preference misalignment, so subjects took more time to figure out what to do. As a side observation to further exemplify the difference in complexity of the environments, subjects in Conflict spent on average 29.6 seconds looking at the feedback screen whereas subjects in Aligned only spent 21.20 seconds. The difference is significant (Wilcoxon ranksum test, $z = 4.108$, $p = 0.005$, $N = 32$), but not affected by rounds.

2.3.2 Treatment effects: Comparing communication after aligned vs conflict

We now turn to the main questions of interest: (i) is communication in the new environment with partially aligned interests more informative for subjects who started with Aligned versus Conflict environment in part one?, (ii) is the effect stronger when subjects face the new environment sporadically (in Rare) versus permanently (in Frequent)?, and (iii) do those differences persist over time?

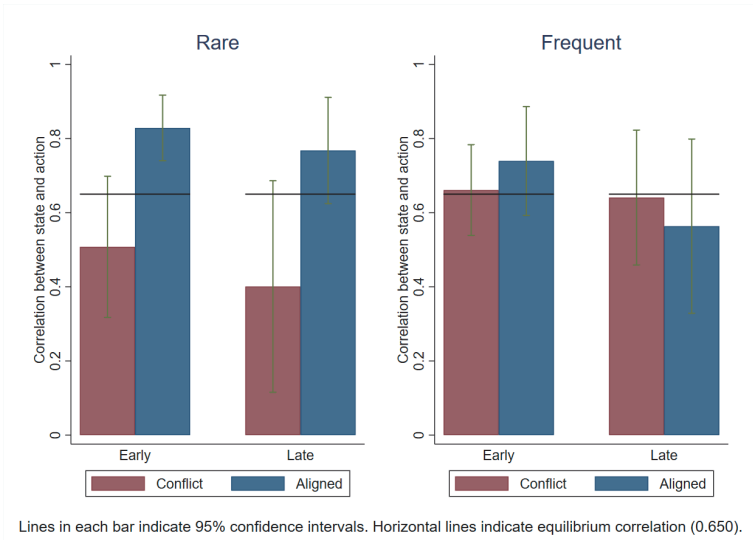
To answer these questions, we compare correlations when $b = 1$ between Aligned and Conflict environments. As illustrated in Prediction 2.1 and Prediction 2.2, the comparison is done separately for early rounds and for late rounds of part two. For the Rare case, subjects faced $b = 1$ only 10 times. We define as early rounds the first five (31, 35, 39, 42, 43) where they did so and as late the last five (46, 47, 52, 55, 58). For the Frequent case, subjects faced $b=1$ in all 30 rounds of the part two. There we define rounds 31-35 as early and 36-40 as late.²²

All comparisons are visualised in Figure 2.3. Within each treatment, correlations are presented separately for early and for late rounds. We note here that both in Figure 2.3 as well as the analysis in this section, we use aggregated observations (both over subjects in a matching group and over rounds) to ensure all comparisons use independent observations. This approach leaves us with eight independent ob-

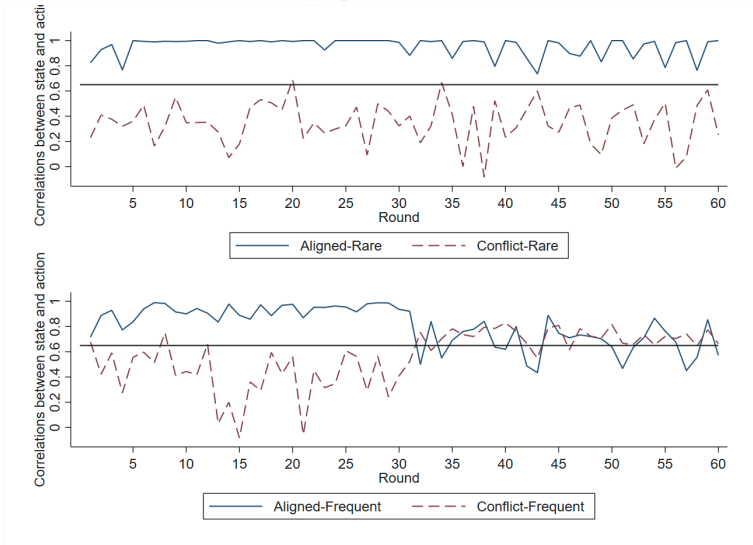
²¹Regressions were also performed on individual level. Those regressions included control variables (risk, CRT, trust, age, gender, study). The conclusions remain the same. Regressions can be found in Table 2.6.

²²The results are qualitatively the same if we define rounds 56-60 as late.

servations per treatment. The upside is that differences which are significant with this conservative approach indicate very high confidence in the treatment effect. The downside is that some comparisons may be underpowered. To address the possible low power issue, in subsection 2.5.1.3 we estimate ordered logistic regressions of receiver action on state (while clustering errors at the subject level) and verify that the conclusions presented here remain valid.



(a) Bar graphs of correlations



(b) Correlations over rounds

Figure 2.3: Treatment effects

We first look at the left part of Figure 2.3a. In the early rounds, the correlation in Aligned-Rare is higher (mean = 0.829) than the correlation in the Conflict-Rare (mean = 0.508) and are significantly different (Wilcoxon ranksum test, $z = -2.731$, $p = 0.0047$, $N = 16$). The effect remains sizeable and significant in the late rounds (Aligned-Rare: mean = 0.768, Conflict-Rare: mean = 0.401, Wilcoxon ranksum test, $z = -2.310$, $p = 0.0207$, $N = 16$). The upper graph of Figure 2.3b shows the correlation over the rounds of Aligned-Rare and Conflict-Rare. We see that the correlation after Aligned-Rare remained higher than after Conflict-Rare, further illustrating the treatment effect when the new environment occurs rarely.

Result 1. *Communication in early rounds is more informative in Aligned-Rare than in Conflict-Rare treatment. The effect persists over time.*

We now turn to the right part of Figure 2.3a. In the early rounds, the correlation in Aligned-Frequent (mean = 0.734) and the correlation in the Conflict-Frequent (mean = 0.661) are not significantly different (Wilcoxon ranksum test, $z = -0.945$, $p = 0.3823$, $N = 16$). In the late rounds, the correlation in Aligned-Frequent treatment (mean = 0.564) and the correlation in the Conflict-Frequent treatment (mean = 0.641) are also not significantly different (Wilcoxon ranksum test, $z = 0.525$, $p = 0.6454$, $N = 16$). The null effect is further illustrated in the bottom part of Figure 2.3b.

Result 2. *There is no difference in the informativeness of communication between the Aligned-Frequent and the Conflict-Frequent treatments, neither in early nor in late rounds.*

2.3.3 Overcommunication and undercommunication

We now turn our attention to the absolute levels of the correlations and test for overcommunication and undercommunication. We compare the observed correlations in all treatments with the equilibrium predicted correlation. As seen in Table 2.1, when $b = 1$, the most informative equilibrium has a correlation of 0.650. Prediction 2.3 suggests that the observed correlations will be higher than 0.650 after Aligned and lower than 0.650 after Conflict. The comparison is also visualised in Figure 2.3, where the horizontal black lines are at the equilibrium level of 0.650.

The comparison is performed by a signtest.²³ In early rounds of the Aligned-Rare treatment, the correlation is higher (mean = 0.829, signtest, $p = 0.0039$, $N = 8$) whereas for Conflict-Rare the correlation is lower (mean = 0.508, signtest, $p = 0.1445$, $N = 8$) than 0.650. The same pattern is observed in later rounds (Aligned-Rare: mean = 0.768, signtest, $p = 0.1445$, $N = 8$, Conflict-Rare: mean = 0.401,

²³Results from the regression method suggested by (Cai and Wang, 2006, footnote 12) are presented in subsection 2.5.1.4. All conclusions remain valid with this alternative method.

signtest, $p = 0.0352$, $N = 8$). All tests find no evidence that the correlation differs from 0.650 (p-values are between 0.363 and 0.634) in either early or late of Aligned-Frequent and Conflict-Frequent treatments.

Result 3. *Overcommunication is observed in Aligned-Rare treatment and undercommunication in Conflict-Rare. The informativeness of communication in the Aligned-Frequent and Conflict-Frequent treatments does not differ from equilibrium predictions.*

2.3.4 Habit formation and inattention at the individual level

The results presented so far are based on aggregate data. In this subsection, we look more closely in individual decisions to better understand habitual behaviour. We are interested in two sets of comparisons. First, we want to compare the tendency to behave habitually across treatments. More specifically, to answer the questions (i) does starting from the simpler aligned environment result in stronger reliance on habits?, and (ii) do subjects rely more on habits when the new environment occurs rarely compared to frequently?. Second, we want to compare individual characteristics between habitual and non-habitual subjects such as (iii) do habitual subjects make decisions faster, (iv) do habitual subjects have lower cognitive ability, (v) is relying on habits financially costly?, and (vi) are there more habitual receivers than senders?

To make those comparisons, we first need a method to classify subjects into habitual and non-habitual. We apply a two-step procedure to do so. In the first step, we apply the psychology definition of habits. Habits are characterised by high automaticity and reduced dependence on goals (Wood and Runger, 2016). We operationalise the definition into two requirements. High automaticity requires that subjects converge to a stable strategy in part one. The habit formation process takes time (Lally et al., 2010). To account for this, we ignore the first 10 rounds where subjects could potentially still be using trial and error. Reduced goal dependence requires that subjects relied on the same strategy in part two as they did in part one, despite the change in the preference alignment. A subject is classified as *habitual* if their choices satisfy *both* requirements.

We take a data-driven approach to identify behavioural strategies. The set of possible strategies we consider is not restricted to a particular theoretical model. For example, in similar experiments, individual decision analysis typically focused on level- k type classification of behavioural types (Cai and Wang, 2006; Wang et al., 2010). With our procedure, additional strategies are also included. For example, when $b = 2$, no level- k prediction would imply that senders should exaggerate the true state by one. L0 senders would tell the truth, L1 senders should exaggerate by two since they believe they are facing credulous receivers, and higher levels would

exaggerate even more.²⁴ We apply our classification method on rounds 11-30 of the first part and on the 10 rounds of part two where subjects played in the new environment.

We consider all possible pure strategies that can exist in the game. For senders, for each of the five possible states, they can choose among five possible messages, resulting in 3,125 possible strategies. Symmetrically, for receivers, for each of the five possible messages they receive, they can choose among five possible actions, also resulting in 3,125 possible strategies. Next, we compute the percentage of decisions consistent with each of the strategies. Eligible strategies are those that are consistent with at least 60% of subject choices. This threshold is used as a compliance rate in behavioural type analysis of sender-receiver games in (Cai and Wang, 2006; Wang et al., 2010). Among the eligible strategies (if any), we select the one with the highest percentage. The compliance rate of 60% is used for both part one and part two.

We can successfully identify behavioural strategies for 228 subjects (out of 256) for part one and for 236 subjects for part two. In total six subjects remain unclassified in both part one and part two and, consequently, are classified as non-habitual. However, those subjects could have formed the habit of *being unpredictable* by using a mixed strategy. To account for the possibility of habitual mixing, we augment our procedure with a second step which attempts to correct for this limitation. We estimate a regression of choice on cue (for senders this is message on state, for receivers this is action on message) including data from both parts and incorporate an interaction effect to allow for different slopes across parts. Formally, we estimate the following regression

$$\text{Choice}_i = \beta_0 + \beta_1 * \text{Cue}_i + \beta_2 * \text{Part}_i + \beta_3 * \text{Part}_i * \text{Cue}_i + \epsilon$$

If β_2 and β_3 are jointly significant, then the subject changed strategy. If not, then the subject used the same strategy and is classified as habitual. Our second step essentially equates habitual behaviour with (statistically) similarly informative choices between part one and part two.

Table 2.2 below shows the number of habitual subjects across treatments.²⁵ We remind the reader that there are 32 senders and 32 receivers in each treatment. In total 112 subjects are classified as habitual.

²⁴Other econometric methods to estimate behavioural strategies are the Structural Frequency Estimation Method of Dal Bó and Fréchet (2011) and the spike-logit model of Costa-Gomes and Crawford (2006). In those methods, the set of candidate strategies is predefined. Costa-Gomes and Crawford (2006) consider whether alternative strategies (pseudo-types) provide a better fit than the original strategies as a robustness check for their classifications. Our method has a similar intuition in the sense that we consider every possible strategy and choose the best fitting one.

²⁵The full lists of strategies (for both habitual and non-habitual subjects, and for both part one and part two) are presented in subsection 2.5.1.5. In subsection subsection 2.5.1.6 we also redo our analysis using a threshold of 80%. With the higher threshold, essentially we require an even higher automaticity. All results presented here are qualitatively the same.

Role	Treatment			
	A-F	A-R	C-F	C-R
Sender	14	11	10	13
Receiver	16	25	11	12
Total	30	36	21	25

Treatment abbreviations:

A-F = Aligned-Frequent

A-R = Aligned-Rare

C-F = Conflict-Frequent

C-R = Conflict-Rare

Table 2.2: Habitual subjects per treatment

First, we look at the effect of the complexity of the initial environment on habit formation. Aggregated, in Aligned-Frequent and Aligned-Rare 66 out of 128 subjects behaved habitually compared to 46 out of 128 Conflict-Frequent and Conflict-Rare (proportion test, $z = 2.5198$, $p = 0.0117$, $N = 256$). Thus, more subjects relied on habits if they started with common-interest environments compared to conflicting-interest environments. This observation suggests that the simplicity of the common interest environment facilitated the formation of stronger habits and is in line with psychology findings on the effect of complexity on habit formation (Wood et al., 2002; Verplanken, 2006). In more complex environments, like Conflict-Frequent and Conflict-Rare in our experiment, reaching a stable strategy is harder.

Second, we are interested to see whether subjects are more likely to rely on habit when they face the new environment rarely compared to frequently. Our data are in the expected direction, but the difference is not significant. Taken together, in Aligned-Rare and Conflict-Rare 61 out of 128 subjects behaved habitually compared to 51 out of 128 in Conflict-Frequent and Conflict-Rare (proportion test, $z = 1.2599$, $p = 0.2077$, $N = 256$). Third, we would expect habitual subjects to decide faster in the new environment. This is clearly supported by our data. When facing the new environment, habitual subjects on average made decisions in 13.47 seconds whereas non-habitual subjects made decisions in 16.47 seconds (Wilcoxon ranksum test, $z = 2.799$, $p\text{-value} = 0.0051$, $N = 256$). To have a benchmark on their decision times from part one, we can look at the difference in decision time between part one and part two. Overall subjects who started in common-interests environment *increased* their decision times by 5.75 seconds whereas subjects who started in conflicting-interests environment *decreased* their decision times by 3.67 seconds.

Separately comparing time differences between habitual and non-habitual subjects for each treatment reveals an interesting pattern. In Aligned-Frequent and Aligned-Rare, decision times of non-habitual subjects increased significantly more than decision times of habitual subjects (Wilcoxon ranksum test, Aligned-Frequent: $z = 2.153$, $p\text{-value} = 0.0313$; Aligned-Rare: Wilcoxon ranksum test, $z = 3.126$, p -

value=0.0018). The pattern is not observed for subjects who started with the conflicting environment as there is no significant difference between subjects who did and subjects who did not rely on habit (Conflict-Frequent, Wilcoxon ranksum test, $z = -1.108$, p-value=0.2678; Conflict-Rare, Wilcoxon ranksum test, $z = 0.777$, p-value=0.4369). This pattern suggests that noticing a change in the environment, which would imply an increase in decision time, was easier for subjects who started with the simple aligned environment compared to subjects who started with the complex conflicting environment.

Fourth, we would expect habitual subjects to have lower CRT scores. CRT is a proxy for the tendency to rely on intuitive choices versus deliberate thinking. Given that overriding habits requires conscious effort, subjects with higher CRT are more likely to adapt their strategies. In line with our expectations, we find that habitual subjects have (weakly) lower CRT scores than non-habitual subjects (habitual: mean = 2.06, $N = 112$, non-habitual: mean = 2.24, $N = 144$, Wilcoxon ranksum test, $z = 1.729$, p-value= 0.0838, $N = 256$).

Fifth, we are interested in whether relying on habits financially hurt subjects. When interacting in the new environment, habitual subjects earned (on average per round) 89.51 points whereas non-habitual subjects earned 90.51. The difference is not statistically significant (Wilcoxon ranksum test, $z = 0.544$, p-value=0.5861, $N = 256$), but more importantly is not economically large.²⁶ This suggests that habits worked relatively well for subjects who relied on them. Thus their choice to not adapt their decision can be considered rational.

Finally, we find that habits persist more among receivers as there are more habitual receivers (62) than habitual senders (40). The difference is significant (Wilcoxon ranksum test, $z = 2.8086$, p-value=0.0050, $N = 256$). The majority of habitual senders are truth-tellers (27) and the majority of habitual receivers are believers (43). It is illustrating to compare earnings between habitual and non-habitual subjects separately for senders and for receivers. For senders, there is significant difference in earnings (habitual: mean = 80.82, $N = 48$, non-habitual: mean = 88.33, $N = 80$, Wilcoxon ranksum test, $z = 3.239$, p-value= 0.0012, $N = 128$). For receivers there is no difference (habitual: mean = 96.64, $N = 64$, non-habitual: mean = 93.24, $N = 64$, Wilcoxon ranksum test, $z = 1.303$, p-value= 0.1924, $N = 128$). This suggests that receivers are not harmed by being credulous due to the presence of habitual truth-tellers.

As a side observation, it is illustrating to further break the group of habitual subjects based on whether they noticed the change in the environment (positive time difference) or not (negative time difference). 41 out of 112 habitual subjects did not

²⁶The same conclusion holds when comparing habitual and non-habitual subjects on the basis of the loss from not playing empirical best response (Wilcoxon ranksum test, $z = 0.861$, p-value=0.3893, $N = 256$).

increase their decision time. 71 out of 112 subjects did increase their decision time but kept using the same strategy. Arguably, failing to even notice a change cannot be rational, even if it did not hurt subjects financially. At the same time, noticing a change and consciously using the same strategy can be rational exactly because it did not hurt subjects financially. Thus, this decomposition suggests that inattention can be both rational and irrational.

Taken together, these observations suggest the following interpretation of the data. Subjects found the common-interests environment simple, quickly stabilized their behaviour (into truth-telling and message-following), and had fast decision times. When the underlying environment changed, the change in payoffs was salient to the subjects that did pay attention as the variance in earnings in part one was very small. Hence, those subjects who changed strategy increased their decision times as thinking about how to adapt requires cognitive effort. Subjects found the conflicting-interests environment complex in part one and had overall higher decision times. Given the difficulty converging to a stable strategy together with the large variance of their round per round earnings in part one, noticing the change in the underlying bias was less salient, resulting in overall even faster decision times than part one, despite the change in preference alignment.

2.4 Concluding remarks

The key takeaways from our chapter are: (i) habits affect strategic communication, and (ii) reliance on communication habits in atypical environments is moderated by the salience of the change in the environment. By randomizing subjects into environments that support either more informative or less informative communication, we facilitated the formation of different communication habits. When communicating in a new unfamiliar environment, roughly one third of our subjects relied on their acquired habit and did not adapt their strategy. We varied the salience of the change in the environment by varying how often subjects communicated in the unfamiliar environment. When the change was salient, we observed a strong treatment effect as subjects familiar with the honest environment communicate more informatively than subjects familiar with the dishonest environment. When the salience was low, we found no significant effect. This pattern suggests that inattention rather than preference formation can explain our data.

Our results provide support for the conjecture that overcommunication is partially attributed to the fact that in daily interactions telling the truth and believing what you hear work well most of the time. Hence, familiarity with environments that support informative communication (outside of the lab) may lead to excessively informative communication when subjects communicate in an experiment (inside the lab). By

creating a counterfactual environment where communicating honestly does not pay off, we observed undercommunication.

Our results suggest that habit formation can explain how differences in honesty can solidify in different groups. To illustrate, different occupations are characterised by different levels of preference alignment. Doctors typically have aligned preferences with their patients whereas judges often have misaligned preferences with suspects. Habit formation suggests that doctors may develop the habit of believing information whereas judges may develop the habit of mistrusting information. When communicating outside of their familiar work environment, they may carry their disposition with them.²⁷

A wealth of evidence shows that people are not much better than chance at accurately judging the truthfulness of information (Bond Jr and DePaulo, 2006). In a recent experiment, (Serra-Garcia and Gneezy, 2021) find that conditional on judging a piece of information as truthful, senders are more likely to share it, and conditional on a piece of information being shared, receivers are more likely to believe it. Having shown that receivers who are mostly exposed to truthful information may form the habit of believing information, our results suggest that their habit can make receivers overly credulous and more susceptible to believing fake news and misinformation. Thus, studying the effect of habits on believing and sharing false information is an interesting avenue for future research.

More broadly, our results suggest that habit formation plays an important role in economic decision making (in our case, strategic information transmission). Thus, it is important to take into account whether a given economic situation we are studying resembles a situation with which agents may be more familiar. Especially when we study less frequent phenomena, reliance on past habits may be a good predictor of behaviour. To illustrate, we discuss two empirical questions that build on the key takeaway from the current chapter. A real estate agent who works in a seller's market (where demand exceeds supply) may develop the habit of negotiating hard as they have high bargaining power. Would they adapt their strategy in situations when supply exceeds demand and how does this depend on how salient the increase in supply is? An investor during prosperous times may develop the habit of investing in high-risk high-return assets. Would they adjust their risk portfolio differently when they rarely receive signals that the economy is slowing down compared to a salient media covered emerging crisis?

²⁷Anecdotally, the competition for the World's Biggest Liar is annually held in a pub in England. Contestants from across the world try to come up with the most convincing lie. The rules forbid lawyers and politicians from participating because "they are judged to be too skilled at telling porkies"(Source: BBC, accessed 03-06-2021.)

2.5 Appendix to Chapter 2

2.5.1 Additional results and robustness checks

This appendix consists of five subsections. First, we list all Bayesian equilibria of the game. Next, we compare our results from part one to previous literature and show that past findings replicate. In the next subsection, we present econometric evidence for our main treatment effects via ordered logistic regressions. We then apply the econometric method of Cai and Wang (2006) as a robustness check for our results on overcommunication and undercommunication. Finally, we present the full classification of subjects in behavioural strategies from both part one and part two and also repeat our analysis with a different threshold for classifying behaviour (80%).

2.5.1.1 All Bayesian equilibria of the game

Table 2.3 lists the complete set of all perfect Bayesian equilibria of the game for all possible values of b . The equilibria are ranked in order of informativeness –as captured by the correlation between state and action– with the last in each parameter range being the most informative.

Table 2.3: All perfect Bayesian Nash equilibria for all values of b

Ranking	Messages	Actions	Corr(S,A)
1	{1, 2, 3, 4, 5}	{3}	0.00
2	{1, 2}, {3, 4, 5}	{1, 2}, {4}	0.84
-	{1, 2, 3}, {4, 5}	{2}, {4, 5}	0.84
-	{1}, {2}, {3, 4, 5}	{1}, {2}, {4}	0.84
3	{1}, {2, 3}, {4, 5}	{1}, {2, 3}, {4, 5}	0.90
-	{1, 2}, {3}, {4, 5}	{1, 2}, {3}, {4, 5}	0.90
-	{1, 2}, {3, 4}, {5}	{1, 2}, {3, 4}, {5}	0.90
4	{1}, {2}, {3}, {4, 5}	{1}, {2}, {3}, {4, 5}	0.95
-	{1}, {2}, {3, 4}, {5}	{1}, {2}, {3, 4}, {5}	0.95
-	{1}, {2, 3}, {4}, {5}	{1}, {2, 3}, {4}, {5}	0.95
-	{1, 2}, {3}, {4}, {5}	{1, 2}, {3}, {4}, {5}	0.95
5	{1}, {2}, {3}, {4}, {5}	{1}, {2}, {3}, {4}, {5}	1.00

(a) $b \in [0, 0.22)$

Ranking	Messages	Actions	Corr(S,A)
1	{1, 2, 3, 4, 5}	{3}	0.00
2	{1}, {2, 3, 4, 5}	{1}, {3, 4}	0.65
3	{1, 2}, {3, 4, 5}	{1, 2}, {4}	0.84
4	{1}, {2}, {3, 4, 5}	{1}, {2}, {4}	0.90
-	{1}, {2, 3}, {4, 5}	{1}, {2, 3}, {4, 5}	0.90
-	{1}, {2}, {3}, {4, 5}	{1}, {2}, {3}, {4, 5}	0.90
5	{1}, {2}, {3}, {4}, {5}	{1}, {2}, {3}, {4}, {5}	1.00

(b) $b \in [0.22, 0.50)$

Ranking	Messages	Actions	Corr(S,A)
1	{1, 2, 3, 4, 5}	{3}	0.00
2	{1}, {2, 3, 4, 5}	{1}, {3, 4}	0.65
3	{1, 2}, {3, 4, 5}	{1, 2}, {4}	0.84

(c) $b \in [0.50, 0.73)$

Ranking	Messages	Actions	Corr(S,A)
1	{1, 2, 3, 4, 5}	{3}	0.00
2	{1}, {2, 3, 4, 5}	{1}, {3, 4}	0.65

(d) $b \in [0.73, 1.28)$

Ranking	Messages	Actions	Corr(S,A)
1	{1, 2, 3, 4, 5}	{3}	0.00

(e) $b \in [1.28, \infty)$

2.5.1.2 Replicating past cheap-talk experimental findings

This subsection serves two goals. First, it illustrates the differences in the behaviour of subjects across aligned and conflict treatments in more detail. Second, it provides evidence replicating past findings in experiments testing the comparative statics of Crawford and Sobel (1982).

Crawford and Sobel (1982) predicts that communication will be more informative with more aligned preferences. Table 2.4 shows the correlations between states and actions, states and messages, and messages and actions in part one. The first pair of columns was presented and discussed in subsection 2.3.1. The other two pairs of columns exhibit the same patterns and serve as a robustness check for the manipulation check. All correlations differ significantly between Aligned and Conflict environment.

N	Environment	Correlation(S,A)		Correlation(S,M)		Correlation(M,A)	
		Observed	Predicted	Observed	Predicted	Observed	Predicted
16	Aligned	0.953	1.000	0.967	1.000	0.982	1.000
16	Conflict	0.387	0.000	0.528	0.000	0.647	0.000

Table 2.4: Correlations between states, messages and actions in part one

At the same time, we observe overcommunication in the conflict treatment as all correlations are significantly larger than zero. This can be seen by comparing actual with predicted correlations in Table 2.4. The results are in line with past experimental findings (Cai and Wang, 2006; Wang et al., 2010). Table 2.5 provides a comparison of results from earlier articles and the current one.²⁸

Bias	Correlation	Current	CW	WSC	Predicted
Low	Corr(S,A)	0.959	0.876	0.86	1.000
	Corr(S,M)	0.972	0.916	0.93	1.000
	Corr(M,A)	0.983	0.965	0.92	1.000
High	Corr(S,A)	0.402	0.207	0.32	0.000
	Corr(S,M)	0.560	0.391	0.34	0.000
	Corr(M,A)	0.650	0.542	0.58	0.000

Notes: CW=Cai and Wang (2006), WSC=Wang et al. (2010)

Table 2.5: Correlations between states, messages and actions in part one

Table 2.6 shows regressions of decision times and time spent on feedback screen, both on individual and on matching group level. They provide the evidence for the conclusions from subsection 2.3.1 that (i) decision times differ between treatments

²⁸Previous articles reported correlations computed based on choices of pairs of subjects (not on matching group level or aggregated over rounds as current chapter). To facilitate comparisons, we do the same in Table 2.5. Comparison data are from Table 3 in Cai and Wang (2006) and from Table 2 in Wang et al. (2010).

and decrease over rounds, and (ii) the feedback times differ between treatments and do not decrease over rounds.

Table 2.6: Decision and feedback times in part one

	Decision Time		Feedback Time	
	Group	Individual	Group	Individual
Aligned	-11.31*** (1.49)	-11.04*** (1.21)	-10.52*** (1.94)	-10.71*** (1.14)
Round	-0.41*** (0.05)	-0.40*** (0.03)	-0.02 (0.07)	-0.02 (0.06)
Risk		-0.07 (0.34)		0.21 (0.35)
CRT		-0.37 (0.66)		0.65 (0.56)
Trust sender		-0.47 (0.73)		-0.04 (0.82)
Trust receiver		-0.34 (0.91)		-0.33 (0.92)
Constant	28.40*** (1.84)	30.36*** (3.90)	30.09*** (1.92)	37.28*** (2.85)
Controls	No	Yes	No	Yes
R^2	0.428	0.109	0.096	0.016
Observations	960	7680	960	7680

Controls: Age Gender Study

Std. Err. adjusted for 256 (32) individual (matching group) clusters

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

2.5.1.3 Econometric tests for treatment effects

In this subsection, we are interested in testing whether starting from then aligned environment in part one leads to more informative communication when interacting in the new environment in part two compared to starting from the conflict environment. To do so, we estimate ordered logistic regressions of action on state and interact state with part one environment. A significant interaction ($State \times Aligned$) translates to more informative communication after aligned environment compared to after conflicting. We estimate separate regressions for when the new environment occurs rarely or frequently, and separate for early and late rounds that it does so. Each regression is estimated using individual choices with errors clustered at subject

level. The regressions control for risk, CRT, trust towards strangers and demographics.

Table 2.7: Ordered logistic regression of action on state

	Action			
	Rare Early	Rare Late	Frequent Early	Frequent Late
State	1.15*** (0.10)	1.16*** (0.11)	1.49*** (0.12)	1.65*** (0.10)
State×Aligned	0.34*** (0.06)	0.32*** (0.06)	0.09 (0.07)	-0.04 (0.06)
Round	0.02 (0.02)	-0.02 (0.02)	-0.05 (0.05)	0.08 (0.05)
Risk	-0.06 (0.06)	0.08 (0.06)	0.05 (0.06)	0.06 (0.06)
CRT	0.08 (0.10)	-0.15 (0.10)	0.12 (0.09)	0.02 (0.12)
Trust sender	0.06 (0.12)	-0.08 (0.13)	0.04 (0.14)	0.15 (0.14)
Trust receiver	0.06 (0.16)	0.18 (0.17)	0.11 (0.15)	0.04 (0.15)
Controls	Yes	Yes	Yes	Yes
Pseudo R^2	0.228	0.227	0.248	0.259
Observations	640	640	640	640

Action refers to receiver’s part two behaviour under partially aligned interests

Controls: Age Gender Study

Std. Err. adjusted for 128 subject clusters

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Our results reveal a treatment effect when the new environment occurs rarely (columns 1 and 2) and a null effect when the new environment occurs frequently (columns 3 and 4). Thus the results in main text are robust.

2.5.1.4 Econometric tests for overcommunication and undercommunication

This section provides robustness checks for the results on overcommunication and undercommunication presented in subsection 2.3.3. To do so, we use the regression method utilised by Cai and Wang (2006).

This method has a standard regression as a starting point. Consider a model $Y = \alpha + \beta X + \epsilon$. The estimator for β is given by $b = \frac{SD_Y}{SD_X} \text{Corr}(X, Y)$, where SD_Y, SD_X are the sample standard deviations of X and Y and $\text{Corr}(X, Y)$ is the correlation between X and Y . To test whether the estimated correlation differs from a theoretical

one (denote the theoretical by σ_{XY}), it suffices to estimate the adjusted model $Y - r_{XY}X = \alpha + \beta X + \epsilon$, where $r_{XY} = \frac{SD_Y}{SD_X} \sigma_{XY}$. The t-test on the estimate of β in the adjusted model allows us to precisely test whether $Corr(X, Y) = \sigma_{XY}$. We estimate those regressions separately for each of the four treatments and separately for early and late rounds. For all regressions, we use the correlation of the most informative equilibrium as the theoretical prediction ($\sigma_{XY} = 0.650$).

Table 2.8: Regressions of (adjusted) action on state

	Action							
	CR Early	CR Late	AR Early	AR Late	CF Early	CF Late	AF Early	AF Late
State	-0.14* (0.05)	-0.15** (0.05)	0.19*** (0.03)	0.17*** (0.03)	0.03 (0.04)	0.09** (0.03)	0.10 (0.05)	0.07 (0.04)
Risk	-0.04 (0.04)	0.05 (0.04)	-0.04 (0.03)	-0.02 (0.03)	0.01 (0.04)	-0.02 (0.04)	0.06 (0.04)	0.03 (0.03)
CRT	0.08 (0.09)	-0.09 (0.09)	-0.00 (0.05)	-0.08 (0.05)	0.08 (0.06)	-0.04 (0.06)	0.01 (0.07)	0.11 (0.06)
Trust sender	0.06 (0.08)	-0.07 (0.09)	0.01 (0.07)	0.04 (0.06)	0.06 (0.09)	0.11 (0.07)	0.02 (0.09)	0.04 (0.07)
Trust receiver	0.15 (0.13)	-0.01 (0.13)	-0.09 (0.08)	0.11 (0.09)	0.04 (0.09)	0.06 (0.07)	0.08 (0.10)	-0.10 (0.10)
Constant	1.33* (0.59)	0.75 (0.46)	1.15*** (0.33)	0.99** (0.33)	1.40*** (0.24)	1.26** (0.37)	0.94* (0.44)	1.10*** (0.30)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.073	0.092	0.140	0.133	0.036	0.069	0.047	0.069
Observations	320	320	320	320	320	320	320	320

Action refers to receiver's part two behaviour under partially aligned interests

Controls: Age Gender Study, Std. Err. adjusted for 64 subject clusters, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

When the new environment occurs rarely (first four columns), we see significant differences from equilibrium predictions. When subjects started with conflicting preferences (columns 1 and 2), we observe undercommunication as the coefficient on state is negative. When subjects started with aligned preferences (columns 3 and 4), we observe overcommunication as the coefficient on state is positive. We observe no significant differences when the new environment occurs frequently (with the exception of column 6).

2.5.1.5 Full classification of behavioural strategies

This subsection presents the behavioural strategies in part one and part two of the experiment. Before presenting the results of the classification, we describe all strategies and (in parentheses) their coding.

We present our classification results separately for each treatment, and separately for senders and receivers. We remind the reader that this classification uses 60% as

Strategy	Coding	Strategy	Coding
Tell the truth	Truth	Follow message	Believer
Exaggerate state by 1	State+1	Discount message by 1	Message-1
Exaggerate state by 2	State+2	Discount message by 2	Message-2
Exaggerate state by 3	State+3	One more than message	Message+1
Always send message 4	Always 4	Always choose action 3	Always 3
Always send message 5	Always 5	Always choose action 4	Always 4

(a) Sender strategies

(b) Receiver strategies

Table 2.9: All strategies used

a threshold to classify a subject as using a particular strategy. Habitual subjects are in bold.

	Truth	State+1	State+2	Mixing
Truth	12	17	1	-
State+1	-	1	-	-
Mixing	-	-	-	1

(a) Sender strategies

	Believer	Message-1	Message-2	Unclassified
Believer	16	10	2	3
Unclassified	1	-	-	-

(b) Receiver strategies

Table 2.10: Strategies used in Aligned-Frequent

	Truth	State+1	State+2	Unclassified
Truth	11	17	1	3

(a) Sender strategies

	Believer	Message-1	Unclassified
Believer	25	4	3

(b) Receiver strategies

Table 2.11: Strategies used in Aligned-Rare

	Truth	State+1	State+2
Truth	6	-	-
State+1	2	4	-
State+2	-	10	-
State+3	-	-	1
Always 4	-	1	-
Always 5	-	-	2
Unclassified	1	4	1

(a) Sender strategies

	Believer	Message-1	Message-2	Always 3	Mixing	Unclassified
Believer	6	3	-	-	-	1
Message-1	2	1	-	-	-	-
Message-2	1	7	1	-	-	-
Always 3	-	-	-	2	-	-
Mixing	-	-	-	-	1	-
Unclassified	4	3	-	-	-	-

(b) Receiver strategies

Table 2.12: Strategies used in Conflict-Frequent

	Truth	State+1	State+2	State+3	Always 4	Always 5	Mixing	Unclassified
Truth	2	-	-	-	-	-	-	1
State+1	2	1	1	-	-	-	-	-
State+2	1	8	4	-	-	1	-	-
State+3	2	1	1	-	-	-	-	-
Always 4	1	-	-	-	-	-	-	-
Always 5	1	-	-	1	-	2	-	-
Unclassified	-	1	-	-	1	1	-	-

(a) Sender strategies

	Believer	Message-1	Message-2	Always 3	Unclassified
Believer	6	-	-	-	-
Message-1	-	4	1	-	1
Message-2	3	7	2	-	1
Message+1	1	-	-	-	-
Always 3	-	1	-	-	-
Always 4	-	-	-	-	1
Unclassified	-	1	-	3	-

(b) Receiver strategies

Table 2.13: Strategies used in Conflict-Rare

2.5.1.6 Robustness of habitual classification with respect to threshold

This subsection briefly discusses the results if we increase the threshold to classify a subject into a behavioural strategy from 60% to 80%. Increasing the threshold naturally reduces the number of subjects classified into a behavioural strategy. With 60% as a threshold, we classify 112 subjects as habitual, whereas with 80% we classify 102. This already suggests that the classification is not very sensitive to the chosen threshold.

Consistent with the observations from main text, we find: i) more habitual subjects after aligned environment compared to conflicting (57 VS 45), more habitual subjects when the new environment is rare compared to frequent (57 VS 45), (iii) habitual subjects making faster decisions compared to non-habitual (16.97 seconds VS 12.43 seconds), (iv) habitual having lower CRT scores (2.06 VS 2.23), (v) habitual subjects having slightly lower earnings (88.3 VS 91.2), and (vi) more habitual receivers than senders (62 VS 40).

2.5.2 Payoff tables for different values of bias parameter

	Action is 1	Action is 2	Action is 3	Action is 4	Action is 5
State is 1	110, 113	98, 93	67, 60	28, 19	-16, -26
State is 2	87, 93	110, 113	98, 93	67, 60	28, 19
State is 3	52, 60	87, 93	110, 113	98, 93	67, 60
State is 4	11, 19	52, 60	87, 93	110, 113	98, 93
State is 5	-36, -26	11, 19	52, 60	87, 93	110, 113

Figure 2.4: Payoff tables when $b = 0$

	Action is 1	Action is 2	Action is 3	Action is 4	Action is 5
State is 1	110, 113	98, 93	67, 60	28, 19	-16, -26
State is 2	87, 93	110, 113	98, 93	67, 60	28, 19
State is 3	52, 60	87, 93	110, 113	98, 93	67, 60
State is 4	11, 19	52, 60	87, 93	110, 113	98, 93
State is 5	-36, -26	11, 19	52, 60	87, 93	110, 113

Figure 2.5: Payoff tables when $b = 1$

	Action is 1	Action is 2	Action is 3	Action is 4	Action is 5
State is 1	55, 108	88, 88	108, 55	88, 14	55, -31
State is 2	14, 88	55, 108	88, 88	108, 55	88, 14
State is 3	-31, 55	14, 88	55, 108	88, 88	108, 55
State is 4	-82, 14	-31, 55	14, 88	55, 108	88, 88
State is 5	-137, -31	-82, 14	-31, 55	14, 88	55, 108

Figure 2.6: Payoff tables when $b = 2$

2.5.3 Experimental instructions

Welcome to the session

Welcome!

Thank you for participating in this study. Please make sure that you are in the Zoom meeting throughout the experiment. You were admitted to the session from the waiting room, renamed, and send back to the waiting room. This was to ensure your privacy. If you have any questions, you can message the experiment during the experiment. The Zoom session only allows participants to message the experimenter. Any question you ask and the answer from the experimenter will **not** be shown to any other participant. Please keep your video off and stay muted throughout the experiment.

Payment registration

Please enter your IBAN below. This will be used for payment after the experiment. You will **not** be able to change this at a later point. We will delete this number after making the payment.

Overview of the experiment

Welcome!

Welcome to this experiment. Please read the following instructions carefully. We ask that you do not communicate with other participants during the experiment. The use of mobile phones is not allowed during this experiment. If you have any questions, or need assistance of any kind, at any time, please message the experimenter privately in the Zoom session and he/she will assist you. The data collected throughout this experiment does not include your name or any other information that would allow your identification. All of the data you provide during the experiment cannot be traced back to you.

Your earnings in today's session will be paid to you at the end of the experiment. Your earnings will depend on your own and other participants' decisions. You will play **60** rounds in total. For each round, your earnings will be in points. At the end of the experiment, your **accumulated** points will be converted to euros at a rate of 1 euro per 200 points. You will receive your earnings at the end of the experiment at the bank account you provided.

In the next page you will receive the relevant instructions. Thank you for your participation.

Rules of the sender-receiver game

Please read the following instructions carefully.

Matching & roles

In each round, all participants are matched in pairs. One participant within a pair has the role of player A and the other participant has the role of player B. The matching scheme is chosen to guarantee the following:

- In each round you will be randomly matched to another participant.
- You will never learn with whom you are matched with.
- You will never be paired to the same participant in subsequent rounds.
- You will always have the same role in all rounds.

Sequence of actions

1. In each round of the experiment, the computer will randomly roll a die with numbers between 1 and 5. All numbers are equally likely. This outcome of the die is called the *state*. Player A will observe the state, whereas player B will not.
2. Player A moves first and has to choose between the following 5 options.
 - Send the message "The state is 1"
 - Send the message "The state is 2"
 - Send the message "The state is 3"
 - Send the message "The state is 4"
 - Send the message "The state is 5"

If player A decides to send a message, it does not have to match the state. This is the only decision of player A.

-
3. Player B will observe the message and choose an action between 1 and 5. The decision of player B ends the round.

Earnings

In each round you can earn or lose points. The earnings of both players **depend only** on the state and the action of player B. The earnings **do not depend** on the message sent by player A. The earnings of both players for all possible combinations of state and action will be provided to you in a table. The table will be shown to both of you in the decision screen.

Understanding questions

Each cell of the table contains two numbers which correspond to the earnings of the two players.

- For player A, the earnings are the number on the left (shown in blue).
- For player B, the earnings are the number on the right (shown in red).

Remember that earnings depend **only** on the combination of state and action and **not** on the message.

Below there is an example of such a table to make you familiar with the format. All the scenarios described in the questions are purely hypothetical. Answering all questions correctly will make sure you fully understand the rules of the game and how points are earned.

	Action is 1	Action is 2
State is 1	10, 20	20, 10
State is 2	30, 30	40, 40

1. The state is 1. Player A send the message “The state is 1”. Player B chose action 2. **What are the earnings of each player?**
 - Player A gets 10 and player B gets 20
 - Player A gets 20 and player B gets 10
 - Player A gets 30 and player B gets 30
 - Player A gets 40 and player B gets 40
2. The state is 1. Player A send the message “The state is 2”. Player B chose action 2. **What are the earnings of each player?**

-
- Player A gets 10 and player B gets 20
 - Player A gets 20 and player B gets 10
 - Player A gets 30 and player B gets 30
 - Player A gets 40 and player B gets 40

3. The state is 2. Player A send the message “The state is 2”. Player B chose action 2. **What are the earnings of each player?**

- Player A gets 10 and player B gets 20
- Player A gets 20 and player B gets 10
- Player A gets 30 and player B gets 30
- Player A gets 40 and player B gets 40

4. The state is 2. Player A send the message “The state is 1”. Player B chose action 2. **What are the earnings of each player?**

- Player A gets 10 and player B gets 20
- Player A gets 20 and player B gets 10
- Player A gets 30 and player B gets 30
- Player A gets 40 and player B gets 40

5. When player A chooses the message to send to player B, both players know the state. **Is this statement True of False?**

- True
- False

6. Player A can send the message “The state is 2” when the state is 1. **Is this statement True of False?**

- True
- False

7. Player A sent the message “I don’t want to send a message” when the state is 1. Player B chose action 2. **What are the earnings of each player?**

Click the “Check” button below to check your answers. You can only proceed to the next page if all answers are correct.

Decision screen

Round X of 60

Below you see the table containing the earnings for both players for every combination of state and action.

- For player A, the earnings are the number on the left (shown in blue).
- For player B, the earnings are the number on the right (shown in red).

	Action is 1	Action is 2	Action is 3	Action is 4	Action is 5
State is 1	55, 108	88, 88	108, 55	88, 14	55, -31
State is 2	14, 88	55, 108	88, 88	108, 55	88, 14
State is 3	-31, 55	14, 88	55, 108	88, 88	108, 55
State is 4	-82, 14	-31, 55	14, 88	55, 108	88, 88
State is 5	-137, -31	-82, 14	-31, 55	14, 88	55, 108

[SENDER] You are **player A**. The randomly drawn state is DIE PHOTO (2). Please choose a message to send to player B by clicking the corresponding button below.

[RECEIVER AFTER ACTIVE SENDER] You are **player B**. Player A sent you the message “The state is 5”. Please choose a message to send to player B by clicking the corresponding button below.

[RECEIVER AFTER INACTIVE SENDER] You are **player B**. Player A was inactive in this round due to technical/connectivity issues. Hence, click Next to proceed.

Feedback screen

Results from round X of 60

Below you see the table containing the earnings for both players for every combination of state and action.

- For player A, the earnings are the number on the left (shown in blue).
- For player B, the earnings are the number on the right (shown in red).

	Action is 1	Action is 2	Action is 3	Action is 4	Action is 5
State is 1	55, 108	88, 88	108, 55	88, 14	55, -31
State is 2	14, 88	55, 108	88, 88	108, 55	88, 14
State is 3	-31, 55	14, 88	55, 108	88, 88	108, 55
State is 4	-82, 14	-31, 55	14, 88	55, 108	88, 88
State is 5	-137, -31	-82, 14	-31, 55	14, 88	55, 108

[PLAYER ACTIVE, PARTNER ACTIVE] The state was 2. Player A send the message “The state is 5”. Player B chose action 3.

[SENDER ONLY] You were **player A**. Therefore, in this round you earned 88 points.

[RECEIVER ONLY] You were **player B**. Therefore, in this round you earned 88 points.

[PLAYER ACTIVE, PARTNER INACTIVE] Your partner was inactive in this round so you automatically earned 100 points.

[PLAYER INACTIVE, PARTNER (IN)ACTIVE] You were inactive in this round and automatically earned 0 points.

Survey

Lottery Task

In the following task, **5 different lotteries** will be presented on your screen. In each of these lotteries, **both rewards A and B are equally likely**, i.e. have a probability of exactly 50%. The rewards are denoted in points.

You are asked to **choose exactly one** of the lotteries, which subsequently will be implemented. A random generator will determined whether you win reward A or reward B, respectively. At the end of the experiment, your reward will be added to your earnings.

	Reward A	Reward B	
No.	50% Probability	50% Probability	Your Choice
1.	140	140	
2.	120	180	
3.	100	220	
4.	80	260	
5.	60	300	
6.	10	350	

CRT elicitation

Please answer the following questions. Each correct answer is worth 50 points.

1. The ages of Mark and Adam add up to 28 years in total. Mark is 20 years older than Adam. How many years old is Adam?
2. If it takes 10 seconds for 10 printers to print out 10 pages of paper, how many seconds will it take for 50 printers to print out 50 pages of paper?

-
3. On a loaf of bread, there is a patch of mould. Every day the patch doubles in size. If it takes 12 days for the patch to cover the entire loaf of bread, how many days would it take for the patch to cover half the loaf of bread?

Trust attitudes

Please answer the following questions.

- When I communicate with strangers, I tell them the truth.
(Strongly disagree, Disagree, Neither agree or disagree, Agree, Strongly agree)
- When I communicate with strangers, they tell me the truth.
(Strongly disagree, Disagree, Neither agree or disagree, Agree, Strongly agree)

Demographics

Please answer the following questions.

- Please indicate your age.
- Please indicate your field of study.
(Economics, Social Sciences, Natural Sciences, Humanities, Applied Sciences, Other)
- Please indicate your gender.
(Male, Female, Prefer not to answer)

Payment information and debriefing (example)

Thank you!

The experiment is completed. Thank you for your participation.

From the main game, you earned in total 94 points. For the other tasks you additionally earned 154 points. The exchange rate is €1 for 200 points, so you earned €0.77.

You will receive your payment to your bank account using the IBAN you provided in the beginning of the experiment. You can now leave the Zoom session and close your browser.



Chapter 3

Anchoring and markets

This chapter is based on Ioannidis et al. (2020).

3.1 Introduction

A wealth of evidence has accumulated questioning some of the foundations of expected utility theory, and behavioural theorists have shown how these challenges can be accommodated (Wakker, 2010). At the core of standard and behavioural economic modelling remains the assumption that people are endowed with well articulated and stable preferences. This fundamental assumption, however, has also been challenged Ariely et al. (2003), who have shown that preferences are initially malleable by normatively irrelevant anchors. People subsequently choose consistently with these initial preferences, and thereby end up with preferences that are characterised by what Ariely et al. (2003) call "coherent arbitrariness". For a series of products that range from familiar (like an average bottle of wine) to unfamiliar (like listening to an unpleasant sound), they find substantial anchoring effects.

Economists often assign less weight to behavioural anomalies when they are obtained in non-repeated individual decision making tasks. The line of reasoning is that anomalies may be eroded when people have relevant experience, for instance as a result of trading in markets. To counter such scepticism, Ariely et al. (2003) included a treatment where subjects, after being exposed to an anchor, submitted a bid to avoid listening to an annoying sound. In the uniform-price sealed-bid auction, the three lowest bidders had to listen to the sound and each of them received a payment equal to the fourth lowest bid. Like in the individual decision making treatment, sizeable (and lasting) anchoring effects were observed in this treatment.

This chapter aims to make two contributions. A first contribution is that we investigate the effects of uninformative anchoring on valuations for a familiar good in a large sample. This is important because previous articles have provided mixed evidence, from sizeable anchoring effects (Ariely et al., 2003) to no anchoring effects (Fudenberg et al., 2012). Our chapter stands out because of the combination of two features. First, we have a large sample of 316 subjects who are all exposed to the same anchoring protocol, while previous studies have often been based on rather small samples. Second, we use a transparently random anchor that subjects know to be uninformative because they generate it themselves with a ten-sided die.

A second contribution of our chapter is that we investigate how elicited preferences are affected in a richer market setting than the one of Ariely et al. (2003), where subjects could not learn from others' bids during the auction. We employ a standard double auction where traders are continuously updated about other traders' bids and asks. We believe that a double auction provides a much better chance for market forces to erode initial traces of anchoring.

Our experiment consists of three phases. In the first phase, we apply a typical anchoring protocol: we ask whether subjects are willing to sell a bottle of wine for an individually drawn, random price. Then we elicit their valuation (Willingness-To-

Accept) for the bottle of wine with the Becker-DeGroot-Marschak (BDM) procedure (Becker et al., 1964). In the second phase, we randomly assign subjects to either a small double auction market ($n=2$) or a large double auction market ($n=8$). Subjects participate in two trading periods, once as a buyer and once as a seller. In the third phase, we elicit each subject's valuation once more.

The first phase of the experiment allows us to test whether a random anchor influences elicited valuations. We hypothesise that subjects' valuations correlate positively with their anchors. We further conjecture that market experience will affect subjects' elicited preferences. Subjects who are not completely sure about their preference may move into the direction of the preferences exhibited by other traders. This way, anchoring effects may diminish or even disappear. Thus, we hypothesise that the valuations elicited in the third phase will exhibit smaller (if any) anchoring effects. We also hypothesise that the large market will have a stronger effect on subjects' preferences than the small market, and that anchoring effects are eroded more efficiently in the former.

Contrary to our first hypothesis, we observe no effect of the random anchor on subjects' valuations. We believe our null result contributes to the literature on the robustness of anchoring effects. In the discussion section, we position our chapter in the literature and elaborate on what we can learn from our null result. There, we discuss the results of a concise meta-analysis of experimental articles that cite Ariely et al. (2003) and investigate the effects of anchoring on preferences.

We do find support for the idea that market participation affects how people value the bottle of wine. The variance in subjects' elicited valuations after the market shrinks within trading groups. As expected, the effect of other traders' behaviour on a subject's preference is stronger in the large market. These results underline the potential power that markets may play in eroding individual biases and noise. However, in this study the double auction is not needed to avoid anchoring effects on valuations.

The remainder of the chapter is organised as follows. Section 3.2 describes our experimental design and the hypotheses to be tested. Section 3.3 presents the results of the experiment. Section 3.4 provides a discussion of how our results fit in the literature.

3.2 Experimental design and implementation

We preregistered our study on the American Economic Association's registry for randomised controlled trials (Ioannidis et al., 2018).¹ The experiment was run at the

¹If we had found that elicited valuations are affected by anchoring and that markets diminish the role of anchoring, a confounding explanation would be that the effect of anchoring generally fades

CREED communication Lab of the University of Amsterdam. The communication lab has 16 soundproof, closed cubicles. The experiment was programmed in oTree (Chen et al., 2016). Subjects read the computerised instructions at their own pace (subsection 4.5.2). No communication was allowed during the experiment. Subjects were informed that they could earn money as well as a bottle of wine. It was explained that the experiment consisted of three phases during which they would make five decisions. Subjects knew that one of those five decisions would randomly be selected for payment at the end of the experiment. In phase I, subjects made two decisions, in phase II they made two decisions and in phase III they made one decision. They only received the instructions for the next phase after a previous phase was finished.

There were two treatments which were varied between subjects. The Small market consisted of two subjects and the Large market of eight subjects. In each session, we simultaneously ran the two treatments. Subjects were randomly assigned to either one of them.

Phase I was identical for both treatments. At the start of phase I, the experimenter entered each subject's cubicle with a ten sided die (numbered from zero to nine). Subjects determined their own random anchor by rolling the die twice. The first outcome was the integer part and the second was the decimal part of the anchor price. For example, if a subject rolled six and four, the price was 6.4€. Hence, subjects knew that the anchor price was an uninformative draw in the range from 0€ up to 9.9€. This procedure took place in the presence of the experimenter to guarantee that the subjects entered the correct numbers.² We used this procedure of subjects generating the anchor themselves to make it fully transparent to our subjects that the anchor price was truly random.

The first decision of phase I was the anchoring question. The subjects were endowed with a bottle of wine, a picture of which was shown to them on their screen. Consequently, they were asked whether they were willing to sell the bottle to the experimenter for a price that corresponded to the anchor price that they had just drawn. For the second decision of phase I, each subject was asked to submit the minimum price for which they were willing to sell the same bottle of wine. This Willingness-To-Accept decision (WTA) was incentivised via the BDM procedure. The application of the BDM procedure aimed at minimising the chance that subjects form any kind of inference from the elicitation process itself. The instructions included a description of the BDM mechanism and emphasised that it is optimal to provide the true valuation of the bottle. The explanation did not include a numerical example as we did not want any number to operate as an additional anchor. For the same rea-

out over time. Our preregistration mentions a control treatment to isolate the part of the reduction of the anchoring effect due to market forces and the part due to time fading. Given that we do not find an effect of anchoring, we did not run this treatment.

² Four out of 316 subjects did not wait until the experimenter arrived and entered numbers of their own.

son, the upper bound of the distribution from which the BDM price was drawn was not revealed. The subjects knew that a number would be randomly drawn between 0 and two times the (unknown) price of the bottle of wine in the store. To avoid outliers, we bounded the WTA from above. Subjects were given an error message if they entered a WTA above two times the price of the wine and were asked to re-submit their decision.³ The message did not inform them of the actual upper bound, but simply stated that their price was higher than what the experimenters believe is a reasonable price for the wine.⁴

In phase II, the market treatment was implemented. In the Large market, eight subjects participated in a double auction with four buyers and four sellers. In the Small market, two subjects participated in a market with one buyer and one seller. In a typical session of 16 subjects, half were randomly assigned to the Large market (one trading group) and half to the Small market (four trading groups). The market lasted for two periods. The trading group remained the same across the two periods, but buyer and seller roles were swapped. This way all subjects were exposed to both sides of the trade before they continued to phase III.

Except for the number of traders, the market treatments were identical. Each seller was endowed with a bottle of wine and each buyer was endowed with an amount equal to the price of the bottle of wine that we paid in the store. Traders were unaware of the size of this amount. At the end of the experiment, the amount was revealed only if the market decision was chosen for payment and only to buyers.

Buyers could submit bids to buy the bottle of wine. They could increase their bid multiple times, but not decrease (or withdraw) their current highest bid. Sellers could submit asks to sell the bottle of wine. They could decrease their ask multiple times, but not increase (or withdraw) their current lowest ask. All bids and asks were automatically recorded in the Order Book, which was visible to everyone and updated in real time. A trade occurred automatically whenever any of the following two rules was satisfied. (i) When a buyer submitted a bid that was higher than or equal to the lowest ask of the sellers in the Order Book, this buyer bought from the seller with the lowest ask and the corresponding ask was the transaction price. (ii) When a seller submitted an ask that was lower than or equal to the highest bid of the buyers in the Order Book, this seller sold to the buyer with the highest bid and the corresponding bid was the transaction price. All realised trades and their corresponding prices were automatically recorded in the publicly visible Trade Book. Subjects who had already traded still saw live updated Order and Trade Books. Before the trading

³ This message was shown to only five out of 316 subjects.

⁴Bohm et al. (1997) showed that selling prices elicited via a BDM mechanism are sensitive to the upper bound of the BDM distribution. They use three treatments varying the bound, namely standard (market price), high (unrealistic price) and unspecified (upper bound as "not to exceed what we believe any real buyer would be willing to pay"). They observe no difference between the standard and unspecified, whereas bidding is higher in the high treatment.

period opened, the subjects had to correctly answer 6 multiple choice questions to make sure they understood the rules of the market.

In phase III, we again elicited subjects' WTA for the bottle of wine with the BDM mechanism. After that, subjects were asked to complete a standard demographics survey asking for their age, gender and field of study. The experiment ended at this point and the final screen shown to the subjects informed them about which of the five decisions was chosen for payment as well as their payoff. If a WTA decision was implemented, they were informed of the random BDM draw and whether this random price meant that they sold the bottle of wine or kept it.

This design allows us to test the following hypotheses. To that purpose, we use the anchors to assign subjects to a High anchor group and a Low anchor group on the basis of either a median split or a quartile split. We use the data of phase I to test for anchoring.

Hypothesis 1. *The phase I WTA in the High anchor group is larger than the phase I WTA in the Low anchor group.*

We use the data of phases I and III to test whether the market affects subjects' elicited preferences and alleviates the anchoring effect.

Hypothesis 2. *The difference in WTA between the High and the Low anchor group is smaller in phase III than in phase I.*

Hypothesis 3. *The reduction in the difference in WTA between phase I and III is larger in the Large market treatment than the Small market treatment.*

In total 316 subjects participated in the experiment, 160 in the Large market treatment (20 trading groups) and 156 in the Small market treatment (78 trading groups). The experiment lasted approximately 75 minutes. Depending on their decisions during the experiment, subjects received on average 12.17€ including the participation fee of 8€ (excluding the bottle of wine). On top of their payment, 160 of the subjects physically received a bottle of wine. We used four different bottles of wine across sessions to avoid that prospective subjects could potentially learn the price from subjects that had participated already. Two of the bottles were priced at 6.00€ and two at 7.50€. Our subjects are on average 21 years old. Most of them (67%) are economics students, and they are evenly balanced across genders (females 53%, males 47%).

3.3 Results

3.3.1 Anchoring manipulation

In this subsection, we shed light on the question whether anchoring affects subjects' valuation of the bottle of wine. Figure 3.1 plots subjects' WTA in phase I as a function of their anchor. The figure suggests that subjects' WTA is fairly independent of their anchor.⁵

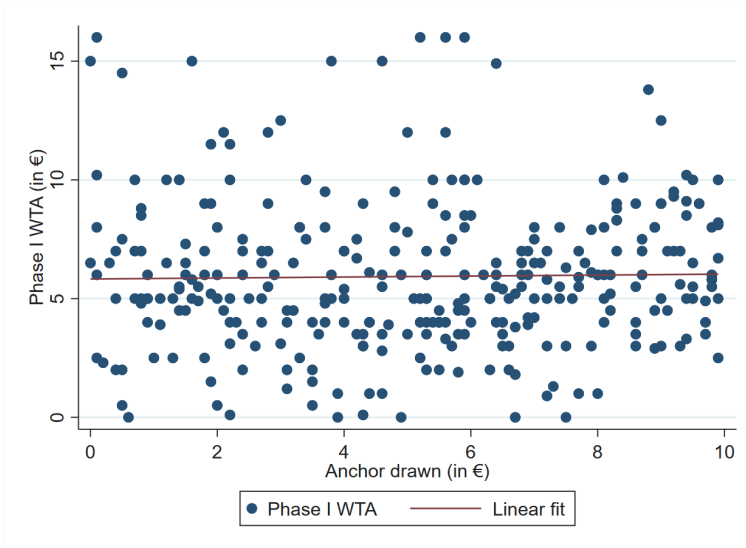


Figure 3.1: Scatter plot of phase I WTA on anchor with fitted regression line

Table 3.1 makes the results more precise. Hypothesis 1 states that the WTA in the group with high anchors will be larger than the WTA in the group with low anchors. First we do a median split of our data. Contrary to the hypothesis, the WTA for the Low anchor group does not significantly differ from the High anchor group. The evidence is in the expected direction, but the effect size is very small and far from economically significant. The magnitudes of our anchoring effects as measured by the ratio of the valuations in the top and bottom part of the distribution varies between 1.04 (for the ratio of the quartiles) to 1.07 (for the ratio of the quintiles).⁶ In comparison, for the series of products in Ariely et al. (2003) the ratio of top and bottom quintiles ranges from 2.16 to 3.03. The lack of support for an anchoring effect is further illustrated by a regression of the reported WTA on the anchor, while

⁵All the results presented in this subsection are robust to discarding the subjects that did not wait for the experimenter to record their anchor (see section 2) and the subjects that reported an unreasonable high initial WTA (see section 3).

⁶To have enough observations in each group, we preregistered to run the tests on the top versus the bottom half, and on the top versus the bottom quartile. The literature focuses on quintiles instead of quartiles. For comparison, we have included these statistics as well.

controlling for the price of the wine. The estimation reveals a very small and far from significant slope ($b = 0.019, SE = 0.061, CI = [-0.102, 0.140], t = 0.31, p = 0.755, N = 316$).

	WTA before		
	Median	Quartile	Quintile
High anchor group	6.07 (0.23)	6.35 (0.29)	6.74 (0.32)
Low anchor group	5.79 (0.27)	6.11 (0.37)	6.29 (0.44)
Ratio (High/Low)	1.048	1.039	1.072
z	1.314	1.137	1.249
p-value	0.189	0.257	0.213
Observations	316	163	123

Notes: z and p-values refer to Wilcoxon-Mann-Whitney rank-sum tests. Standard errors in parentheses.

Table 3.1: Mean WTA by anchor group

The previous literature has suggested some robustness checks. For instance, Fudenberg et al. (2012) include an analysis where they test for anchoring effects after leaving out inconsistent responses. We define a response as inconsistent if the WTA is higher than the anchor price that was accepted or lower than an anchor price that was rejected. In our sample, we have 51 (16.14%) inconsistent observations from subjects resulting in a reduced sample size of 265. Using rank-sum tests, we find no anchoring effect for either median split (ratio = 1.177, $z = 0.239, p = 0.811, N = 265$) or quartile split (ratio = 1.027, $z = 0.971, p = 0.429, N = 135$) or quintile split (ratio = 1.022, $z = 0.738, p = 0.460, N = 105$). A regression of valuation on anchor - again controlling for price - reveals an insignificant slope ($b = -0.015, SE = 0.063, CI = [-0.140, 0.110], t = -0.24, p = 0.813, N = 265$). Hence, focusing only on consistent answers does not affect our main result of no anchoring effects.

Another approach that has been used in the literature is to replace valuations above the BDM range by the maximum of the BDM range. One reason to do so is that all reports higher than the BDM range yield the same outcome. So very high reports need not reflect very high valuations, which may bias the analysis. Ariely et al. (2003) and Maniadis et al. (2014) truncate valuations in this way and find that it does not affect their results. The same approach is not directly applicable for our study as our subjects did not know the exact range of the BDM, and higher valuations than the maximum were not allowed. However, in the same spirit we can investigate whether our results are sensitive to replacing valuations above 10 by 10, the highest possible anchor. Rank-sum tests reveal no anchoring effects for either median split (ratio = 1.077, $z = 1.359, p = 0.174, N = 316$) or quartile split (ratio = 1.081, $z = 1.169, p = 0.242, N = 163$) or quintile split (ratio = 1.177, $z =$

1.278, $p = 0.201$, $N = 123$). A regression of valuation on anchor and price confirms the result ($b = 0.055$, $SE = 0.051$, $CI = [-0.046, 0.156]$, $t = 1.07$, $p = 0.285$, $N = 316$). Hence, the truncation of valuations also does not qualify our null result.

As a final robustness check, we test for anchoring across demographic characteristics of our subjects (field of study and gender) as well as across different types of wine. We regress valuation on anchor, controlling for the market price of the wine and find no anchoring effect both for economic students ($b = 0.049$, $SE = 0.075$, $CI = [-0.099, 0.198]$, $t = 0.66$, $p = 0.511$, $N = 211$) as well as non-economic students ($b = -0.037$, $SE = 0.106$, $CI = [-0.247, 0.173]$, $t = -0.35$, $p = 0.735$, $N = 105$). We repeat the same exercise and find no anchoring effect both for male students ($b = -0.023$, $SE = 0.089$, $CI = [-0.200, 0.154]$, $t = -0.26$, $p = 0.797$, $N = 148$) as well as female students ($b = 0.061$, $SE = 0.085$, $CI = [-0.108, 0.229]$, $t = 0.71$, $p = 0.477$, $N = 168$). Similarly, we find no anchoring effect for any of the four types of wine we used.⁷

Result 1. *There is no anchoring effect in our data.*

Two factors may play a role in our null-result for the effect of anchoring on valuations. The first is the familiarity of the product. Most subjects are likely to be familiar with a bottle of wine, and it may be that anchoring effects occur more easily for unfamiliar products for which people lack a clear initial preference. The other is that in our study the anchoring procedure is transparently uninformative. In the concluding discussion, we present the results of a small meta-analysis that sheds light on these factors.

In light of Result 1, any analysis on whether market experience reduces anchoring effects is meaningless. Still, it remains interesting to investigate whether the market affects people's preferences. Previous work showed that markets can affect people's preferences for unfamiliar goods for which people might not have a clear initial preference to start with, such as tasting an unpleasant liquid (Tufano, 2010) and lotteries (Isoni et al., 2016). It is not clear that markets can affect people's preferences for more familiar goods like a bottle of wine. For the remainder of the analysis, we group subjects on the basis of their phase I WTA instead of their anchor and reshape the remaining two hypotheses accordingly.

3.3.2 Market effect on valuations

To investigate how markets affect elicited preferences, we define new groups based on the valuations. For each trading group separately, we use a median split of the

⁷We regress valuation on anchor separately for each type of wine and the coefficient of anchor is never significant. The p-values range from 0.131 to 0.791.

phase I WTA to assign each subject to a Low or High WTA group.⁸

Hypothesis 2 deals with the question whether the information revealed during the market affects the valuation of subjects. Figure 3.2 illustrates the results. Focusing on the aggregate results of the Small and the Large market, it is clear that subjects' valuations move in the direction of the other WTA group.⁹

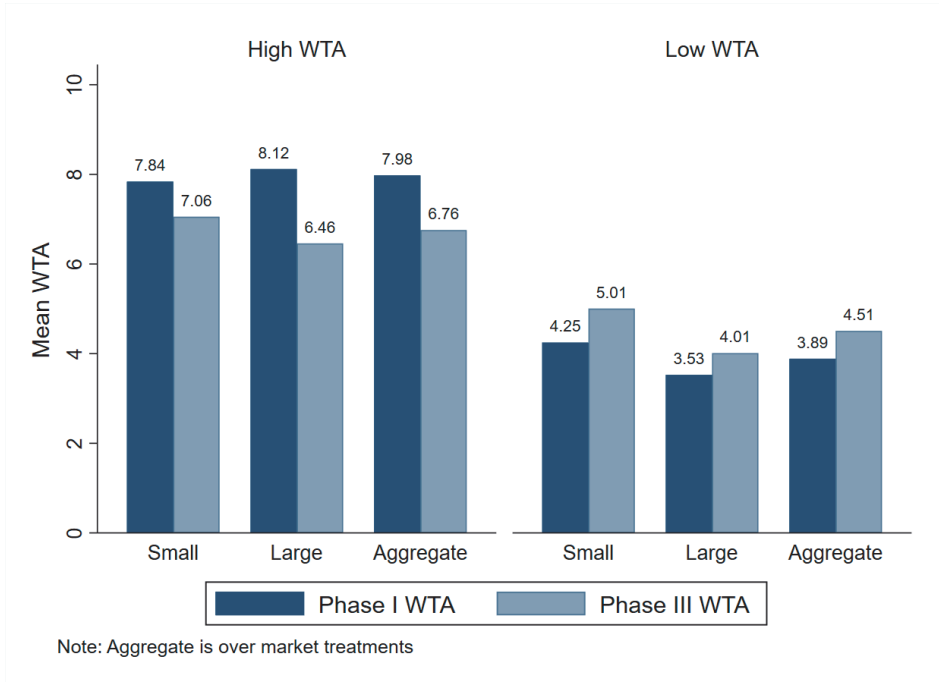


Figure 3.2: Average WTA before and after the market

To test whether the change is a statistical artefact due to regression to the mean, we compare for each subject the absolute difference between their WTA from phase I and the average WTA (from phase I) in their trading group with the same variable from phase III. The average absolute difference is 2.06€ in phase I and 1.64€ in phase III. A Wilcoxon signed-rank test reveals that traders' WTA vary less after the market than before ($z = -5.321, p < 0.001, N = 316$).

Result 2. *Subjects change their WTA in the direction of the average WTA in their own trading group. WTA's elicited in phase I vary more within their trading group than WTA's in phase III do.*

⁸If for example in the Small market, one subject in a trading group submits a valuation of 1€ and the other a valuation of 2€ we classify the latter in the High WTA group, even though his WTA is low in comparison to the overall sample of subjects.

⁹Subjects in the Low WTA group increased their valuation significantly by 0.62€ ($z = 3.720, p < 0.001, N = 158$) and in the high WTA group decreased it significantly by 1.23€ ($z = -5.511, p < 0.001, N = 158$).

We now turn to the question whether subjects change their preferences more in the Large market than in the Small market. Figure 3.2 also displays the results for each market separately. In agreement with Hypothesis 3, we observe that the average decrease in WTA for the High WTA group is larger in the Large market than in the Small market. In contrast to Hypothesis 3, subjects in the Low WTA group increase their WTA to a somewhat larger extent in the Small market than in the Large market.

We test the differential impact of the Large market on preferences in a regression that explains the phase III WTA by the phase I WTA and observed market information, with and without interaction term for the treatment. We define observed market information as the average of the last observed actions of the other market subjects. For market subjects that traded, the last observed action is the price they agreed on. For market subjects that did not trade, it is the last bid/ask they submitted. The results in Table 3.2 provide supportive evidence for the idea that subjects attach more weight to their own WTA in the Small market as compared to the Large market.¹⁰ In the first two columns, we present regressions for each market separately and in the last column we present a regression with both markets and an interaction term.

Table 3.2: Effect of observed market information on WTA change by treatment

	Small Phase III WTA	Large Phase III WTA	Both Phase III WTA
Phase I WTA	0.64*** (0.08)	0.47*** (0.08)	0.50*** (0.06)
Observed market information	0.33*** (0.08)	0.36* (0.14)	0.34*** (0.07)
Small*Phase I WTA			0.10* (0.05)
Wine price	0.32 (0.24)	0.08 (0.25)	0.25 (0.17)
Constant	1.80 (2.32)	-2.03 (1.46)	-0.52 (1.39)
Controls	Yes	Yes	Yes
R^2	0.520	0.421	0.470
Observations	156	160	316

Notes: Controls are gender, age and field of study. Standard errors in parentheses are adjusted for 98 trading group clusters.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

¹⁰In Table 3.2 we included the market price of the wine as a control variable. Using wine fixed effects instead produces qualitatively similar results.

Result 3. *Subjects change their valuation more in the direction of other trading group members in the Large market than in the Small market.*

3.4 Concluding discussion

In this chapter we find supportive evidence for the potential of markets to reduce the effects of anchoring on valuations.¹¹ However, in our study the market was not needed to correct a potential individual bias due to anchoring. We find no effect of anchoring on reported valuations. This result raises doubt on the robustness of the anchoring effect on people's preferences. Before we discuss differences between our design and other similar studies, we emphasise that our null result is not a consequence of an under-powered study. The observations of the first two sessions were used as a pilot to conduct a power analysis. We conducted a power analysis with an aim to obtain a significant result at the 5% level with 80% power in each of our market treatments separately. The power analysis resulted in an estimated sample size of 148 subjects per market treatment, so 296 subjects in total. To be on the safe side, we aimed for 320. Given that the anchoring hypothesis is based on the total sample (as no treatment has been introduced yet), we have a very high power of 99%.

The existing literature provides a mixed picture of whether anchors affect people's valuations. Studies differ in details in how they were run, and it is possible that anchoring effects on preferences occur in some circumstances but not in others. When trying to make sense of previous results on anchoring, a complicating factor is that many of these are based on rather small samples which makes it impossible to distinguish between true results, false positives and false negatives. However, even the large studies provide mixed results.

To make sense of previous results, we carried out a limited meta-analysis. In this analysis, we restricted our attention to the 1493 articles that cite Ariely et al. (2003). Among those, we selected the ones that reported an incentivised experiment investigating the effect of anchoring on elicited valuations. We left out results based on hypothetical payments, which includes a very large literature on the effects of anchoring on contingent valuations.¹² Table 3.3 lists the 19 selected studies on the

¹¹Subjects change their valuations in the direction of the others in their trading group. Our findings corroborate the results of Tufano (2010) and Isoni et al. (2016) who show that markets shape preferences for tasting an unpleasant liquid and preferences for lotteries, respectively. Our results show that markets not only change elicited preferences for unfamiliar goods, i.e. goods where people might not have a clear preference to start with, but also for familiar goods. Our results do not shed light on the question whether the shaping of preferences is a rational process or not. Behavioural conformism may drive the changes in elicited preferences. However, it may also be that preferences for the wine are partly determined by an estimate of the price of the wine in the store, and that people use others' trading decisions to rationally form a better estimate of the retail price.

¹²Some studies combine incentivised treatments and hypothetical treatments. In those cases, our selection only includes results from the incentivised treatments.

anchoring of preferences, together with their main features and the reported effect. Each study result is summarised in two ways: (i) as a ratio of valuations of high over low anchor group, and (ii) as Hedge's g , defined as the difference in valuations between high and low anchor group divided by the pooled standard deviation.

The first feature describes the type of good for which a valuation was elicited. *Familiar* goods are ordinary goods that most people now and then consume, like wine, chocolate and books. *Unfamiliar* goods are goods for which people lack daily life experience, like consuming badly tasting liquids and listening to unpleasant sounds. People's preferences may be more affected by anchors when they report their value for an unfamiliar product for which they do not have well-articulated preferences. The second feature describes the extent to which the anchor may have been perceived as being informative about the price of the good. Some studies use *informative* anchors. One such example is provided by Jung et al. (2016), who investigate the effect of a default anchor on people's donation in a Pay-What-You-Want pricing scheme. Naturally, a default may be perceived as a recommended donation. There are also studies that intended to provide an uninformative anchor, but which may unintentionally have been interpreted as informative by subjects. We categorise these anchors as *questionable*. One such approach is to let a subject's anchor be determined by the last two digits of their social security number (SSN) (e.g., Ariely et al. (2003), Bergman et al. (2010)). About one-third of the subjects of Chapman and Johnson (1999) mention that they thought that the SSN anchor was informative. Likewise, Yoon and Fong (2019) and Yoon et al. (2019) use randomly generated uninformative prices, but leave subjects in the dark about the nature of the random number. Their instructions do not exclude the possibility that the random number is somehow correlated to the true price.¹³ Studies that use a randomly generated anchor, and clearly communicate the whole procedure to the subjects, are categorised as using an *uninformative* anchor.

The third feature in which studies differ is whether subjects' willingness-to-accept (WTA) or willingness-to-pay (WTP) is elicited. We include this variable because initially there was some support for the idea that anchoring effects are more easily observed for WTP than for WTA (Simonson and Drolet, 2004).

Before presenting the results of this concise meta-analysis, we motivate some methodological choices. First, all results need to be weighted appropriately with respect to their precision. Studies are typically weighted by the inverse of the variance of the estimated effect size. We use this approach here and weigh each study according to the variance of Hedge's g .¹⁴ Second, to test whether the effect size varies

¹³They instructed their subjects about the anchoring in the following way: "First, we will ask whether you would like to buy the item at a particular price. That price will be determined randomly by having you convert the numbers on the card you received into a whole-dollar price."

¹⁴Hedge's g is computed as $g = \frac{m_1 - m_2}{s}$ where m_1, m_2 are the means of the two groups and s is

Table 3.3: Anchoring on valuation studies (ordered by publication year)

(1) Study	(2) Good	(3) Anchor	(4) WTP/WTA	(5) #Results	(6) Sample	(7) Ratio(H/L)	(8) Hedge's g
Ariely et al. (2003)	Familiar [0.2%]	Questionable [2.4%]	WTP [0.2%]	1	55	1.71	0.841
—" —"	Unfamiliar [15.0%]	Informative [0.9%]	WTA [8.9%]	4	61	2.16	1.311
—" —"	Unfamiliar [5.7%]	Questionable [3.7%]	WTA [3.4%]	1	90	1.62	1.137
Simonson and Drolet (2004)	Familiar [1.2%]	Questionable [13.2%]	WTA [12.0%]	2	139	1.23	0.331
Ariely et al. (2006)	Unfamiliar [11.7%]	Informative [0.7%]	WTP [0.7%]	2	164	1.36	2.905
—" —"	Unfamiliar [5.2%]	Questionable [3.4%]	WTP [0.3%]	1	81	3.95	1.016
Bergman et al. (2010)	Familiar [0.5%]	Questionable [5.2%]	WTP [0.5%]	1	116	1.43	0.794
Tufano (2010)	Unfamiliar [9.8%]	Informative [0.6%]	WTA [5.9%]	1	134	1.05	0.061
Sugden et al. (2013)	Familiar [3.9%]	Informative [4.1%]	WTA [39.1%]	9	100	1.11	0.165
—" —"	Familiar [4.4%]	Informative [4.5%]	WTP [4.6%]	9	111	1.09	0.071
—" —"	Unfamiliar [13.1%]	Informative [0.8%]	WTA [7.8%]	9	20	1.06	0.166
—" —"	Unfamiliar [14.6%]	Informative [0.9%]	WTP [0.9%]	9	22	1.09	0.086
Koçaş and Demir (2014)	Familiar [0.1%]	Informative [0.2%]	WTP [0.2%]	1	46	5.00	1.766
Maniadis et al. (2014)	Unfamiliar [8.5%]	Informative [0.5%]	WTA [5.0%]	1	116	1.29	0.221
Alevy et al. (2015)	Familiar [0.8%]	Questionable [8.9%]	WTA [8.2%]	1	187	1.06	0.120
Ma et al. (2015)	Unfamiliar [1.5%]	Informative [0.1%]	WTA [0.9%]	1	48	1.34	3.228
Shah et al. (2015)	Familiar [0.4%]	Questionable [4.1%]	WTP [0.4%]	1	95	1.53	0.934
Isoni et al. (2016)	Unfamiliar [14.9%]	Informative [0.9%]	WTA [8.9%]	1	204	1.19	0.186
Jung et al. (2016)	Familiar [82.2%]	Informative [84.9%]	WTP [85.7%]	14	1383	1.13	0.140
Li et al. (2017)	Familiar [0.4%]	Questionable [4.0%]	WTP [0.4%]	1	88	1.21	0.624
Ifcher and Zarghamee (2020)	Familiar [0.8%]	Informative [0.8%]	WTP [0.9%]	1	190	1.31	0.315
Yoon et al. (2019)	Familiar [1.5%]	Questionable [16.0%]	WTP [1.5%]	3	117	1.30	0.615
Yoon and Fong (2019)	Familiar [3.6%]	Questionable [39.1%]	WTP [3.7%]	4	215	1.33	0.581

(a) Informative and questionable anchors (corresponds to Figure 3.3a)

(1) Study	(2) Good	(3) Anchor	(4) WTP/WTA	(5) #Results	(6) Sample	(7) Ratio(H/L)	(8) Hedge's g
Fudenberg et al. (2012)	Familiar [28.6%]	Uninformative [24.0%]	WTA [27.2%]	2	79	1.00	0.001
—" —"	Familiar [14.1%]	Uninformative [11.8%]	WTP [100%]	1	78	0.99	0.009
—" —"	Unfamiliar [100%]	Uninformative [16.4%]	WTA [18.6%]	1	108	0.96	0.086
Current study (2019)	Familiar [57.2%]	Uninformative [47.9%]	WTA [54.3%]	1	316	1.05	0.089

(b) Uninformative anchors (corresponds to Figure 3.3b)

Notes to columns (column number in parentheses):

(2-4): The percentages indicate the weight each study receives (see Table 14 for formulas). The percentages are normalised to add up to 100% within each category. For aggregated results, we present the sum of weights.

(5): #Results displays the number of treatments in an article that share the same features for the variables listed in columns (2), (3) and (4) and that differ in features that are not included in the table.

(6): For aggregated results, we present the average sample size.

(7): For aggregated results, we present the sample-weighted average ratio.

(8): For aggregated results, we present the sample-weighted average Hedge's g .

across different subgroups, we use the Q statistic. The Q statistic is a measure of the weighted variance of the effect sizes and is compared with the variance that would have been observed if all effect sizes were sampled from a population with the same mean.¹⁵ Third, we use a random-effects model. Given that we are accumulating data from different studies that were carried out in different ways, we believe that the random-effects model is more appropriate than the fixed-effects model.

Figure 3.3 provides an overview of how these three dimensions affect the effect of anchoring on elicited valuations. We present the results in two separate forest-plots, the lower one for studies that use uninformative anchors and the upper one for studies that use informative or questionable anchors. Overall, whether an anchor is uninformative or not has a strong effect on whether anchoring affects elicited valuations or not.¹⁶ With uninformative anchors, we find a precise null-result for the effect of anchoring on elicited valuations. In contrast, with informative or questionable anchors there is a sizeable and significant effect of anchoring. The difference in anchoring effects between studies with uninformative anchors and the other studies is significant ($Q = 27.67, p < 0.001, N = 83$).

Within the class of studies that use informative or questionable anchors, we find the following results of the mediating variables on the empirical relevance of anchoring; (i) anchors have a significantly stronger effect for unfamiliar goods for which people do not have a clear initial preference to start with than for familiar goods, whereas (ii) whether WTA or WTP is used to elicit subjects' preferences does not matter; (iii) somewhat surprisingly, we find a significantly stronger effect of questionable anchors than familiar anchors, which supports the view that many of the questionable anchors were actually interpreted as informative by subjects.

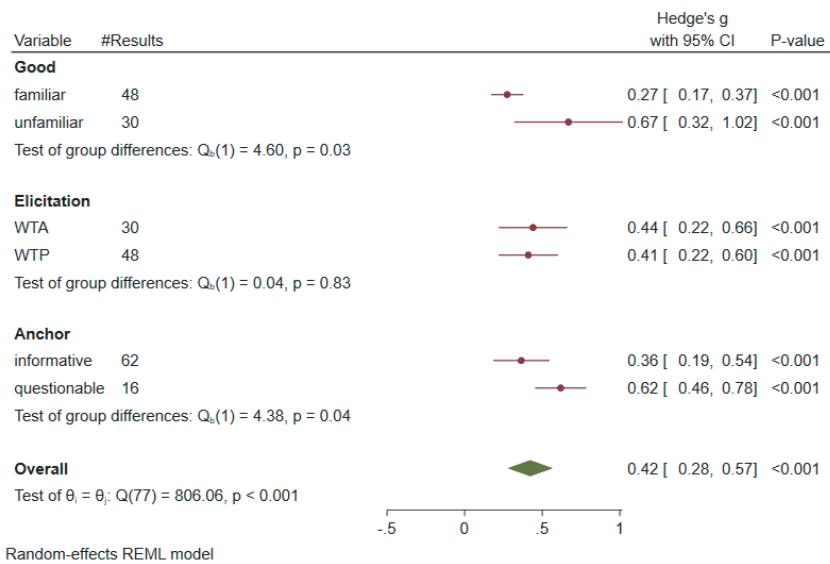
Interestingly, in studies that use transparently uninformative anchors, anchoring never has an effect on elicited valuations. So far, whether a familiar or unfamiliar good is used does not matter when the anchor is clearly uninformative. Although these studies are based on relatively many data, there are only a couple of them, and clearly more studies in this category are welcome.

In the class of studies that use familiar goods, a final interesting comparison is between studies that use clearly uninformative anchors and those that do not. Anchoring effects on elicited valuations are only observed in the latter category, and the

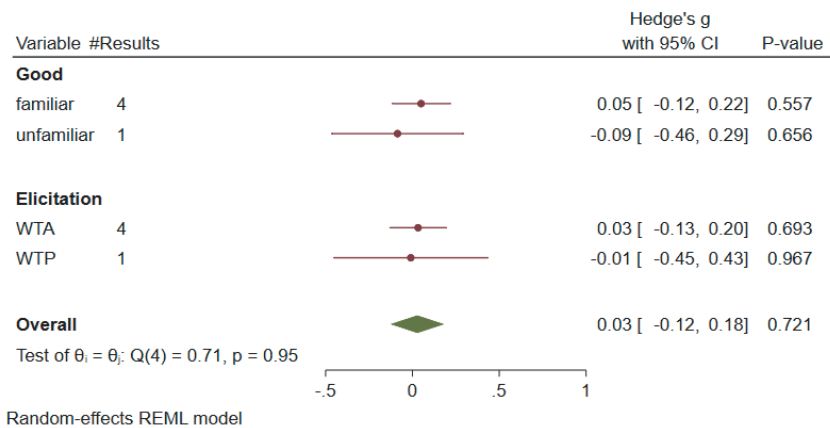
their pooled standard deviation. The variance of the estimator is given by $\text{Var}(g) = \frac{n_1+n_2}{n_1n_2} + \frac{g^2}{2(n_1+n_2)}$, where n_1, n_2 are the sample sizes of the two groups. The weight of each study is given by $w = \frac{1}{\text{Var}(g)}$.

¹⁵The Q statistic is computed as $Q = \sum_{i=1}^k (w_i(ES_i - \bar{ES})^2)$, where k is the number of subgroups compared, w_i is the weight of each study, ES_i is the effect size of each study and \bar{ES} is the mean effect size across all studies. Under the null hypothesis that all effect sizes are equal, the Q statistic follows a X^2 distribution with $(k - 1)$ degrees of freedom. For more details, we refer to Card (2015, Chapter 8).

¹⁶As a robustness check, we repeated the analysis without Jung et al. (2016) which has very large weight (due to the total number of participants exceeding 19,000); all conclusions remain qualitatively the same.



(a) Informative and questionable anchors



(b) Uninformative anchors

Figure 3.3: Forest-plots of results by anchor type

Notes: For each subgroup, the plot includes a dot centred at the mean of the effect size of the corresponding subgroup with lines extended to indicate the confidence intervals. The overall effects are represented by diamonds centred on their estimated values with the width corresponding to the confidence interval length. The height of the diamond is not relevant.

difference in the effect is significant ($Q = 23.21, p < 0.001, N = 52$). So, preferences for familiar goods can be anchored, but this requires the use of an anchor that is perceived as informative.

Overall, this meta-analysis yields the following picture: (i) if anchors are informative or perceived to be informative, then (unsurprisingly) anchoring has an effect, and mediating variables play mostly a sensible role, that is, no difference in the effect

of anchoring on WTA and in the effect on WTP, while anchors have a stronger effect for valuations of unfamiliar products than familiar products; (ii) in the few studies in which anchors are uninformative there is a quite precise null-effect, and so far none of the mediating variables plays any role.

In many cases with real world relevance, evaluations may be made in the presence of seemingly (though not truly) informative anchors. In our view, our study sheds light on what makes anchoring of valuations actually have an impact. Originally, Tversky and Kahneman (1974) demonstrated the effect of anchors on people's *judgements* with transparently uninformative numbers. On page 1128, they write "subjects were asked to estimate various quantities, stated in percentages (for example, the percentage of African countries in the United Nations). For each quantity, a number between 0 and 100 was determined by spinning a wheel of fortune in the subjects' presence. The subjects were instructed to indicate first whether that number was higher or lower than the value of the quantity, and then to estimate the value of the quantity by moving upward or downward from the given number. Different groups were given different numbers for each quantity, and these arbitrary numbers had a marked effect on estimates." Our study suggests that the source of the anchoring of *valuations* may not be that a random uninformative number is imprinted in a subject's mind. Instead, the problem seems to be that people can be tricked into believing that an uninformative piece of information is actually a relevant piece of information. In this sense, the anchoring of valuations is more about people being too gullible when they process information. If the anchoring of preferences only reliably appears when subjects perceive the anchor as informative, then it may be less appropriate to think of the anchoring of preferences as an anchoring bias. Instead, it seems to be driven by a perception bias.

3.5 Appendix to Chapter 3

3.5.1 Experimental instructions

Welcome to the session

Welcome!

Welcome to this experiment. Please read the following instructions carefully. We ask that you do not communicate with other participants during the experiment. The use of mobile phones is not allowed during this experiment. If you have any questions, or need assistance of any kind, at any time, an experimenter will assist you privately. The data collected through this experiment does not include your name or any other information that would allow your identification. All of the data you provide during

the experiment cannot be traced back to you.

Your earnings in today's session will consist of money and/or a bottle of wine. You will start the experiment with a capital of €8.0. Besides this starting capital, your earnings will depend on your own decisions and may depend on other participants' decisions.

The bottle of wine is showed below. The bottle of wine is **different** in every session of this experiment.



In this experiment you will make 5 decisions in total.

At the end of the experiment, one of your 5 decisions will be selected at random. Only this decision will determine your earnings (possibly in combination with the decisions of other participants). Your earnings for this decision will be added to your starting capital of €8.0. You will privately receive your earnings at the end of the experiment.

For each decision you will receive some instructions. You will only receive the instructions of a subsequent decision if a previous decision is completely finished.

Thank you for your participation.

Anchor generation screen

For this part of the experiment, you are required to roll a 10-sided die twice and report the outcomes. The outcomes will be converted to a price that you will see in the next page. The price is constructed in the following way: the first roll is the integer (euros) part and the second roll is the decimal part of the price you will see. Please remain seated and the experimenter will come to your room with the die shortly.

Submit the outcome of first die roll and click Submit.

Submit the outcome of second die roll and click Submit.

Please confirm that your rolls where X and Y. If the outcomes were different, please click Resubmit.

The outcomes of the rolls are registered. The corresponding price is €X.Y. The experiment will continue shortly after the experimenter leaves the room. Please wait.

Decision 1

[Photo of wine bottle]

You are given the bottle of wine. You are asked to decide if you want to sell it back to the experimenter for a price equal to €X.Y.

If the decision is selected at the end of the experiment, then

- A. If you click **NO**, you keep the bottle of wine.
- A. If you click **YES**, you get the €X.Y in return for the bottle of wine.

Decision 2

[Photo of wine bottle]

You are given the bottle of wine. You are asked to report the lowest price (rounded to the nearest 10 cents) for which you are willing to sell the bottle of wine to the experimenter. The lowest price is the one that makes you indifferent between keeping the bottle and selling it.

If this decision is selected at the end of the experiment, a random price will be drawn between €0.00 and what the experimenter estimates to be the maximum price any buyer would be willing to pay. Then one of the following will happen:

- a. If the random price is smaller than the price you reported, then you keep the bottle of wine.
- a. If the random price is larger or equal to the price you reported, then the experimenter will buy the bottle of wine from you and you will receive the random price.

You will not receive the price you reported. Instead you may receive the random price drawn. The reported price had no impact on the random price as the price was drawn before the experiment started.

It is in your best interest to report the price that equals your true valuation for the bottle of wine.

Please confirm that your price was €[price reported by participant].

If you'd like to change the price, please click Resubmit.

Market Instructions

You will now participate in a market running for 2 periods. In this market, every participant is a trader. The item for trade is the same bottle of wine. The duration of each period of the market is 300 seconds.

Roles

In the first period of the market, every participant is assigned to the role of either a buyer or a seller.

There are 4 buyers and 4 sellers in the market. In the second period of the market, all participants will change role. [Large market treatment]

There is 1 buyer and 1 seller in the market. In the second period of the market, all participants will change role. [Small market treatment]

Buyers

Each buyer is endowed with an amount equal to the price of the bottle of wine that we paid in the store. The size of this amount will be revealed to you only if it affects your payoffs.

Sellers

Each seller is endowed with a bottle of wine.

Offers

All the Bids and Asks will automatically be recorded in the **Order Book**, which is visible to all participants in the market.

Trades

When a trade occurs, it gets automatically recorded in the **Trade Book**, which is visible to everyone.

The buyer and the seller that traded will receive a message informing them about the trade as well. They can no longer submit offers, but they continue to see the live-updated offer book and trade book. [Large market treatment]

Rules

The rules of the market are presented in the next page.

Market rules and control questions

Buyers

Buyers are asked to submit the price they are willing to pay for the bottle of wine. Offers submitted by the buyers are called **Bids**. Buyers can *increase* their bid multiple times, but not decrease (or withdraw) their current bid.

Sellers

Sellers are asked to submit the price they are willing to accept for the bottle of wine. Offers submitted by sellers are called **Asks**. Sellers can *decrease* their ask multiple times, but not increase (or withdraw) their current ask.

Trades

A trade occurs automatically whenever any of these two rules are satisfied.

- When a buyer submits a bid that is higher than or equal to the lowest ask of the sellers in the **Order Book**, this buyer buys from the seller with the lowest ask and the corresponding ask is the transaction price.
- When a seller submits an ask that is lower than or equal to the highest bid of the buyers in the **Order Book**, this seller sells to the buyer with the highest bid and the corresponding bid is the transaction price.

Earnings

If a trade occurs, the seller receives the transaction price, while the buyer receives the bottle of wine and the transaction price is subtracted from his/her endowment. If a seller does not trade, the seller keeps the bottle of wine. If a buyer does not trade, the buyer keeps his/her endowment.

Please answer the questions on the right to make sure you fully understand the rules of the market.

- You are a buyer. Your bid is the second highest bid. You increase your bid and become the highest bidder. Your current bid is lower than the lowest ask.
Does this action result in a trade for you? And if yes, at which price?
 - Yes, for a price equal to buyer's bid
 - Yes, for a price equal to the seller's ask
 - No

-
- You are a seller. You want to submit an ask higher than your current ask.

Is this ask allowed?

Yes

No

- You are a seller. Your ask is the lowest ask. Your current ask is higher than the highest bid. You decrease your ask below the highest bid.

Does this action result in a trade for you? And if yes, at which price?

Yes, for a price equal to buyer's bid

Yes, for a price equal to the seller's ask

No

- You are a buyer. Your bid is the highest bid. Your bid is lower than the lowest ask. A seller decreases his/her ask below your bid.

Does this action result in a trade for you? And if yes, at which price?

Yes, for a price equal to buyer's bid

Yes, for a price equal to the seller's ask

No

- You are a seller. Your ask is the second lowest ask. A buyer increases his/her bid above your ask.

Does this action result in a trade for you? And if yes, at which price?

Yes, for a price equal to buyer's bid

Yes, for a price equal to the seller's ask

No

- You are a buyer. You want to submit a bid higher than your current bid.

Is this bid allowed?

Yes

No

Click the "Check" button below to check your answers. You can only proceed to the next page if you have answered all questions correctly.

Decisions 3 and 4

[Photo of wine bottle]

Time left to the end of the market:

Screen of buyer who has a currently active bid before any trades

Order	Book	Trade	Book
Asks	Bids	Number	Price
8.2	6.5		
9.8	5.3		
11.2	2.6		
15.8	0.8		

You are a **buyer** in the market. Your current bid is €5.3.

Screen of buyer and seller who completed a trade

Order	Book	Trade	Book
Asks	Bids	Number	Price
9.8	2.6	1	8.2
11.2	0.8	2	6.5

You are a **seller** in the market. [Seller screen]

You are a **buyer** in the market. [Buyer screen]

You agreed to trade for a price of €6.5. Please wait until the auction is over.

Screen of buyer and seller who are active after trades took place

Order	Book	Trade	Book
Asks	Bids	Number	Price
9.8	2.6	1	8.2
11.2	0.8	2	6.5

You are a **buyer** in the market. Your current bid is €0.8. [Buyer screen]

You are a **seller** in the market. Your current ask is €11.2. [Seller screen]

Market Summary of periods 1/2

Summary screen for seller who did not trade

Period 1/2 of the market is now over.

In period 1/2, you did not agree on a trade.

The market will start again shortly. In the next period of the market, you will be a buyer. [Period 2 only]

Summary screen for buyer who did not trade

Period 1/2 of the market is now over.

In period 1/2, you did not agree on a trade.

The market will start again shortly. In the next period of the market, you will be a seller. [Period 2 only]

Summary screen for seller who traded

Period 1/2 of the market is now over.

In period 1/2, you were a seller and sold the bottle of wine for €6.5.

The market will start again shortly. In the next period of the market, you will be a buyer. [Period 2 only]

Summary screen for buyer who traded

Period 1/2 of the market is now over.

In period 1/2, you were a buyer and bought the bottle of wine for €6.5.

The market will start again shortly. In the next period of the market, you will be a seller. [Period 2 only]

Decision 5

[Photo of wine bottle]

You are given the bottle of wine. You are asked to report the lowest price for which you are willing to sell the bottle of wine to the experimenter.

If this decision is selected at the end of the experiment, a random price will be drawn between €0.00 and what the experimenter estimates to be the maximum price any buyer would be willing to pay. Then one of the following will happen:

- a. If the random price is smaller than the price you reported, then you keep the bottle of wine.
- a. If the random price is larger or equal to the price you reported, then the experimenter will buy the bottle of wine from you and you will receive the random price.

You will not receive the price you reported. Instead you may receive the random price drawn. The reported price had no impact on the random price as the price was drawn before the experiment started.

It is in your best interest to report the price that equals your true valuation for the bottle of wine.

Please confirm that your price was €[price reported by participant].
If you'd like to change the price, please click Resubmit.

Demographics

Please answer the following questions.

- Please indicate your age.
- Please indicate your field of study.
(Economics, Social Sciences, Natural Sciences, Humanities, Applied Sciences, Other)
- Please indicate your gender.
(Male, Female, Prefer not to answer)

Debrief screen example (Decision 1 selected for payment)

Thank you!

Thank you for your participation in the experiment. Your starting capital was €8.0. Additionally, Decision 1 was randomly chosen for payment. You were endowed with the bottle of wine.

You reported you wanted to sell the bottle of wine to the experimenter for €6.2. Hence, you receive €6.2 from the sale of the bottle of wine.

Your total payment is €14.2. [Participant who sold the bottle to experimenter]

You reported you did not want to sell the bottle of wine to the experimenter for €2.1. Hence, you receive no money, but you get to keep the bottle of wine.

Your total payment is €8.0. You will also receive the bottle of wine. [Participant who did not sell the bottle to experimenter]

Please remain seated. The experimenter will come to your cubicle for the payment. After that you may leave the room. Please remember to pick your personal belongings that you stored in the lockers.



Chapter 4

Whistleblowing under competition

4.1 Introduction

Corporate fraud presents a pressing concern for various stakeholders, impacting the economy, society, and eroding public trust in the financial system. Despite its substantial welfare costs, only a third of all corporate fraud cases are ever detected (Dyck et al., 2023). Whistleblowing emerges as a vital tool to expose and deter fraudulent activities within organisations (Leder-Luis, 2023). Whistleblowing entails reporting illegal or unethical behaviour by employees of a firm, leading to a conflict between moral responsibility and loyalty to the firm (Mesmer-Magnus and Viswesvaran, 2005; Waytz et al., 2013; Dungan et al., 2019). Consequently, the study of whistleblowing behaviour and its determinants has garnered significant academic interest and policy relevance in the field of economics.

Examining whistleblowing empirically is inherently complex and faces identification and measurement challenges, as only detected fraud and blown whistles can be observed.¹ To address these issues, researchers have increasingly adopted experimental approaches to study whistleblowing. The experimental literature is primarily motivated by providing empirical evidence for emerging whistleblowing laws that safeguard whistleblowers from retaliation or offer financial incentives for uncovering fraud.² A common limitation in existing studies is that they focus on the behaviour of experimental firms in isolation, overlooking industries where firms impact each other's revenues. Competitive pressures may provide strategic incentives undermining whistleblowing and facilitating lawbreaking, as well as erode moral values. Thus, the primary goal of our chapter is to fill this gap by examining whistleblowing and lawbreaking in a competitive setting. The second goal of our chapter is to study whether beliefs about the frequency of and judgements about the appropriateness of whistleblowing and lawbreaking moderate the effect of competition on unethical behaviour (i.e., breaking the law and not blowing the whistle).

Our experiment builds upon the whistleblowing game introduced by Butler et al. (2020). In this game, managers are provided with the chance to break the law which can yield personal gains for themselves and their employees, but at the detriment of other participants who act as members of the public. Employees, on the other hand, are not victims of the manager's unlawful conduct but rather benefit from it. They are given the choice to report their manager's wrongdoing, which incurs a cost for the employee and results in an automatic monetary penalty imposed on the manager. Our treatments vary whether firms operate independently or compete for market revenue. We predict that competition will decrease whistleblowing and

¹Whistleblowing has also been examined from a theoretical perspective (see for instance Heyes and Kapur (2009) and Givati (2016)).

²Examples of such laws include the Public Interest Disclosure Act in the United Kingdom (1998), the Dodd-Frank Act in the United States (2010), and the more recent EU Whistleblower Protection Directive in the European Union (2019).

increase lawbreaking.

We also employ a post-experiment incentivised survey eliciting beliefs about the frequency of whistleblowing and lawbreaking, and appropriateness ratings of such actions. We conjecture that the direct effect of competition on whistleblowing and lawbreaking will be mediated by the indirect effect of beliefs and judgements on behaviour.

Overall, we find little evidence that competition affects unethical behaviour. We find an insignificant decrease on whistleblowing and a marginally significant increase in lawbreaking under competition. Having found a null treatment effect, there is no scope for beliefs and morality judgements to mediate the effect. Thus, we reformulate the second goal of our chapter to investigate whether beliefs and morality judgements are correlated with whistleblowing and lawbreaking behaviour, and whether competition affects them. We find that beliefs significantly correlate with behaviour, whereas morality judgements do so less strongly. We also find that competition moves beliefs in the direction of observed behaviour whereas morality judgements are not significantly affected by competition.

Our chapter contributes to the existing experimental literature on whistleblowing within firms, which has primarily focused on identifying financial factors that encourage individuals to blow the whistle. Prior studies have highlighted the positive impact of incentives such as monetary rewards (Breuer, 2013; Schmolke and Utikal, 2018; Butler et al., 2020), protection from dismissal (Wallmeier, 2019; Mechtenberg et al., 2020), and the threat of fines for non-reporting (Schmolke and Utikal, 2018). Our research sheds light on a previously overlooked aspect – the potentially negative effect of competition on whistleblowing. If competition does decrease whistleblowing, then the existing literature's conclusions should be seen as upper bounds on the prevalence of whistleblowing. However, our findings suggest that competition has a minimal effect on whistleblowing.

Experimental work has also studied non-pecuniary motivations of whistleblowers. Bartuli et al. (2016) investigated personality factors and attitudes, and found that employees who score higher in the Honesty-Humility factor, who are more altruistic, and are more aware of ethical issues are more likely to blow the whistle. Antinyan et al. (2020) found that higher trust in the government and institutions also increases the likelihood to blow the whistle. Motivated by the fact that whistleblowers are sometimes seen as heroes and sometimes as snitches, Butler et al. (2020) provide experimental evidence that the expected social approval or disapproval of whistleblowers by the public affects their behaviour. We contribute to this literature by studying how beliefs and morality judgements are used to justify not blowing the whistle and breaking the law.

Relevant to our study is also a well established literature on whistleblowing be-

tween firms in the context of cartels. Since cartels by definition involve multiple firms, competition between firms is always present. The rich experimental literature has provided evidence that various forms of leniency programs on cartel reporting, such as offering financial rewards (Apesteguia et al., 2007; Hinloopen and Soetevent, 2008; Bigoni et al., 2012) or providing full or partial amnesty (Bigoni et al., 2015; Feltovich and Hamaguchi, 2018) to whistleblowing firms who report a cartel, are effective against cartel formation and price fixing, and promote cartel discovery.³ Notably, (Hamaguchi et al., 2009) vary the level of competition between firms by comparing cartels of two or of seven firms, and find that competition increases the effectiveness of leniency programs as cartels are less sustainable among seven firms.

Furthermore, our study is closely related to the broader literature exploring the effect of competition on unethical behaviour. Previous research has operationalised competition through competitive payment schemes (Schwieren and Weichselbaumer, 2010; Gill et al., 2013; Savikhin and Sheremeta, 2013; Cartwright and Menezes, 2014; Buser and Dreber, 2016; Schurr and Ritov, 2016; Vadera and Pathki, 2021) or market-based settings (Falk and Szech, 2013; Bartling et al., 2015, 2019, 2023; Ziegler et al., 2020). A recent meta-analysis by Huber et al. (2023) examines 45 experimental designs and reports a small overall effect of competition on morality. The general consensus is that competition tends to lead to more unethical behaviour. Our experimental design aligns more closely with the competitive payment schemes, and our results reveal that competition leads to a small insignificant increase of unethical behaviour.

The remaining of the chapter is organised as follows. section 4.2 provides a detailed presentation of the whistleblowing game, the experimental design, our predictions, and the implementation. All results are presented in section 4.3. We end the chapter with section 4.4 which interprets the results, and suggests areas for future research.

4.2 Experimental design and predictions

4.2.1 The baseline whistleblowing game

The whistleblowing game is based on Butler et al. (2020) and has been adapted for this study. In the game, nine participants are randomly assigned to three firms, each consisting of one manager and two employees. Additionally, six participants play the role of members of the public. The inclusion of a larger number of members of the public compared to the size of each firm aims to recreate, in a laboratory setting, the

³Hinloopen and Normann (2009) and Hinloopen et al. (2023) provide a comprehensive overview of experiments on leniency programs for cartel reporting.

sentiment that society, which may be harmed by corporate fraud, is larger than the firm committing the fraud.

The employees are given an addition task, where they are provided with six pairs of two-digit numbers and asked to report the sum of each pair. For each correctly reported sum, they earn 20 Experimental Currency Units (ECU) as private earnings and contribute 10 ECU to the firm's surplus. They have 120 seconds to complete this task. The members of the public also engage in the same addition task, but they only earn private earnings.

The manager receives a fixed income of 120 ECU. Furthermore, they have the opportunity to double the firm's surplus by choosing one of two options. The first option involves a multiplication task, where the manager is provided with six pairs of two-digit numbers and asked to report the product of each pair. If they report at least three correct products, the firm's surplus is doubled. If they fail to do so, the surplus remains unchanged. The second option is to break the law, which automatically doubles the firm's surplus without the manager engaging in the multiplication task. However, breaking the law results in a loss of 20 ECU for each of the six members of the public. Importantly, when making their decision, the manager is unaware of the size of the surplus created by the employees. This setup prevents managers from basing their decision to break the law on the performance of their employees, ensuring comparability of manager decisions across firms. The final surplus is distributed among the firm members, with the manager keeping half of the surplus and each employee receiving a quarter.

The employees have the option to blow the whistle if their manager breaks the law. Their willingness to blow the whistle is elicited using the strategy method. At the moment of deciding to blow the whistle, the employees are unaware of whether their manager has broken the law. If the manager does break the law, one of the two employees is randomly selected, and their decision to blow the whistle is implemented. If the manager does not break the law, the employee's decision is not implemented. Blowing the whistle comes at a cost of 25 ECU for the selected employee and imposes a penalty of 70 ECU on the manager.

4.2.2 Treatments

Our experiment consists of two primary treatments. The Baseline treatment follows the design described in the previous subsection, where the firms operate independently of each other. As highlighted in the introduction, this design serves as our baseline comparison.

In the Competition treatment, we introduce a modification to break the independence among firms in a straightforward manner. Before redistributing the surpluses of the firms, we rank the surpluses. The firm with the largest surplus emerges as

the winner of the competition and has its surplus further increased by 50%, while the each of the other two firms experience a decrease of 25% in their surpluses. In case of a tie, the winner is selected randomly with uniform probability. This treatment creates a tournament incentive structure and introduces strategic uncertainty as firm members must form beliefs about the behaviour of other firms. The competition treatment is designed to resemble industries with relatively few firms, where the decisions of each firm can significantly impact the distribution of market revenues among them.

Our design included a set of treatments which are only insightful as robustness checks if a treatment effect is established. Since we do not observe a significant treatment effect, we relegate detailed description and analysis of those additional treatments to subsection 4.5.1. In short, the additional treatments included a treatment where the winner of the competition was determined randomly, and a replication of our main two treatments in a setting whether the members of the public receive passive income instead of performing the task.

4.2.3 Predictions

Denoting the baseline and the competition treatments as B and C respectively, we denote the probability with which the manager breaks the law in treatment j by b^j , and the probability with which the selected employee blows the whistle in treatment j by w^j , where $j \in \{B, C\}$. We also denote by d_b^j and d_s^j respectively, the moral cost of the manager for breaking the law and the moral cost of the employee for not blowing the whistle in treatment j . Finally, we denote the expected firm surplus produced by the two employees before the manager makes any decision by S , and the probability that the manager can correctly solve the six multiplication problems by a .

In the baseline treatment, breaking the law yields the manager their fixed income and half of the doubled surplus, but their utility is reduced by the moral costs of breaking the law and by the expected punishment if the selected employee blows the whistle.

$$U_b^B = 120 + \frac{1}{2} \times 2S - 70w^B - d_b^B = 120 + S - 70w^B - d_b^B$$

Doing the multiplication task yields the fixed income and half of the surplus; the surplus may or may not be doubled depending on the manager's ability.

$$U_t^B = 120 + \frac{1}{2} [a \times 2S + (1 - a) \times S] = 120 + \frac{1 + a}{2} S$$

The manager weakly prefers to break the law if

$$U_b^B \geq U_t^B \Rightarrow \frac{1-a}{2}S \geq 70w^B + d_b^B \quad (4.1)$$

Intuitively, a manager is more likely to break the law if they are of low ability, if whistleblowing is less likely, or if their moral costs of doing so are low.

To derive the equivalent condition for the competition treatment, we further denote the probability of winning the competition if the surplus of the firm is not doubled by π_{1S} , and the probability of winning the competition if the surplus of the firm is doubled by π_{2S} . By definition we have $\pi_{2S} \geq \pi_{1S}$. Given that the surplus before the manager makes their decision is expected to be similar between all firms,⁴ and that ties are broken with uniform probability, π_{2S} can never be lower than one third. This is evident as the lowest probability of winning after having doubling the surplus is obtained in the case when all other firms also doubled their surplus, where all firms have exactly one third chance of winning. Similarly, if the surplus is not doubled, π_{1S} can never exceed one third. π_{1S} is zero if any other firm doubled their surplus and is positive only if no other firm doubled their surplus. In that case firms are tied in terms of surplus and each firm has one third chance of winning. Thus, $\pi_{2S} \geq \frac{1}{3} \geq \pi_{1S}$.

In the competition treatment, breaking the law yields the manager

$$U_b^C = 120 + \frac{1}{2} \times \left(\pi_{2S} \times \frac{3}{2} + (1 - \pi_{2S}) \times \frac{3}{4} \right) 2S - 70w^C - d_b^C = \\ 120 + \frac{3(1 + \pi_{2S})}{4} S - 70w^C - d_b^C$$

whereas doing the multiplication task yields

$$U_t^C = 120 + \frac{1}{2} \left[a \left(\pi_{2S} \times \frac{3}{2} + (1 - \pi_{2S}) \times \frac{3}{4} \right) 2S + (1 - a) \left(\pi_{1S} \times \frac{3}{2} + (1 - \pi_{1S}) \times \frac{3}{4} \right) S \right] = \\ 120 + \left[\frac{3a(1 + \pi_{2S})}{4} + \frac{3(1 - a)(1 + \pi_{1S})}{8} \right] S$$

The manager weakly prefers to break the law if

$$U_b^B \geq U_t^B \Rightarrow \frac{1-a}{2} \left[\frac{3}{2} \left(\frac{1}{2} + \pi_{2S} - \frac{\pi_{1S}}{2} \right) \right] S \geq 70w^C + d_b^C \quad (4.2)$$

Since $\pi_{2S} \geq \frac{1}{3} \geq \pi_{1S}$, the term between brackets is necessarily weakly larger than 1. Thus, by comparing the multipliers of the surplus in Equation 4.1 and Equation 4.2, we see that the expected benefits of breaking the law are larger under competition

⁴The addition task is relatively straightforward and almost all employees are expected to solve all number adding problems correctly. In fact, out of 2,016 sets of the six addition problems that the employees in our experiment worked on, they solved all six correctly in 1,888 of them (93.6%) with almost all other instances resulting in five correct. Thus, the surplus of each firm before the manager decision was roughly the same (120, 60 from each employee).

than under baseline. Assuming that employees blow the whistle less under competition $w^C \leq w^B$, we predict that managers will break the law more frequently under competition. The tendency of managers to break the law more under competition may be further strengthened if competition reduces the moral costs of breaking the law ($d_b^C \leq d_b^B$).

Prediction 4.1. *The propensity of managers to break the law will be higher under competition.*

For the employee in treatment j , the expected cost of blowing the whistle is $\frac{1}{2} \times 25 \times b^j$, whereas the expected benefit is avoiding the moral costs of staying silent d_s^j . Thus under baseline the employee prefers to blow the whistle if

$$d_s^B \geq \frac{25}{2}b^B \tag{4.3}$$

Under competition, the employee prefers to blow the whistle if

$$d_s^C \geq \frac{25}{2}b^C \tag{4.4}$$

Given the first prediction that managers will be more likely to break the law under competition ($b^C \geq b^B$), we predict that employees will be less likely to blow the whistle under competition. This tendency may be further facilitated if under competition the moral cost of staying silent is lower ($d_s^C \leq d_s^B$).

Prediction 4.2. *The propensity of employees to blow the whistle will be lower under competition.*

We continue with a discussion of how beliefs about the frequency of whistleblowing and lawbreaking are expected to correlate with whistleblowing and lawbreaking behaviour. We inform our discussion based on the predictions above, but also by discussing non-pecuniary motivations.

We begin with the beliefs about the frequency of behaviour of participants in the other role, i.e., the beliefs of managers about the frequency of whistleblowing, and the beliefs of employees about the frequency of lawbreaking. From Equation 4.1 and Equation 4.2, we note that managers are more likely to break the law if they expect fewer employees to blow the whistle as this reduces the expected punishment they may receive. Similarly, from Equation 4.3 and Equation 4.4, we note that employees are more likely to blow the whistle when they expect fewer managers to break the law as this reduces the expected cost of blowing the whistle. Thus, we would expect the propensity to break the law to be negatively correlated with the perceived likelihood that employees will blow the whistle, and the propensity to blow the whistle to be negatively correlated with the perceived likelihood that managers will break the law.

The effect may be even more pronounced if managers use low whistleblowing and employees use high lawbreaking as evidence that unethical behaviour is prevalent around them, allowing them to further justify their actions.⁵

Next, we look at beliefs about the behaviour of participants in the same role. From a strategic point of view, the perceived frequency of whistleblowing should not affect the decision to blow the whistle as shown in Equation 4.3 and Equation 4.4. However, employees may have non-pecuniary reasons related to the belief about the frequency of whistleblowing. More specifically, employees may use their belief that fewer employees will blow the whistle to justify their own decision to stay silent. In other words, employees may reason that staying silent is not so bad since everyone is doing it.⁶ Thus, we would expect the propensity to blow the whistle to be positively correlated with the belief that other employees are blowing the whistle.

For managers in the baseline treatment, the perceived frequency of lawbreaking has no strategic effects in their behaviour as shown in Equation 4.1. On the contrary, in the competition treatment, lawbreaking of other firms becomes relevant. While in both treatments a manager is better off by doubling the surplus, in the competition treatment doubling the surplus has the additional benefit of increasing the probability of winning the competition. However, the expected benefits of doing so are decreasing with the probability that other managers double their surplus too. To illustrate, if neither of the other managers double their surplus, then doing so increases the manager's winning probability from $\frac{1}{3}$ to 1. However, if both other managers double their surplus, then doing so only increases the manager's winning probability from 0 to $\frac{1}{3}$.⁷ Since the probability of doubling the surplus is an increasing function of the probability of breaking the law, it follows that the expected benefits of breaking the law are also decreasing with the probability that other managers break the law. In short, best-responding would suggest a negative correlation between lawbreaking and the perceived frequency of lawbreaking only in the competition treatment. However, non-pecuniary motivations are also present in both treatments, and similarly to employees, managers may use the higher perceived prevalence of lawbreaking as a

⁵This reasoning resembles a form of social weighting, a rationalisation under which people are find examples of others that are similarly or more corrupt than themselves in order to justify their own corrupt behaviour while shielding their moral identity (Ashforth and Anand, 2003).

⁶This reasoning resembles a form of denial of responsibility, a rationalisation under which people rationalise their corrupt behaviour with the belief that others in their position are also engaging in the same behaviour (Ashforth and Anand, 2003). It can also be interpreted as evidence for a preference for conformity if participants are motivated by following what others are doing (Fatas et al., 2018), or as evidence for false consensus if they project their own behaviour and expect it to be more prevalent among the general population (Aronson et al., 2016).

⁷Formally, denoting the probability that another manager doubles the surplus by π , we can rewrite the winning probabilities as $\pi_{1,S} = \frac{1}{3}(1-\pi)^2 = \frac{1}{3}\pi^2 - \frac{2}{3}\pi + \frac{1}{3}$ and $\pi_{2,S} = \frac{1}{3}\pi^2 + \frac{1}{2}[2\pi(1-\pi)] + (1-\pi)^2 = \frac{1}{3}\pi^2 - \pi + 1$. Both probabilities are decreasing in π . Further substituting the probabilities in the multiplier of the surplus in Equation 4.2, we obtain the multiplier as $\frac{1-a}{8}(\pi^2 - 4\pi + 8)$, which is also decreasing in π .

justification to break the law themselves.⁸

Finally, as indicated in Equation 4.1 and Equation 4.2, managers are more likely to break the law if they suffer a lower moral cost from doing so. Similarly, employees are less likely to blow the whistle if they suffer a lower moral cost from staying silent (Equation 4.3 and Equation 4.4). Thus, we expect lawbreaking to be negatively correlated with moral costs, and whistleblowing to be positively correlated with moral costs. If competition erodes morals, we would expect the moral costs to be lower under competition.

4.2.4 Post-experiment survey

The survey consists of several parts. In the first part, we elicit beliefs about our primary behavioural outcomes. Participants are asked to estimate the frequency of employees blowing the whistle (between 0 and 72) and the frequency of managers breaking the law (between 0 and 36) in their session. If their estimates fall within three units of the correct values, they earn 20 ECU.

Next, we elicit morality judgements by asking participants to rate the appropriateness of three actions: a manager breaking the law, an employee not blowing the whistle, and the public losing part of their earnings. Participants rate the morality of employee and manager behaviour on a Likert scale ranging from very immoral to very moral, and the acceptability of the public losing earnings on a scale from very unacceptable to very acceptable. Additionally, participants indicate their level of loyalty to their firm on a scale from not loyal at all to very loyal. We incentivise participants with the Krupka and Weber (2013) method. The participants earn 10 points for each judgement if their responses match the modal answer in their session.

In the third part, we assess participants' risk preferences using the lottery task introduced by Eckel and Grossman (2002). Participants are presented with a series of lotteries with increasing expected payoffs and greater variance. Their choices are incentivised and realised by the computer. Next, we explore participants' social preferences using the Social Value Orientation method developed by Murphy et al. (2011). Participants make six decisions. In each decision, they are provided with nine pairs of payoffs for themselves and another participant, and they must select one option for each decision. Participants are informed that one decision will be randomly chosen for payment, that they will be randomly paired with another participant, and that the decision of one of the two participants will be implemented. Finally, we gather standard demographic information such as age, gender, and field of study.

⁸As for employees, this reasoning can be interpreted as evidence for denial of responsibility, preference for conformity, or false consensus.

4.2.5 Implementation

Each session consisted of exactly 15 participants. The main deviation from the original one-shot game of Butler et al. (2020) is the fact that the game was played repeatedly for twelve rounds. In the first round, all participants were randomly assigned their role. The three managers retained their role for the remaining of the experiment. This reflects the fact that managers, who on average receive high wages, face smaller variability in their income over time.⁹ Employees and members of the public alternated roles in each round. Changing roles between rounds simulated the fact that employees of firms in one industry are members of the public with respect to other industries. The employees were randomly assigned to a firm. The matching ensured that employees and members of the public would not be part of the same firm in consecutive rounds.

Following the original design of Butler et al. (2020), we use a framed experiment. As can be seen in the instructions on subsection 4.5.2, the labels we used for the participant roles and the actions of the game matched the description provided here.¹⁰ Given the nuance associated with whistleblowing, we believe providing contextual cues is necessary for our experiment. Alekseev et al. (2017) summarised existing evidence and concluded that “using evocative language either does not affect behaviour or affects it in a desirable way by evoking the desired emotional response.”

We did not provide feedback on the decisions of the employees and managers between rounds. This design feature, together with the role switching, and the random reassignment of employees in firms between rounds, aimed at mitigating reputation effects. At the same time, we were interested in eliciting beliefs about the prevalence and appropriateness of behaviour at the end of the experiment, and the absence of feedback provided us with a cleaner measure. We did provide feedback on individual performance (the number of correct answers provided) and feedback on whether their firm won the competition (only in the competition treatment).

The experimental sessions took place between April and October of 2023. Half of the participants were recruited from the participant pool of the Birmingham Experimental Economics Laboratory at the University of Birmingham, and half from the participant pool of the Cambridge Experimental and Behavioural Economics Group at the University of Cambridge. The experiment was programmed in oTree (Chen et al., 2016) and preregistered (Ioannidis, 2023). Ethical approval was obtained from the University of Amsterdam, the University of Birmingham, and the University of Cambridge.¹¹ Informed consent was collected from all participants at the beginning of

⁹Previous whistleblowing experiments with repeated games also had managers keeping their role for the duration of the experiment (Mechtenberg et al., 2020).

¹⁰Framing effects have been documented to influence behaviour in a range of environments such as public good games (Sonnemans et al., 1998; Cookson, 2000; Cartwright, 2016) and dictator games (Dreber et al., 2013; Goerg et al., 2020).

¹¹Ethics approval from the University of Amsterdam was provided by the Ethics Committee of Eco-

each session. Clear instructions were provided to participants on screen as well as in print, and they had to answer a series of comprehension questions correctly before making decisions.

Each treatment arm involved 120 participants across eight sessions, resulting in a total of 240 participants. The participants were on average 22.59 years old (SD = 4.45, min = 18, max = 41) and came from various fields of study (27% Social Sciences, 22% Natural and Applied Sciences, 17% Humanities, 34% Other). The gender distribution was relatively balanced, with 55% female, 42% male, and 3% non-binary gendered participants. Table 4.3 in the appendix provides a break down of demographic variables across all treatments and participant pools. Each participant took part in only one experimental session and received an average payment of £11.21 (SD = 1.71, min = £5.70, max = £15.5) for approximately 75 minutes, including a participation fee of £2.00.

4.3 Results

4.3.1 The effect of competition on whistleblowing and lawbreaking

Figure 4.1 presents an initial overview of whistleblowing and lawbreaking across our treatments. The figure shows bar graphs separately for employee whistleblowing (left) and manager lawbreaking (right).

To support each result, we conduct two sets of tests. Firstly, we aggregate observations from the eight sessions of each treatment and conduct Mann-Whitney ranksum tests. While aggregation limits us to only eight observations per treatment and may reduce statistical power, this approach ensures using truly independent observations. Importantly, any significant comparison using this conservative method offers robust evidence of a treatment effect. To further validate our ranksum tests, we employ econometric estimations of treatment effects using linear probability models, with errors clustered at the session level.¹²

nomics and Business on January 30, 2023 (Reference: EB-953). Ethics approval from the University of Birmingham was provided by the Humanities and Social Sciences Committee on March 6, 2023 (Reference: ERN 0894-3). Ethics approval from the University of Cambridge was provided by the Cambridge Judge Business School Departmental Ethics Review Group on September 7, 2023 (Reference: ERN 23-32).

¹²Our preregistration included a power analysis which was based on the linear probability model. We assumed that the probability of whistleblowing in the baseline treatment would be 0.23 as in Butler et al. (2020). Based on this assumption, our minimum detectable effect size with 120 participants per treatment would be 0.10, i.e., we would be able to reject the hypothesis of no difference in whistleblowing between baseline and competition if the competition decreases whistleblowing by at least 0.10. In our baseline, we observe higher whistleblowing (around 30%) than in Butler et al. (2020). This difference may be driven by the fact that our experiment was repeated whereas the original was one-shot.

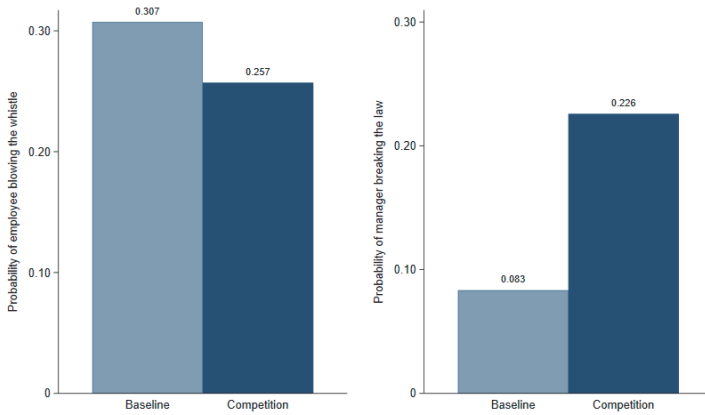


Figure 4.1: Whistleblowing (left) and lawbreaking (right) across treatments

We first focus on whistleblowing behaviour. 30.7% of employees blew the whistle in Baseline and 25.7% did so in Competition. The difference is insignificant (Mann-Whitney ranksum test, $z = 1.582, p = 0.1136, N = 16$). Our null result is further illustrated econometrically. We regress the decision to blow the whistle on the treatment variable, while controlling for risk, social value orientation, and demographics. The estimation reveals a small insignificant decrease of whistleblowing under competition ($b = -0.030, SE = 0.038, CI = [-0.118, 0.039], t = -0.82, p = 0.432, N = 1152$).

Result 1. *Competition results in an insignificant decrease in employee whistleblowing.*

Next, we analyse the lawbreaking decisions of managers. 8.3% of managers broke the law in Baseline and 22.6% in Competition. The difference is marginally significant (Mann-Whitney ranksum test, $z = 1.954, p = 0.0507, N = 16$). Our econometric estimation provides a similar picture. An analogous regression of the decision to break the law on treatment provides evidence for a marginally significant increase in lawbreaking under competition ($b = 0.146, SE = 0.070, CI = [-0.004, 0.295], t = 2.07, p = 0.056, N = 576$).

Result 2. *Competition results in a marginally significant increase in manager lawbreaking.*

4.3.2 The effect of beliefs and morality judgements on behaviour

This subsection aims at investigating two questions. First, we document whether beliefs about the frequency and judgements about the appropriateness of whistleblowing and lawbreaking are correlated with observed behaviour. Second, we study whether beliefs and judgements were affected by competition.

For our first question, we augment the econometric analysis from before by adding beliefs and judgements to the model. Results are shown in Table 4.1. The table reports estimates from linear probability models using as dependent variable the decision to blow the whistle (first two columns) and the decision to break the law (last three columns).

	Blow the whistle		Break the law		
	(1)	(2)	(3)	(4)	(5)
Competition	-0.030 (0.038)	0.016 (0.024)	0.146* (0.070)	0.043 (0.035)	-0.014 (0.074)
Belief about frequency of whistleblowing		0.707*** (0.046)		-0.170 (0.134)	-0.117 (0.158)
Belief about frequency of lawbreaking		-0.306*** (0.099)		0.705*** (0.222)	0.538* (0.276)
Competition*Belief about lawbreaking					0.234 (0.341)
Appropriateness of employee staying silent		-0.238* (0.124)			
Appropriateness of manager breaking the law				-0.089 (0.155)	-0.101 (0.160)
Appropriateness of public losing earnings		-0.140* (0.075)		0.401* (0.198)	0.411** (0.192)
Loyalty to the firm		-0.130** (0.051)		-0.108 (0.109)	-0.111 (0.117)
Controls	Yes	Yes	Yes	Yes	Yes
Observations	1152	1152	576	576	576

Standard errors in parentheses, clustered on matching group level.

Significance levels * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Controls: Age, Gender, Study, Risk, Social Value Orientation

Table 4.1: The effect of beliefs and judgements on whistleblowing and lawbreaking behaviour

We find that mostly employees –and to a lesser extent also managers– respond to their beliefs about the behaviour of their counterpart in the expected direction. The probability that an employee blows the whistle is significantly negatively correlated with the expected belief about the frequency of lawbreaking (column 2), suggesting that the higher expected cost of whistleblowing associated with more frequent lawbreaking deterred employees from blowing the whistle. While not significantly so, the probability of a manager breaking the law is negatively correlated with the expected belief about the frequency of whistleblowing (column 4), suggesting that managers also break the law less when the expected punishment from doing so is higher, but the effect is weak.

On the contrary, beliefs about the behaviour of participants in the same role do not follow the direction of best-responding. As discussed in the prediction section earlier, best-responding would suggest that the probability to blow the whistle would be uncorrelated with the perceived frequency of whistleblowing, whereas the proba-

bility to break the law would be negatively correlated with the perceived frequency of lawbreaking only in the competition treatment. Both for employees and for managers, non-pecuniary motivations would suggest a positive correlation. Our estimates are consistent with the latter motivation as both for employees (column 2) and for managers (column 4), we observe a significant positive correlation between their own behaviour and their stated belief about the prevalence of the same behaviour in their session. In column 5, we explicitly check if there is a negative correlation between breaking the law and belief about frequency of lawbreaking in the competition treatment by interacting the treatment dummy with the belief of lawbreaking and find no evidence of a negative correlation.¹³

Finally, we find suggestive evidence that moral costs influenced the behaviour of participants. Employees are less likely to blow the whistle when the moral cost of staying silent is lower, when they believe that the public losing part of their earnings is more morally acceptable, and when they are more loyal to their firm. Managers are more likely to break the law if they find that the public losing part of their earnings is more morally acceptable.

Observation 4.1. *Whistleblowing and lawbreaking are significantly correlated with beliefs about the frequency of such behaviour, and marginally significantly correlated with morality judgements.*

Finally, we test whether competition affected beliefs and morality judgements. Table 4.2 presents the elicited beliefs and judgements across our treatments. For beliefs about whistleblowing and judgement of an employee who did not blow the whistle, we only consider observations from participants in the role of employee. Symmetrically, for beliefs about lawbreaking and judgement of a manager who broke the law, we only consider observations from participants in the role of manager.¹⁴

We observe that competition affected the elicited beliefs in the direction of the observed behaviour suggesting that our participants were roughly accurate in their estimations. Employees expected significantly less whistleblowing and managers expected marginally significantly more lawbreaking.¹⁵ We find no evidence that moral

¹³When eliciting beliefs, we asked employees to estimate the frequency of whistleblowing including their own behaviour which accounted for 6 out of 72 decisions. Similarly, for managers the elicitation included 12 out of 36 of their own decisions. Thus, our belief variable may produce biased results as it is endogenous. Repeating the analysis using modified beliefs, i.e., the originally stated beliefs after subtracting the decisions of each participant, provides qualitatively similar results.

¹⁴Comparing beliefs and judgements between employees and managers, we only find that managers report higher loyalty to their firm. This can arguably be attributed to the our implementation as managers were acting as members of a firm for all twelve rounds of the whistleblowing game whereas employees were members of a firm for only six out of twelve rounds.

¹⁵To check for the presence of false consensus, we repeat our tests using observations from participants in a different role. We observe that managers do not expect more whistleblowing under competition ($p = 0.7645$, $N = 48$) and employees do not expect more lawbreaking under competition ($p = 0.990$, $N = 192$), suggesting that indeed participants do project to some extent their own behaviour when estimating the frequency of behaviour of participants in the same role.

	Belief about frequency of whistleblowing	Belief about frequency of lawbreaking	Appropriateness of an employee staying silent	Appropriateness of a manager breaking the law	Appropriateness of the public losing earnings	Loyalty to the firm
Baseline	0.3539	0.1968	0.4688	0.3125	0.1646	0.5896
Competition	0.2804	0.3148	0.4687	0.2396	0.2042	0.6167
p-value	0.0254	0.0787	0.9045	0.1734	0.1504	0.4984
Observations	192	48	192	48	240	240

All beliefs and judgements are standardised to be between 0 and 1. The p-values are from Mann-Whitney ranksum tests.

Table 4.2: Beliefs and morality judgements across treatments

costs were affected by competition as all appropriateness ratings are similar between treatments (columns 3-6).

Observation 4.2. *There is suggestive evidence that competition affected the beliefs of employees and managers, whereas morality judgements were unaffected by competition.*

4.4 Concluding discussion

This chapter investigates the determinants of whistleblowing, focusing on conditions that pose threats to the act of whistleblowing itself. While existing literature has predominantly explored factors that facilitate whistleblowing, our study sheds light on whether a competitive environment hinders it. Given that promising policy interventions incur costs, either in terms of monetary rewards to whistleblowers or implementation of protective laws, our study can be interpreted as answering the question of whether policy makers should prioritise more competitive industries where whistleblowing may be less prevalent and consequently interventions may improve social welfare the most. However, we find that competition does not decrease whistleblowing significantly.

Our results further reveal that morality judgements are not heavily relied upon when deciding whether to blow the whistle or whether to break the law, whereas beliefs about the prevalence of those behaviours are. When operating in a competitive environment, we find no evidence that unethical behaviour is more acceptable, whereas we find evidence that it is perceived to be more common. Thus, our findings roughly suggest that competition does not imply that unethical behaviour per se is perceived as morally less bad, but that engaging in unethical behaviour is a lesser threat to one's image when more others are believed to behave unethically.

We end our discussion with a brief comment on the importance of generalisability and replication. Even though we had no reason to expect ex-ante different behaviour across participant pools, competition affected whistleblowing behaviour differently across the university of Birmingham and the university of Cambridge participant pools.¹⁶ While in both cases the level of whistleblowing in the baseline was

¹⁶For more details on the results commented here, we refer to subsection 4.5.1.

similar, competition decreased whistleblowing in the first case whereas whistleblowing was unaffected in the latter. One plausible reason for this discrepancy is statistical randomness as any analysis within each participant pool has lower power and is prone to produce either false positives or false negatives. Another plausible reason is unobserved differences between participant pools that our experiment and survey were not designed to capture. To illustrate, our framed experiment used terms such as competition, whistleblowing, and breaking the law. It is conceivable that our framing differentially affected participants that may differ in their ethnic and cultural background, political orientation, general attitudes towards competition, or general attitudes towards cheating.

Consider as a thought experiment that two researchers had run sufficiently powered replications of our treatments using different participant pools, and behaviour within each participant pool followed the distinct patterns we comment on here. One researcher would claim that competition decreases whistleblowing, and the other researcher would claim that competition does not affect whistleblowing. With this perspective in mind, our study suggests that further research is needed to establish the conditions under which the effect (if any) of competition on whistleblowing emerges.

4.5 Appendix to Chapter 4

4.5.1 Additional treatments and exploratory results

This appendix serves two purposes. We first present an exploratory comparison of behaviour between the two participant pools. Next, we describe in detail the additional robustness treatments we ran and briefly present observations from those treatments.

4.5.1.1 Behaviour between participant pools

Table 4.3 breaks down demographic characteristics of participants in all our sessions across treatments and participant pools. Our sessions and treatments are balanced.¹⁷

We summarise behaviour in Table 4.4, which reports whistleblowing and law-breaking across treatments and participant pools.

Observation 4.3. *Comparing behaviour across participant pools, we find that*

- (a) *whistleblowing is similar in baseline, but under competition we observe more whistleblowing within university of Cambridge participants;*

¹⁷Formally, we test whether demographics differ per treatment-participant pool combination. We find no evidence for differences in either age (ANOVA test, $F = 0.31, p = 0.8200, N = 240$), gender (Pearson chi-squared test, $\chi^2_6 = 6.04, p = 0.419, N = 240$) or field of study (Pearson chi-squared test, $\chi^2_9 = 15.78, p = 0.072, N = 240$).

Treatment	Participant pool	Age		Gender			Field of study		
		Male	Female	Other	Social	Natural	Humanities	Other	
Baseline	Birmingham	22.4	0.38	0.57	0.05	0.22	0.20	0.10	0.48
	Cambridge	22.4	0.47	0.50	0.03	0.28	0.23	0.18	0.31
Competition	Birmingham	23.1	0.48	0.52	0.00	0.23	0.22	0.13	0.42
	Cambridge	22.5	0.33	0.63	0.04	0.33	0.23	0.25	0.19

Each row corresponds to one combination of treatment and participant pool. Each combination involves 60 participants.

Table 4.3: Demographics across participant pools for primary treatments

	Whistleblowing		Lawbreaking	
	Baseline	Competition	Baseline	Competition
Birmingham	0.319	0.191	0.007	0.188
Cambridge	0.295	0.323	0.159	0.264
p-value	0.468	0.083	0.018	0.561
p-value	0.594	0.054	0.038	0.424

Notes on rows:

Row (3): Mann-Whitney tests between participant pools

Row (4): Linear Probability Model of participant pool coefficient.

Table 4.4: Whistleblowing and lawbreaking across participant pools

(b) *lawbreaking is similar under competition, but in baseline we observe more lawbreaking within university of Cambridge participants.*

Repeating the analysis from the main text, we observe that within the university of Birmingham participants, competition significantly decreased whistleblowing (Mann-Whitney: $p = 0.0209$, $N = 8$) and significantly increased lawbreaking (Mann-Whitney: $p = 0.0247$, $N = 8$), whereas within university of Cambridge participants, competition affected neither whistleblowing (Mann-Whitney: $p = 0.7702$, $N = 8$) nor lawbreaking (Mann-Whitney: $p = 0.5590$, $N = 8$). Estimates from linear probability models within each participant pool show the same pattern.

Observation 4.4. *Within each participant pool, we find that*

(a) *competition significantly decreased whistleblowing and significantly increased lawbreaking within the university of Birmingham participants;*

(b) *competition did not affect whistleblowing or lawbreaking within the university of Cambridge participants;*

In the next subsection we focus on the university of Birmingham participant pool and the robustness treatments we ran there.

4.5.1.2 Robustness treatments

In our baseline treatment firms operated independently of each other whereas in our competition treatment the firm with the largest surplus would get a 50% bonus in

market revenue. In contrast, there are industries where the distribution of market revenues is less sensitive to the strategic behaviour of the firms. This can be due to factors such as the size of the industry, where many small firms have a smaller influence on aggregate market outcomes, or the presence of exogenous shocks. In our Random treatment, we incorporate such market structures. This treatment is similar to the competition treatment, with the only difference being the method used to determine the winning firm. Instead of ranking the firm surpluses, we randomly select one of the firms as the winner. It is important to note that the total amount of ECU available to the firms remains the same across treatments; the only difference lies in how it is distributed among the firms. In the random treatment, the probabilities of winning the competition are $\pi_{1S} = \pi_{2S} = \frac{1}{3}$; for these winning probabilities the multiplier of the surplus (see Equation 4.2) collapses to 1. Assuming that in the random treatment the moral costs of breaking the law and not blowing the whistle are higher compared to baseline, but lower compared to competition, we conjecture that lawbreaking and whistleblowing behaviour would be in between baseline and competition.

We also expand our experimental design by introducing two additional treatments, BaselineNotask and CompetitionNotask.¹⁸ These treatments replicate the baseline and competition treatments with a single key difference. In our main treatments, both members of the public and employees perform the same task. However, in real-world scenarios, employees and managers of firms often hold the belief that the public, which could be negatively affected by corporate fraud, is less deserving. In our additional treatments, the members of the public receive a passive income of 140 ECU without participating in the number adding task. This modification allows us to examine how the attitudes and actions of employees and managers may vary based on their perception of the worthiness of the public's income. If the moral costs of breaking the law and not blowing the whistle are lower when the public receives passive income, then keeping the level of competition constant, we predict that the managers will be more likely to break the law and the employees will be less likely to blow the whistle in our additional treatments compared to our primary treatments.

Table 4.5 indicates that demographic characteristics from our treatments from the university of Birmingham participant pool are roughly balanced in both the three primary treatments and the two additional treatments.¹⁹ Figure 4.2 presents whistle-

¹⁸Hypotheses for these additional treatments (as well as for the random treatment) were not included in the preregistration. The analysis presented in this Appendix is thus entirely exploratory in nature.

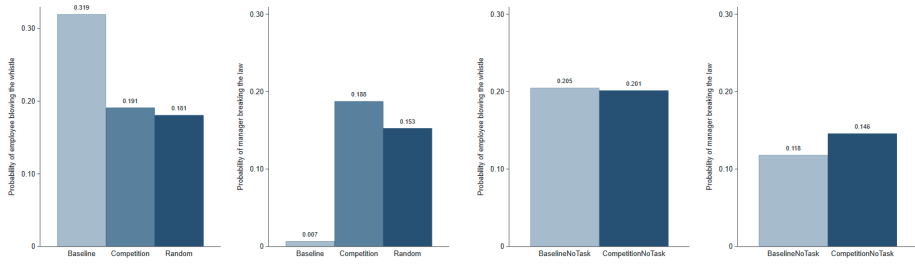
¹⁹In the three primary treatments (Baseline, Competition, Random) there are no differences in either age (ANOVA test, $F = 0.52, p = 0.5956, N = 180$), gender (Pearson chi-squared test, $\chi_4^2 = 5.80, p = 0.214, N = 180$) or field of study (Pearson chi-squared test, $\chi_6^2 = 10.61, p = 0.101, N = 180$). Similarly, for the two additional treatments (BaselineNotask, CompetitionNotask) there are no differences in either age (ANOVA test, $F = 2.92, p = 0.0901, N = 120$), gender (Pearson chi-squared test, $\chi_2^2 = 2.71, p = 0.258, N = 180$) or field of study (Pearson chi-squared test, $\chi_3^2 = 1.45, p =$

Treatment	Age	Gender			Field of study			
		Male	Female	Other	Social	Natural	Humanities	Other
Baseline	22.4	0.38	0.57	0.05	0.22	0.20	0.10	0.48
Competition	23.1	0.48	0.52	0.00	0.23	0.22	0.13	0.42
Random	23.2	0.53	0.45	0.02	0.32	0.35	0.03	0.30
BaselineNotask	21.9	0.38	0.62	0.00	0.28	0.33	0.12	0.27
CompetitionNotask	20.8	0.27	0.72	0.01	0.33	0.30	0.17	0.20

Each row corresponds to a single treatment involving 60 participants.

Table 4.5: Demographics across all treatments of the university of Birmingham participant pool

blowing behaviour and lawbreaking across all treatments. We emphasise that all results from this section are exploratory and have lower power than the results presented in the main text.



(a) When public engages in effort task

(b) When public receives passive income

Figure 4.2: Bar graphs of whistleblowing and lawbreaking over treatments

Repeating the same analyses as in the main text, we find that when members of the public perform an effort task, competition decreases employee whistleblowing (Mann-Whitney: $p = 0.0209$, $N = 8$) and increases manager lawbreaking (Mann-Whitney: $p = 0.0247$, $N = 8$). Whether the winner of the competition is determined based on performance or randomly does not matter. When the members of the public receive passive income, competition does not decrease employee whistleblowing (Mann-Whitney: $p = 0.7674$, $N = 8$) or increase manager lawbreaking (Mann-Whitney: $p = 0.7715$, $N = 8$).

We end the analysis by comparing treatments where the public performed the task with treatments where the public received passive income. Without competition, we find less whistleblowing (Mann-Whitney: $p = 0.0814$, $N = 8$) and more lawbreaking (Mann-Whitney: $p = 0.0172$, $N = 8$) when the public receives passive

0.694, $N = 120$). There are small but significant differences between primary and additional treatments with the latter having younger participants and more female participants (age: ANOVA test, $F = 3.04$, $p = 0.0178$, $N = 300$, gender: Pearson chi-squared test, $\chi^2_6 = 6.04$, $p = 0.419$, $N = 240$, field of study: Pearson chi-squared test, $\chi^2_9 = 15.78$, $p = 0.072$, $N = 240$), but the results presented here are organised separately for those treatments.

income compared to when they perform the task. With competition, whistleblowing and lawbreaking are not affected by whether the public receives passive income or performs the task. We also observe that the public losing part of their earnings due to manager lawbreaking is perceived as less severe when the members of the public perform a task compared to when they receive passive income (Mann-Whitney: $p = 0.0173$, $N = 120$).

Observation 4.5. *Within university of Birmingham participants, we find that*

- (a) *when members of the public perform an effort task, competition decreases employee whistleblowing and increases manager lawbreaking. Whether the winner of the competition is determined based on performance or randomly does not matter;*
- (b) *when the members of the public receive passive income, competition does not affect whistleblowing or lawbreaking;*
- (c) *there is less whistleblowing and more lawbreaking when the public receives passive income than when they engage in the effort task;*
- (d) *the public is perceived as less worthy of their income when they receive passive income.*

4.5.2 Instructions and decision screens

Welcome screen

Welcome

Welcome to this experiment. Please read the following instructions carefully. We ask that you do not communicate with other participants during the experiment. The use of mobile phones is not allowed during this experiment. If you have any questions, or need assistance of any kind, at any time, an experimenter will assist you privately. The data collected through this experiment does not include your name or any other information that would allow your identification. All the data you provide during the experiment cannot be traced back to you.

Payment

In addition to your participation fee, you may earn substantially more money from today's experiment. You will be paid privately and anonymously in cash at the end of the experimental session today. Earnings during the experiment will be denominated in Experimental Currency Units, or ECU. Each ECU is worth £0.01. The participation

fee is £2.00. After the experiment finishes, you will be paid the money you earned plus your participation fee.

Duration

You will be asked to make decisions in 12 rounds.

Tasks and decisions

Employee Task

The two employees have the task of adding two numbers and report their sum. Each correct answer gives them 20 points. Additionally each correct answer by each employee generates 10 points to the firm surplus. In total, each employee will be given 6 pairs of numbers to add. Employees have 120 seconds to solve as many as they can.

Manager Decision & Task

Managers have the opportunity to double the firm surplus. Their decision is to choose how they want to try to double the surplus. They can do so in two ways:

- **Do the Manager Task:** In the task, managers are asked to multiply two numbers and report their product. Managers will be given 6 pairs of numbers to multiply. Managers have 120 seconds to solve as many as they can. If they give at least 3 correct answers, the firm surplus is doubled.
- **Break the law:** If the manager breaks the law, they skip the Manager Task and the firm surplus will be doubled automatically. However, it will also generate a loss of 20 points to each of the 6 members of the public.

Employee Decision

The employees have the option to blow the whistle if their manager broke the law. The decision to blow the whistle will be relevant ONLY if the manager broke the law. If the manager did not break the law, this decision will not be implemented. The employees will make this decision before the manager makes their decision. Each employee can decide whether they are willing to blow the whistle. Blowing the whistle will cost the selected employee 25 points, and will generate a penalty of 70 points to the manager. This penalty will be removed from the manager's earnings. One of the two employees will randomly be selected and their decision will be implemented. The decision of the other employee will not be implemented.

Members of the Public Task

The members of the public have the task of adding two numbers and report their sum. Each correct answer gives them 20 points. In total, each member of the public will be given 6 pairs of numbers to add.

Members of the public have 120 seconds to solve as many as they can. [Task treatments]

Members of the public do not make any decisions for this round. [No task treatments]

Payoff explanations

Winning firm [Not in Baseline treatment]

One of the three firms will win the competition and their surplus will be further increased by 50%. Each of the other two firms that did not win the competition will have their surplus reduced by 25%.

The winning firm will be the firm with the largest surplus. [Competition]

The winning firm will be randomly selected. [Random]

Surplus distribution

The total firm surplus will be distributed as follows: The manager keeps 50% of the surplus, and each employee will keep 25%.

Payoff explanation

- Employees:
 - Earn 20 points for each correct answer in number-adding task.
 - Earn 25% of the final firm surplus.
 - Lose 25 points if they blow the whistle and their decision is implemented.
- Managers:
 - Earn fixed income of 120 points.
 - Earn 50% of the final firm surplus.
 - Lose 70 points if they broke the law and the selected employee blew the whistle.
- Members of the public:
 - Earn 20 points for each correct answer in number-adding task. [Task treatments]
 - Earn a passive income of 140 points. [No task treatments]

-
- Lose 20 points for each manager that broke the law.

Paid rounds

At the end of the session, 4 rounds will be randomly selected for payment.

Survey

Estimates

Please answer the following questions. For each question, you will receive additional 20 points if your answer is within a range of 3 of the actual number. Your earnings from this task will be added to your earnings from the game.

- The managers had the choice between breaking the law and doing the multiplication task 36 times in total during the experiment. How many times do you think they broke the law?
- The employees had the choice to blow the whistle or not 72 times in total during the experiment. How many times do you think they blew the whistle?

Judgements

Please answer the following questions. For each question, you will receive additional 10 points if your answer matches the answer of the majority of the participants in today's experiment. Your earnings from this task will be added to your earnings from the game.

- How would you judge a manager who broke the law?
(Very immoral, Immoral, No judgement, Moral, Very moral)
- How would you judge an employee who did not blow the whistle?
(Very immoral, Immoral, No judgement, Moral, Very moral)
- How do you feel about the members of the public losing earnings because managers broke the law?
(Very unacceptable, Unacceptable, No judgement, Acceptable, Very acceptable)
- In the rounds that you were part of a firm, how loyal did you feel to the firm?
(Not loyal at all, Not very loyal, Neutral, Loyal, Very loyal)

Lottery task

In the following task, 6 different lotteries are presented on your screen. In each of these lotteries, both rewards A and B are equally likely, i.e. have a probability of exactly 50%. The rewards are denoted in points.

You are asked to choose exactly one of the lotteries, which subsequently will be implemented. A random generator will determine whether you win reward A or reward B, respectively. At the end of the experiment, your reward will be added to you earnings.

Reward A 50% Probability	Reward B 50% Probability	Your Choice
140	140	
120	180	
100	220	
80	260	
60	300	

Distribution task

In this task you have been randomly paired with another person, whom we will refer to as the other. This other person is someone you do not know and will remain mutually anonymous. All of your choices are completely confidential. You will be making a series of 6 decisions about allocating resources between you and this other person. For each of the following questions, please indicate the distribution you prefer most by choosing the button along the midline. You can only choose one distribution for each of the 6 questions. Your decisions will yield money for both yourself and the other person.

There are no right or wrong answers, this is all about personal preferences. One of the 6 decisions will randomly be selected and implemented. At the end of the experiment, the outcome will be added to you earnings.

Demographics

Please enter the following information.

- Please indicate your age.
- Please indicate your field of study.
(Economics, Social Sciences, Natural Sciences, Humanities, Applied Sciences, Other)
- Please indicate your gender.
(Male, Female, Prefer not to answer)

You receive	85	85	85	85	85	85	85	85	85
Choose									
Other receives	85	76	68	59	50	41	33	24	15
You receive	85	87	89	91	93	94	96	98	100
Choose									
Other receives	15	19	24	28	33	37	41	46	50
You receive	50	54	58	63	68	72	76	81	85
Choose									
Other receives	100	98	96	94	93	91	89	87	85
You receive	50	54	59	63	68	72	76	81	85
Choose									
Other receives	100	89	79	68	58	47	36	26	15
You receive	100	94	88	81	75	69	63	56	50
Choose									
Other receives	50	56	63	69	75	81	88	94	100
You receive	100	98	96	94	93	91	89	87	85
Choose									
Other receives	50	54	59	63	68	72	76	81	85

Summaries

Summary in English

This thesis consists of four chapters. Each is an independent essay on how information affects how people make decisions. Below is a brief summary of each chapter.

Chapter 1 provides a game-theoretic analysis of the Verifiability Approach. The Verifiability Approach is a verbal deception detection method built the following premise. When asked to provide a statement, a truth-teller will provide as many precise details as possible, whereas a liar will face a dilemma between providing many details in order to appear innocent and providing fewer details to avoid getting exposed. A liar will solve this dilemma by providing many vague details. The ratio of verifiable precise details over unverifiable vague details becomes a signal of truthfulness.

In the chapter we model this interaction between a speaker, who may be truth-teller or liar, and an investigator, who wants to uncover the true type of the speaker. Our equilibrium analysis indicates that the best an investigator can achieve in this setting is a partially-separating equilibrium. In this equilibrium, precise statements are investigated frequently enough so that liars are disincentivised from providing them too often.

Chapter 2 provides experimental data answering the key question of whether habits affect communication. A consistently observed phenomenon in the experimental communication literature is that communication between senders and receivers is more informative than expected under canonical models of strategic communication (Crawford and Sobel, 1982). We conjecture that this pattern can be attributed to the environments most people typically communicate in; environments where senders and receivers often have common interests. In such environments, senders may form the habit of telling the truth and receivers may form the habit of trusting information. If our conjecture is true, then senders and receivers who primarily interact in environments characterised by conflicting interests may form habits of lying and distrusting information respectively, and consequently undercommunicate.

To test our conjecture, we conduct a two-stage experiment. We vary the alignment of interests between senders and receivers in the first stage in order to simulate environments where different communication habits may form. In the second stage of the experiment, senders and receivers interact in an environment with partial interest alignment (i.e., between the two extremes of full alignment and full misalignment of the first stage). We find that senders and receivers communicate more informatively if stage one is a common interest environment compared a conflicting interests environment. We interpret this pattern as evidence that indeed habits affect communication behaviour. We additionally vary how often the new unfamiliar environment occurs, and find that habits developed in stage one affect behaviour only when the new environment occurs rarely, whereas the effect of habits disappears if the new

environment occurs frequently.

Chapter 3 provides an experiment on anchoring. Anchoring is a cognitive bias under which irrelevant information affects decisions, and specifically valuations of goods. Our primary question in this chapter is to investigate participating in a market mitigates (or even eliminates) anchoring effects. Contrary to our predictions, our experiment finds no evidence of anchoring as valuations are unaffected by the random anchor. Naturally, this implies that there is no scope for market interaction to eliminate the effect, though we do find that market participation moves valuations closer to those of other participants. We provide meta-analytic evidence suggesting that our failure to find anchoring effects may be due the way the anchor was determined. Our participants rolled a die whose outcomes formed the anchor, a procedure which may be perceived as transparently uninformative.

Chapter 4 provides an experiment on whistleblowing. The experimental literature on whistleblowing behaviour within firms has primarily studied settings where firms operate independently from each other. We conjecture that competition for market revenue provides motivations against whistleblowing, and is unaccounted for in previous studies. We conduct an experiment with two treatments, with and without competition, and find that competition has little effect as it results in an insignificant reduction on whistleblowing. We also find evidence that behaviour correlates with beliefs, but it does not correlate with morality judgements.

Summary in Dutch

Dit proefschrift bestaat uit vier hoofdstukken. Elk hoofdstuk is een essay over hoe informatie invloed heeft op de besluitvorming van mensen. Hieronder volgt een beknopte samenvatting van elk hoofdstuk.

Hoofdstuk 1 biedt een speltheoretische analyse van de “Verifiability Approach”. Dit is een methode om verbale misleiding te detecteren op basis van de volgende aanname. Wanneer gevraagd wordt om een verklaring, zal een waarheidsgetrouwe persoon zoveel mogelijk precieze details verstrekken, terwijl een leugenaar voor een dilemma staat tussen het verstrekken van veel details om onschuldig over te komen en het verstrekken van minder details om niet ontdekt te worden. Een leugenaar zal dit dilemma oplossen door veel vage details te verstrekken. De verhouding van verifieerbare, precieze details ten opzichte van onverifieerbare vage details wordt daarmee een indicatie van geloofwaardigheid.

In het hoofdstuk modelleren we deze interactie tussen een spreker, die waarheidsgetrouw of leugenachtig kan zijn, en een onderzoeker, die het ware type van de spreker wil onthullen. Onze evenwichtsanalyse geeft aan dat het beste wat een onderzoeker in deze setting kan bereiken, een gedeeltelijk scheidingsevenwicht is. In dit evenwicht worden precieze verklaringen vaak genoeg onderzocht zodat leugenaars worden ontmoedigd om ze al te vaak te verstrekken.

Hoofdstuk 2 bespreekt de resultaten van een laboratorium-experiment opgezet om antwoord te geven op de vraag of gewoontes de communicatie beïnvloeden. Een consistent waargenomen fenomeen in de experimentele communicatieliteratuur is dat communicatie tussen zenders en ontvangers informatiever is dan kan worden verwacht volgens canonieke modellen van strategische communicatie (Crawford and Sobel, 1982). We veronderstellen dat dit patroon kan worden toegeschreven aan de omgevingen waarin de meeste mensen doorgaans communiceren; omgevingen waarin zenders en ontvangers vaak gemeenschappelijke belangen hebben. In dergelijke omgevingen kan het zijn dat zenders de gewoonte hebben om de waarheid te vertellen en ontvangers de gewoonte hebben om informatie te vertrouwen. Als deze hypothese waar is, kunnen zenders en ontvangers die voornamelijk in omgevingen communiceren die worden gekenmerkt door conflicterende belangen, gewoontes vormen van liegen en het wantrouwen van informatie, en bijgevolg ondercommuniceren.

Om onze hypothese te testen, voeren we een tweestapexperiment uit. We variëren de overeenkomst van belangen tussen zenders en ontvangers in de eerste fase om omgevingen te simuleren waarin verschillende communicatiegewoontes kunnen ontstaan. In de tweede fase van het experiment communiceren zenders en ontvangers in een omgeving met gedeeltelijke belangenovereenkomst (dat wil zeggen, tussen de twee uitersten van volledig gemeenschappelijke en volledig conflicterende

belangen uit de eerste fase). We vinden dat zenders en ontvangers informatiever communiceren als de eerste fase een omgeving met gemeenschappelijke belangen was in vergelijking met een omgeving met conflicterende belangen. We interpreteren dit patroon als bewijs dat gewoontes inderdaad van invloed zijn op het communicatiegedrag. We variëren ook hoe vaak de nieuwe onbekende omgeving voorkomt en vinden dat gewoontes ontwikkeld in de eerste fase het gedrag alleen beïnvloeden wanneer de nieuwe omgeving zelden voorkomt, terwijl het effect van gewoontes verdwijnt als de nieuwe omgeving vaak voorkomt.

Hoofdstuk 3 bespreekt een laboratorium-experiment over verankering. Verankering is een cognitieve vertekening waarbij irrelevante informatie beslissingen beïnvloedt, en specifiek waarderingen van goederen. De primaire vraag in dit hoofdstuk is om te onderzoeken of deelname aan een markt verankerings-effecten vermindert (of zelfs elimineert). In tegenstelling tot onze voorspellingen levert ons experiment geen bewijs voor verankering, aangezien waarderingen onaangetast blijven door het willekeurige anker. Dit betekent ook dat er geen ruimte is voor marktinteractie om het effect te elimineren, hoewel we wel constateren dat deelname aan de markt waarderingen dichter bij die van andere deelnemers brengt. We bieden meta-analytisch bewijs dat suggereert dat het niet vinden van verankerings-effecten in ons experiment te wijten kan zijn aan de manier waarop het anker werd bepaald. Onze deelnemers gooiden zelf een dobbelsteen waarvan de uitkomsten het anker vormden, een procedure die als overduidelijk oninformatief kan worden beschouwd.

Hoofdstuk 4, tenslotte, bespreekt een experiment over klokkenluiden. De experimentele literatuur tot nu toe heeft klokkenluidersgedrag bestudeerd in omgevingen waarin bedrijven onafhankelijk van elkaar opereren. Het is echter aannemelijk dat in de praktijk concurrentie om marktinkomsten invloed heeft op klokkeluidersgedrag, omdat het mogelijkere wijs motieven biedt die klokkeluiden minder aantrekkelijk maken. We voeren een experiment uit met twee verschillende groepen, d.w.z. met en zonder concurrentie, en vinden dat concurrentie weinig effect heeft omdat het resulteert in een niet significante afname van klokkenluiden. We vinden daarnaast dat gedrag correleert met overtuigingen, maar niet correleert met morele oordelen.

Summary in Greek

Αυτή η διατριβή αποτελείται από τέσσερα κεφάλαια. Το καθένα είναι ένα ανεξάρτητο δοκίμιο σχετικά με το πώς οι πληροφορίες επηρεάζουν τον τρόπο που οι άνθρωποι λαμβάνουν αποφάσεις. Παρακάτω παρέχεται μια σύντομη περίληψη κάθε κεφαλαίου.

Το Κεφάλαιο 1 παρέχει μια ανάλυση παιχνιδιού για τη Μέθοδο Επαλήθευσης (**Verifiability Approach**). Η Μέθοδος Επαλήθευσης είναι μια μέθοδος ανίχνευσης αληθοφάνειας βασισμένη σε δύο υποθέσεις. Όταν κάποιος καλείται να δώσει μια δήλωση, ένας λέγων αλήθεια θα παρέχει όσο το δυνατόν περισσότερες ακριβείς λεπτομέρειες, ενώ ένας ψεύτης θα αντιμετωπίσει ένα δίλημμα μεταξύ του να παρέχει πολλές λεπτομέρειες για να φανεί αθώος και να παρέχει λιγότερες λεπτομέρειες για να αποφύγει να αποκαλυφθεί. Ένας ψεύτης θα λύσει αυτό το δίλημμα παρέχοντας πολλές μα ασαφείς λεπτομέρειες. Ο λόγος των επαληθεύσιμων ακριβών λεπτομερειών ως προς τις ανεπαλήθευτες ασαφείς λεπτομέρειες γίνεται ένα δείκτης αληθείας.

Στο κεφάλαιο μοντελοποιούμε αυτήν την αλληλεπίδραση μεταξύ ενός ομιλητή, ο οποίος μπορεί να λέει αλήθεια ή ψέματα, και ενός ερευνητή, ο οποίος θέλει να αποκαλύψει τον πραγματικό τύπο του ομιλητή. Η ανάλυσή μας βασίζεται στο σημείο ισορροπίας (**Nash equilibrium**) δείχνει ότι το καλύτερο που μπορεί να επιτύχει ένας ερευνητής σε αυτήν την περίπτωση είναι ένα μερικά-διαχωριστικό σημείο ισορροπίας. Σε αυτό το σημείο ισορροπίας, οι ακριβείς δηλώσεις εξετάζονται αρκετά συχνά, ώστε οι ψεύτες να αποτρέπονται από το να τις παρέχουν πολύ συχνά.

Το Κεφάλαιο 2 παρέχει πειραματικά δεδομένα που απαντούν στην κύρια ερώτηση εάν οι συνήθειες επηρεάζουν την επικοινωνία. Ένα φαινόμενο που παρατηρείται συνεχώς στην πειραματική βιβλιογραφία είναι η υπερεπικοινωνία, δηλαδή η επικοινωνία μεταξύ αποστολέων και παραληπτών είναι περισσότερο ενημερωτική από ό,τι αναμένεται βάσει κανονικών μοντέλων στρατηγικής επικοινωνίας (**Crawford and Sobel, 1982**). Εικάζουμε ότι αυτό το μοτίβο μπορεί να αποδοθεί στα περιβάλλοντα στα οποία οι περισσότεροι άνθρωποι επικοινωνούν συχνότερα, περιβάλλοντα όπου οι αποστολείς και οι παραλήπτες συχνά έχουν κοινά συμφέροντα. Σε τέτοια περιβάλλοντα, οι αποστολείς μπορεί να δημιουργήσουν τη συνήθεια να λένε την αλήθεια και οι παραλήπτες μπορεί να δημιουργήσουν τη συνήθεια να εμπιστεύονται τις πληροφορίες. Εάν η εικασία μας είναι αληθής, τότε οι αποστολείς και οι παραλήπτες που επικοινωνούν κυρίως σε περιβάλλοντα που χαρακτηρίζονται από συγκρούσεις συμφερόντων μπορεί να δημιουργήσουν συνήθειες ψεύδους και δυσπιστίας αντίστοιχα και, συνεπώς, να υποεπικοινωνούν.

Για να ελέγξουμε την υπόθεσή μας, διενεργούμε ένα πείραμα σε δύο στάδια. Στο πρώτο στάδιο, ποικίλλουμε την κατεύθυνση των συμφερόντων μεταξύ αποστολέων και παραληπτών προκειμένου να προσομοιώσουμε περιβάλλοντα όπου ενδέχεται να διαμορφωθούν διαφορετικές συνήθειες επικοινωνίας. Στο δεύτερο στάδιο του πειράματος, οι αποστολείς και οι παραλήπτες αλληλεπιδρούν σε ένα περιβάλλον με μερική ευθυγράμμιση συμφερόντων (δηλαδή, μεταξύ των δύο άκρων της πλήρους ταύτισης και της πλήρους

αντίθεσης του πρώτου σταδίου). Βρίσκουμε ότι οι αποστολείς και οι παραλήπτες επικοινωνούν περισσότερο ενημερωτικά αν το πρώτο στάδιο είναι ένα περιβάλλον με κοινά συμφέροντα σε σύγκριση με ένα περιβάλλον με συγκρουόμενα συμφέροντα. Ερμηνεύουμε αυτό το μοτίβο συμπεριφοράς ως απόδειξη ότι πράγματι οι συνθήκες επηρεάζουν την επικοινωνία. Επιπλέον, ποικίλλουμε το πόσο συχνά εμφανίζεται το νέο άγνωστο περιβάλλον και βρίσκουμε ότι οι συνθήκες που αναπτύχθηκαν στο πρώτο στάδιο επηρεάζουν τη συμπεριφορά μόνο όταν το νέο περιβάλλον εμφανίζεται σπάνια, ενώ το αποτέλεσμα των συνηθειών εξαφανίζεται εάν το νέο περιβάλλον εμφανίζεται συχνά.

Το Κεφάλαιο 3 παρέχει ένα πείραμα για το **anchoring**. Το **anchoring** είναι μια γνωστική προδιάθεση σύμφωνα με την οποία μη-χρήσιμες πληροφορίες επηρεάζουν τις αποφάσεις, και ειδικότερα τις εκτιμήσεις των αγαθών. Η κύρια ερώτηση σε αυτό το κεφάλαιο είναι να εξετάσουμε αν η συμμετοχή σε μια αγορά ελαττώνει (ή ακόμη και εξαλείφει) την προδιάθεση του **anchoring**. Παρά τις προβλέψεις μας, το πείραμά μας δεν βρήκε κανένα αποτέλεσμα **anchoring**, καθώς οι εκτιμήσεις των αγαθών δεν επηρεάζονται από την τυχαία πληροφορία. Αυτό σημαίνει ότι δεν υπάρχει περιθώριο για την αλληλεπίδραση της αγοράς για να εξαλείψει το αποτέλεσμα, αν και βρίσκουμε ότι η συμμετοχή στην αγορά μετακινεί τις εκτιμήσεις πιο κοντά στις εκτιμήσεις των άλλων συμμετεχόντων. Παρέχουμε μετα-αναλυτικά στοιχεία που υποδεικνύουν ότι η αποτυχία μας να βρούμε αποτελέσματα του **anchoring** μπορεί να οφείλεται στον τρόπο που καθορίστηκε η τυχαία πληροφορία. Οι συμμετέχοντες μας έριξαν ένα ζάρι, τα αποτελεσμάτα του οποίου σχημάτισαν την τυχαία πληροφορία, μια διαδικασία που μπορεί να θεωρηθεί διαφανώς μη-ενημερωτική.

Το Κεφάλαιο 4 παρέχει ένα πείραμα σχετικά με τον καταγγελτικό λόγο (**whistleblowing**). Η πειραματική βιβλιογραφία μέχρι σήμερα έχει κυρίως μελετήσει την καταγγελτική συμπεριφορά σε περιβάλλοντα όπου οι επιχειρήσεις λειτουργούν ανεξάρτητα μεταξύ τους. Υποθέτουμε ότι ο ανταγωνισμός για το μερίδιο της αγοράς παρέχει κίνητρα εναντίον του καταγγελτικού λόγου, και δεν λαμβάνεται υπόψη σε προηγούμενες μελέτες. Διενεργούμε ένα πείραμα με δύο συνθήκες, με και χωρίς ανταγωνισμό, και βρίσκουμε ότι ο ανταγωνισμός έχει μικρό αποτέλεσμα καθώς οδηγεί σε μια στατιστικά μη-σημαντική μείωση του καταγγελτικού λόγου. Βρίσκουμε επίσης ενδείξεις ότι η συμπεριφορά συσχετίζεται με τις πεποιθήσεις, αλλά δε συσχετίζεται με τις ηθικές κρίσεις.

Bibliography

- Abeler, J., Becker, A., and Falk, A. (2014). Representative evidence on lying costs. *Journal of Public Economics*, 113(1):96–104.
- Abeler, J., Nosenzo, D., and Raymond, C. (2019). Preferences for truth-telling. *Econometrica*, 87(4):1115–1153.
- Acland, D. and Levy, M. R. (2015). Naiveté, projection bias, and habit formation in gym attendance. *Management Science*, 61(1):146–160.
- Agranov, M. and Schotter, A. (2012). Ignorance is bliss: an experimental study of the use of ambiguity and vagueness in the coordination games with asymmetric payoffs. *American Economic Journal: Microeconomics*, 4(2):77–103.
- Alekseev, A., Charness, G., and Gneezy, U. (2017). Experimental methods: When and why contextual instructions are important. *Journal of Economic Behavior & Organization*, 134:48–59.
- Alevy, J. E., Landry, C. E., and List, J. A. (2015). Field experiments on the anchoring of economic valuations. *Economic Inquiry*, 53(3):1522–1538.
- Anderson, B. A. (2016). The attention habit: How reward learning shapes attentional selection. *Annals of the New York Academy of Sciences*, 1369(1):24–39.
- Antinyan, A., Corazzini, L., and Pavesi, F. (2020). Does trust in the government matter for whistleblowing on tax evaders? Survey and experimental evidence. *Journal of Economic Behavior & Organization*, 171:77–95.
- Apestequia, J., Dufwenberg, M., and Selten, R. (2007). Blowing the whistle. *Economic Theory*, 31(1):143–166.
- Arbel, Y., Bar-El, R., Siniver, E., and Tobol, Y. (2014). Roll a die and tell a lie—what affects honesty? *Journal of Economic Behavior & Organization*, 107(1):153–172.
- Ariely, D., Loewenstein, G., and Prelec, D. (2003). Coherent arbitrariness: Stable demand curves without stable preferences. *The Quarterly Journal of Economics*, 118(1):73–106.

-
- Ariely, D., Loewenstein, G., and Prelec, D. (2006). Tom sawyer and the construction of value. *Journal of Economic Behavior & Organization*, 60(1):1–10.
- Aronson, E., Wilson, T. D., and Sommers, S. R. (2016). *Social psychology*. Pearson.
- Ashforth, B. E. and Anand, V. (2003). The normalization of corruption in organizations. *Research in Organizational Behavior*, 25:1–52.
- Baker, S. and Mezzetti, C. (2001). Prosecutorial resources, plea bargaining, and the decision to go to trial. *Journal of Law, Economics, and Organization*, 17(1):149–167.
- Balbuzanov, I. (2019). Lies and consequences: The effect of lie detection on communication outcomes. *International Journal of Game Theory*, 48(4):1203–1240.
- Banks, J. S. and Sobel, J. (1987). Equilibrium selection in signaling games. *Econometrica*, 55(3):647–661.
- Bartling, B., Fehr, E., and Özdemir, Y. (2023). Does market interaction erode moral values? *The Review of Economics and Statistics*, 105(1):226–235.
- Bartling, B., Valero, V., and Weber, R. (2019). On the scope of externalities in experimental markets. *Experimental Economics*, 22(3):610–624.
- Bartling, B., Weber, R. A., and Yao, L. (2015). Do markets erode social responsibility? *The Quarterly Journal of Economics*, 130(1):219–266.
- Bartuli, J., Mir Djawadi, B., and Fahr, R. (2016). Business ethics in organizations: An experimental examination of whistleblowing and personality. *IZA Discussion Paper*.
- Becker, G. M., DeGroot, M. H., and Marschak, J. (1964). Measuring utility by a single-response sequential method. *Systems Research and Behavioral Science*, 9(3):226–232.
- Becker, G. S. and Murphy, K. M. (1988). A theory of rational addiction. *Journal of political Economy*, 96(4):675–700.
- Bell, B. E. and Loftus, E. F. (1989). Trivial persuasion in the courtroom: The power of (a few) minor details. *Journal of Personality and Social Psychology*, 56(5):669–679.
- Belot, M. and van de Ven, J. (2019). Is dishonesty persistent? *Journal of Behavioral and Experimental Economics*, 83:1–9.
- Bergman, O., Ellingsen, T., Johannesson, M., and Svensson, C. (2010). Anchoring and cognitive ability. *Economics Letters*, 107(1):66–68.

-
- Bigoni, M., Fridolfsson, S.-O., Le Coq, C., and Spagnolo, G. (2012). Fines, leniency, and rewards in antitrust. *The RAND Journal of Economics*, 43(2):368–390.
- Bigoni, M., Fridolfsson, S.-O., Le Coq, C., and Spagnolo, G. (2015). Trust, leniency, and deterrence. *The Journal of Law, Economics, and Organization*, 31(4):663–689.
- Bjerk, D. (2007). Guilt shall not escape or innocence suffer? the limits of plea bargaining when defendant guilt is uncertain. *American Law and Economics Review*, 9(2):305–329.
- Blume, A., Lai, E. K., and Lim, W. (2020). Strategic information transmission: A survey of experiments and theoretical foundations. In *Handbook of Experimental Game Theory*. Edward Elgar Publishing.
- Bohm, P., Lindén, J., and Sonnegård, J. (1997). Eliciting reservation prices: Becker–degroot–marschak mechanisms vs. markets. *The Economic Journal*, 107(443):1079–1089.
- Bond Jr, C. F. and DePaulo, B. M. (2006). Accuracy of deception judgments. *Personality and Social Psychology Review*, 10(3):214–234.
- Breuer, L. (2013). Tax compliance and whistleblowing—the role of incentives. *The Bonn Journal of Economics*, 2(2):7–44.
- Buser, T. and Dreber, A. (2016). The flipside of comparative payment schemes. *Management Science*, 62(9):2626–2638.
- Butler, J. V., Serra, D., and Spagnolo, G. (2020). Motivating whistleblowers. *Management Science*, 66(2):605–621.
- Byrne, D. P., Goette, L., Martin, L. A., Delahey, L., Jones, A., Miles, A., Schob, S., Staake, T., and Tiefenbeck, V. (2021). The habit-forming effects of feedback: Evidence from a large-scale field experiment. Working paper, SSRN.
- Cabrales, A., Feri, F., Gottardi, P., and Meléndez-Jiménez, M. A. (2020). Can there be a market for cheap-talk information? an experimental investigation. *Games and Economic Behavior*, 121(1):368–381.
- Cai, H. and Wang, J. T.-Y. (2006). Overcommunication in strategic information transmission games. *Games and Economic Behavior*, 56(1):7–36.
- Caplin, A. (2016). Measuring and modeling attention. *Annual Review of Economics*, 8(1):379–403.
- Card, N. A. (2015). *Applied meta-analysis for social science research*. Guilford Publications.

-
- Cartwright, E. (2016). A comment on framing effects in linear public good games. *Journal of the Economic Science Association*, 2(1):73–84.
- Cartwright, E. and Menezes, M. L. (2014). Cheating to win: Dishonesty and the intensity of competition. *Economics Letters*, 122(1):55–58.
- Chapman, G. B. and Johnson, E. J. (1999). Anchoring, activation, and the construction of values. *Organizational behavior and human decision processes*, 79(2):115–153.
- Charness, G. and Gneezy, U. (2009). Incentives to exercise. *Econometrica*, 77(3):909–931.
- Charness, G., Gneezy, U., and Imas, A. (2013). Experimental methods: Eliciting risk preferences. *Journal of Economic Behavior & Organization*, 87(1):43–51.
- Chen, D. L., Schonger, M., and Wickens, C. (2016). otree—an open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, 9(1):88–97.
- Chen, Y., Kartik, N., and Sobel, J. (2008). Selecting cheap-talk equilibria. *Econometrica*, 76(1):117–136.
- Chetty, R. and Szeidl, A. (2016). Consumption commitments and habit formation. *Econometrica*, 84(2):855–890.
- Cohen, A. (1992). *The Living Law: A Guide to Modern Legal Research*. Rochester, N.Y.: Lawyers Cooperative.
- Cohen, S. A. and Doob, A. N. (1989). Public attitudes to plea bargaining. *Criminal Law Quarterly*, 32(1):85–109.
- Cohn, A., Fehr, E., and Maréchal, M. A. (2014). Business culture and dishonesty in the banking industry. *Nature*, 516(7529):86–89.
- Cohn, A., Maréchal, M. A., and Noll, T. (2015). Bad boys: How criminal identity salience affects rule violation. *The Review of Economic Studies*, 82(4):1289–1308.
- Cohn, A., Maréchal, M. A., Tannenbaum, D., and Zünd, C. L. (2019). Civic honesty around the globe. *Science*, 365(6448):70–73.
- Cookson, R. (2000). Framing effects in public goods experiments. *Experimental Economics*, 3:55–79.
- Coppock, A. and Green, D. P. (2016). Is voting habit forming? new evidence from experiments and regression discontinuities. *American Journal of Political Science*, 60(4):1044–1062.

-
- Costa-Gomes, M. A. and Crawford, V. P. (2006). Cognition and behavior in two-person guessing games: An experimental study. *American economic review*, 96(5):1737–1768.
- Crawford, V. P. and Sobel, J. (1982). Strategic information transmission. *Econometrica*, 50(6):1431–1451.
- Dal Bó, P. and Fréchette, G. R. (2011). The evolution of cooperation in infinitely repeated games: Experimental evidence. *American Economic Review*, 101(1):411–29.
- Daubert v. Merrell Dow Pharmaceuticals, Inc.* (1993). 509 U.S. 579-601.
- de Groot Ruiz, A., Offerman, T., and Onderstal, S. (2015). Equilibrium selection in experimental cheap talk games. *Games and Economic Behavior*, 91(1):14–25.
- De Haan, T., Offerman, T., and Sloof, R. (2015). Money talks? an experimental investigation of cheap talk and burned money. *International Economic Review*, 56(4):1385–1426.
- De Mel, S., McIntosh, C., and Woodruff, C. (2013). Deposit collecting: Unbundling the role of frequency, salience, and habit formation in generating savings. *American Economic Review*, 103(3):387–92.
- Decker, J. F. (2004). The varying parameters of obstruction of justice in american criminal law. *Louisiana Law Review*, 65(1):49–130.
- DePaulo, B. M., Lindsay, J. J., Malone, B. E., Muhlenbruck, L., Charlton, K., and Cooper, H. (2003). Cues to deception. *Psychological Bulletin*, 129(1):74–118.
- Dickhaut, J. W., McCabe, K. A., and Mukherji, A. (1995). An experimental study of strategic information transmission. *Economic Theory*, 6(3):389–403.
- Dieckmann, A., Grimm, V., Unfried, M., Utikal, V., and Valmasoni, L. (2016). On trust in honesty and volunteering among europeans: Cross-country evidence on perceptions and behavior. *European Economic Review*, 90(1):225–253.
- Dreber, A., Ellingsen, T., Johannesson, M., and Rand, D. G. (2013). Do people care about social context? Framing effects in dictator games. *Experimental Economics*, 16:349–371.
- Dungan, J. A., Young, L., and Waytz, A. (2019). The power of moral concerns in predicting whistleblowing decisions. *Journal of Experimental Social Psychology*, 85:103848.

-
- Dyck, A., Morse, A., and Zingales, L. (2023). How pervasive is corporate fraud? *Review of Accounting Studies*, pages 1–34.
- Dziuda, W. and Salas, C. (2018). Communication with detectable deceit. Working paper, SSRN.
- Eckel, C. C. and Grossman, P. J. (2002). Sex differences and statistical stereotyping in attitudes toward financial risk. *Evolution and human behavior*, 23(4):281–295.
- Ekman, P., Friesen, W. V., and O’sullivan, M. (1988). Smiles when lying. *Journal of Personality and Social Psychology*, 54(3):414–420.
- Fagan, R. W. (1981). Public support for the courts: An examination of alternative explanations. *Journal of Criminal Justice*, 9(6):403–417.
- Falk, A. and Szech, N. (2013). Morals and markets. *Science*, 340(6133):707–711.
- Farrell, J. (1993). Meaning and credibility in cheap-talk games. *Games and Economic Behavior*, 5(4):514–531.
- Fatas, E., Heap, S. P. H., and Arjona, D. R. (2018). Preference conformism: An experiment. *European Economic Review*, 105:71–82.
- Feltovich, N. and Hamaguchi, Y. (2018). The effect of whistle-blowing incentives on collusion: An experimental study of leniency programs. *Southern Economic Journal*, 84(4):1024–1049.
- Fleetwood, S. (2019). A definition of habit for socio-economics. *Review of social economy*, 79(2):1–35.
- Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic perspectives*, 19(4):25–42.
- Fudenberg, D., Levine, D. K., and Maniadis, Z. (2012). On the robustness of anchoring effects in wtp and wta experiments. *American Economic Journal: Microeconomics*, 4(2):131–45.
- Fujiwara, T., Meng, K., and Vogl, T. (2016). Habit formation in voting: Evidence from rainy elections. *American Economic Journal: Applied Economics*, 8(4):160–88.
- Gerber, A. S., Green, D. P., and Shachar, R. (2003). Voting may be habit-forming: evidence from a randomized field experiment. *American journal of political science*, 47(3):540–550.
- Gibson, R., Tanner, C., and Wagner, A. F. (2013). Preferences for truthfulness: Heterogeneity among and within individuals. *American Economic Review*, 103(1):532–48.

-
- Gill, D., Prowse, V., and Vlassopoulos, M. (2013). Cheating in the workplace: An experimental study of the impact of bonuses and productivity. *Journal of Economic Behavior & Organization*, 96:120–134.
- Givati, Y. (2016). A theory of whistleblower rewards. *The Journal of Legal Studies*, 45(1):43–72.
- Glaeser, E. L., Laibson, D. I., Scheinkman, J. A., and Soutter, C. L. (2000). Measuring trust. *The quarterly journal of economics*, 115(3):811–846.
- Glazer, J. and Rubinstein, A. (2012). A model of persuasion with boundedly rational agents. *Journal of Political Economy*, 120(6):1057–1082.
- Glazer, J. and Rubinstein, A. (2014). Complex questionnaires. *Econometrica*, 82(4):1529–1541.
- Gneezy, U. (2005). Deception: The role of consequences. *American Economic Review*, 95(1):384–394.
- Goerg, S. J., Rand, D., and Walkowitz, G. (2020). Framing effects in the prisoner's dilemma but not in the dictator game. *Journal of the Economic Science Association*, 6:1–12.
- Grossman, G. M. and Katz, M. L. (1983). Plea bargaining and social welfare. *The American Economic Review*, 73(4):749–757.
- Hamaguchi, Y., Kawagoe, T., and Shibata, A. (2009). Group size effects on cartel formation and the enforcement power of leniency programs. *International Journal of Industrial Organization*, 27(2):145–165.
- Harvey, A. C., Vrij, A., Nahari, G., and Ludwig, K. (2017). Applying the verifiability approach to insurance claims settings: Exploring the effect of the information protocol. *Legal and Criminological Psychology*, 22(1):47–59.
- Havranek, T., Rusnak, M., and Sokolova, A. (2017). Habit formation in consumption: A meta-analysis. *European Economic Review*, 95(1):142–167.
- Herz, H. and Taubinsky, D. (2018). What makes a price fair? an experimental study of transaction experience and endogenous fairness views. *Journal of the European Economic Association*, 16(2):316–352.
- Herzog, S. (2003). The relationship between public perceptions of crime seriousness and support for plea-bargaining practices in israel: A factorial survey approach. *Journal of Criminal Law and Criminology*, 94(1):103–132.

-
- Heyes, A. and Kapur, S. (2009). An economic model of whistle-blower policy. *The Journal of Law, Economics, & Organization*, 25(1):157–182.
- Hinloopen, J. and Normann, H.-T. (2009). *Experiments and competition policy*. Cambridge University Press.
- Hinloopen, J., Onderstal, S., and Soetevent, A. (2023). Corporate leniency programs for antitrust: Past, present, and future. *Review of Industrial Organization*, 63(2):111–122.
- Hinloopen, J. and Soetevent, A. R. (2008). Laboratory evidence on the effectiveness of corporate leniency programs. *The RAND Journal of Economics*, 39(2):607–616.
- Holm, H. (2010). Truth and lie detection in bluffing. *Journal of Economic Behavior and Organization*, 76(1):318–324.
- Holm, H. J. and Kawagoe, T. (2010). Face-to-face lying—an experimental study in sweden and japan. *Journal of Economic Psychology*, 31(3):310–321.
- Huber, C., Dreber, A., Huber, J., et al. (2023). Competition and moral behavior: A meta-analysis of forty-five crowd-sourced experimental designs. *Proceedings of the National Academy of Sciences*, 120(23):1–10.
- Hugh-Jones, D. (2016). Honesty, beliefs about honesty, and economic growth in 15 countries. *Journal of Economic Behavior & Organization*, 127(1):99–114.
- Hurkens, S. and Kartik, N. (2009). Would i lie to you? on social preferences and lying aversion. *Experimental Economics*, 12(2):180–192.
- Ifcher, J. and Zarghamee, H. (2020). Behavioral economic phenomena in decision-making for others. *Journal of Economic Psychology*, 77(1):102–180.
- Ioannidis, K. (2020). Overcommunication: The role of past experience. Preregistration, AEA RCT Registry. <https://www.socialscienceregistry.org/trials/6387>.
- Ioannidis, K. (2022). Habitual communication. Discussion paper No.2022-016/I, Tinbergen Institute.
- Ioannidis, K. (2023). Whistleblowing and competition. Preregistration, American Economic Association’s registry for randomized controlled trials. <https://www.socialscienceregistry.org/trials/11051>.
- Ioannidis, K., Offerman, T., and Sloof, R. (2018). Anchoring bias in markets. Preregistration, AEA RCT Registry. <https://www.socialscienceregistry.org/trials/3402>.

-
- Ioannidis, K., Offerman, T., and Sloof, R. (2020). On the effect of anchoring on valuations when the anchor is transparently uninformative. *Journal of the Economic Science Association*, 6(1):77–94.
- Ioannidis, K., Offerman, T., and Sloof, R. (2022). Lie detection: A strategic analysis of the verifiability approach. *American Law and Economics Review*, 24(2):659–705.
- Isoni, A., Brooks, P., Loomes, G., and Sugden, R. (2016). Do markets reveal preferences or shape them? *Journal of Economic Behavior & Organization*, 122(1):1–16.
- Ispano, A. and Vida, P. (2021). Designing interrogations. *Working Paper*.
- Jehiel, P. (2005). Analogy-based expectation equilibrium. *Journal of Economic theory*, 123(2):81–104.
- Jehiel, P. (2021). Communication with forgetful liars. *Theoretical Economics*, 16(2):605–638.
- Jiang, Y. V. and Sisk, C. A. (2019). Habit-like attention. *Current opinion in psychology*, 29:65–70.
- Johnson, T. (2019). Public perceptions of plea bargaining. *American Journal of Criminal Law*, 46(1):133–156.
- Jung, M. H., Perfecto, H., and Nelson, L. D. (2016). Anchoring in payment: Evaluating a judgmental heuristic in field experimental settings. *Journal of Marketing Research*, 53(3):354–368.
- Jupe, L. M., Leal, S., Vrij, A., and Nahari, G. (2017). Applying the verifiability approach in an international airport setting. *Psychology, Crime & Law*, 23(8):812–825.
- Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.
- Kartik, N. (2009). Strategic communication with lying costs. *The Review of Economic Studies*, 76(4):1359–1395.
- Kassin, S. M. and Norwick, R. J. (2004). Why people waive their miranda rights: The power of innocence. *Law and Human Behavior*, 28(2):211–221.
- Kawagoe, T. and Takizawa, H. (2009). Equilibrium refinement vs. level-k analysis: An experimental study of cheap-talk games with private information. *Games and Economic Behavior*, 66(1):238–255.
- Kim, J.-Y. (2010). Credible plea bargaining. *European Journal of Law and Economics*, 29(3):279–293.

-
- Kleinberg, B., Nahari, G., and Verschuere, B. (2016). Using the verifiability of details as a test of deception: A conceptual framework for the automation of the verifiability approach. In Fornaciari, T., Fitzpatrick, E., and Bachenko, J., editors, *Proceedings of the 2nd Workshop on Computational Approaches to Deception Detection*, pages 18–25, San Diego, California. Association for Computational Linguistics.
- Koçaş, C. and Demir, K. (2014). An empirical investigation of consumers' willingness-to-pay and the demand function: The cumulative effect of individual differences in anchored willingness-to-pay responses. *Marketing Letters*, 25(2):139–152.
- Krupka, E. L. and Weber, R. A. (2013). Identifying social norms using coordination games: Why does dictator game sharing vary? *Journal of the European Economic Association*, 11(3):495–524.
- Lally, P., Van Jaarsveld, C. H., Potts, H. W., and Wardle, J. (2010). How are habits formed: Modelling habit formation in the real world. *European journal of social psychology*, 40(6):998–1009.
- Leder-Luis, J. (2023). Can whistleblowers root out public expenditure fraud? Evidence from medicare. *The Review of Economics and Statistics*, pages 1–49.
- Leib, M. (2021). *(Dis) honesty in individual and collaborative settings: A behavioral ethics approach*. PhD thesis, University of Amsterdam.
- Leshem, S. (2010). The benefits of a right to silence for the innocent. *The RAND Journal of Economics*, 41(2):398–416.
- Li, J., Yin, X., Li, D., Liu, X., Wang, G., and Qu, L. (2017). Controlling the anchoring effect through transcranial direct current stimulation (tdcs) to the right dorsolateral prefrontal cortex. *Frontiers in psychology*, 8(1):1–9.
- Ma, Q., Li, D., Shen, Q., and Qiu, W. (2015). Anchors as semantic primes in value construction: an eeg study of the anchoring effect. *PloS one*, 10(10):1–18.
- Maniadis, Z., Tufano, F., and List, J. A. (2014). One swallow doesn't make a summer: New evidence on anchoring effects. *American Economic Review*, 104(1):277–90.
- Matthews, S. A., Okuno-Fujiwara, M., and Postlewaite, A. (1991). Refining cheap-talk equilibria. *Journal of Economic Theory*, 55(2):247–273.
- Mazar, A. and Wood, W. (2018). Defining habit in psychology. In *The psychology of habit*. Springer.
- Mechtenberg, L., Muehlheusser, G., and Roeder, A. (2020). Whistleblower protection: Theory and experimental evidence. *European Economic Review*, 126(1):103447.

-
- Meredith, M. et al. (2009). Persistence in political participation. *Quarterly Journal of Political Science*, 4(3):187–209.
- Mesmer-Magnus, J. R. and Viswesvaran, C. (2005). Whistleblowing in organizations: An examination of correlates of whistleblowing intentions, actions, and retaliation. *Journal of Business Ethics*, 62(3):277–297.
- Mialon, H. M. (2008). An economic theory of the fifth amendment. In *Economics, Law and Individual Rights*, pages 264–286. Routledge.
- Miranda v. Arizona* (1966). 384 U.S. 436-545.
- Murphy, R. O., Ackermann, K. A., and Handgraaf, M. J. (2011). Measuring social value orientation. *Judgment and Decision Making*, 6(8):771–781.
- Nahari, G., Vrij, A., and Fisher, R. P. (2014a). Exploiting liars' verbal strategies by examining the verifiability of details. *Legal and Criminological Psychology*, 19(2):227–239.
- Nahari, G., Vrij, A., and Fisher, R. P. (2014b). The verifiability approach: Countermeasures facilitate its ability to discriminate between truths and lies. *Applied Cognitive Psychology*, 28(1):122–128.
- O'Connor, C. (2014). The evolution of vagueness. *Erkenntnis*, 79(4):707–727.
- Pascual-Ezama, D., Fosgaard, T. R., Cardenas, J. C., Kujal, P., Veszteg, R., de Liaño, B. G.-G., Gunia, B., Weichselbaumer, D., Hilken, K., Antinyan, A., et al. (2015). Context-dependent cheating: Experimental evidence from 16 countries. *Journal of Economic Behavior & Organization*, 116(1):379–386.
- Peysakhovich, A. and Rand, D. G. (2016). Habits of virtue: Creating norms of cooperation and defection in the laboratory. *Management Science*, 62(3):631–647.
- Reinganum, J. F. (1988). Plea bargaining and prosecutorial discretion. *The American Economic Review*, 78(4):713–728.
- Robert, I. and Arnab, M. (2013). Is dishonesty contagious? *Economic Inquiry*, 51(1):722–734.
- Romero, J. (2015). The effect of hysteresis on equilibrium selection in coordination games. *Journal of Economic Behavior & Organization*, 111(1):88–105.
- Rosenbaum, S. M., Billinger, S., and Stieglitz, N. (2014). Let's be honest: A review of experimental evidence of honesty and truth-telling. *Journal of Economic Psychology*, 45:181–196.

-
- Royer, H., Stehr, M., and Sydnor, J. (2015). Incentives, commitments, and habit formation in exercise: evidence from a field experiment with workers at a fortune-500 company. *American Economic Journal: Applied Economics*, 7(3):51–84.
- Rozen, K. (2010). Foundations of intrinsic habit formation. *Econometrica*, 78(4):1341–1373.
- Samuelson, L. (2001). Analogies, adaptation, and anomalies. *Journal of Economic Theory*, 97(2):320–366.
- Sánchez-Pagés, S. and Vorsatz, M. (2007). An experimental study of truth-telling in a sender–receiver game. *Games and Economic Behavior*, 61(1):86–112.
- Savikhin, A. C. and Sheremeta, R. M. (2013). Simultaneous decision-making in competitive and cooperative environments. *Economic Inquiry*, 51(2):1311–1323.
- Schaner, S. (2018). The persistent power of behavioral change: Long-run impacts of temporary savings subsidies for the poor. *American Economic Journal: Applied Economics*, 10(3):67–100.
- Schmolke, K. U. and Utikal, V. (2018). Whistleblowing: Incentives and situational determinants. Available at SSRN 3198104.
- Schurr, A. and Ritov, I. (2016). Winning a competition predicts dishonest behavior. *Proceedings of the National Academy of Sciences*, 113(7):1754–1759.
- Schwieren, C. and Weichselbaumer, D. (2010). Does competition enhance performance or cheating? A laboratory experiment. *Journal of Economic Psychology*, 31(3):241–253.
- Seidmann, D. J. (2005). The effects of a right to silence. *The Review of Economic Studies*, 72(2):593–614.
- Seidmann, D. J. and Stein, A. (2000). The right to silence helps the innocent: a game-theoretic analysis of the fifth amendment privilege. *Harvard Law Review*, pages 430–510.
- Serota, K. B., Levine, T. R., and Boster, F. J. (2010). The prevalence of lying in america: Three studies of self-reported lies. *Human Communication Research*, 36(1):2–25.
- Serra-Garcia, M. and Gneezy, U. (2021). Mistakes, overconfidence, and the effect of sharing on detecting lies. *American Economic Review*, 111(10):3160–83.
- Serra-Garcia, M., Van Damme, E., and Potters, J. (2011). Hiding an inconvenient truth: Lies and vagueness. *Games and Economic Behavior*, 73(1):244–261.

-
- Shah, A. K., Shafir, E., and Mullainathan, S. (2015). Scarcity frames value. *Psychological Science*, 26(4):402–412.
- Shenhav, A., Rand, D. G., and Greene, J. D. (2012). Divine intuition: Cognitive style influences belief in god. *Journal of Experimental Psychology: General*, 141(3):423.
- Simonson, I. and Drolet, A. (2004). Anchoring effects on consumers' willingness-to-pay and willingness-to-accept. *Journal of consumer research*, 31(3):681–690.
- Sims, C. A. (2003). Implications of rational inattention. *Journal of monetary Economics*, 50(3):665–690.
- Sobel, J. (2020). Lying and deception in games. *Journal of Political Economy*, 128(3):907–947.
- Sonnemans, J., Schram, A., and Offerman, T. (1998). Public good provision and public bad prevention: The effect of framing. *Journal of Economic Behavior & Organization*, 34(1):143–161.
- Sorochinski, M., Hartwig, M., Osborne, J., Wilkins, E., Marsh, J., Kazakov, D., and Granhag, P. A. (2014). Interviewing to detect deception: when to disclose the evidence. *Journal of Police and Criminal Psychology*, 29(2):87–94.
- Stigler, G. J. and Becker, G. S. (1977). De gustibus non est disputandum. *The American Economic Review*, 67(2):76–90.
- Suchotzki, K., Verschuere, B., Van Bockstaele, B., Ben-Shakhar, G., and Crombez, G. (2017). Lying takes time: A meta-analysis on reaction time measures of deception. *Psychological Bulletin*, 143(4):428.
- Sugden, R., Zheng, J., and Zizzo, D. J. (2013). Not all anchors are created equal. *Journal of Economic Psychology*, 39(1):21–31.
- Toplak, M. E., West, R. F., and Stanovich, K. E. (2014). Assessing miserly information processing: An expansion of the cognitive reflection test. *Thinking & reasoning*, 20(2):147–168.
- Tsur, Y. (2017). Bounding reasonable doubt: implications for plea bargaining. *European Journal of Law and Economics*, 44(2):197–216.
- Tufano, F. (2010). Are 'true' preferences revealed in repeated markets? an experimental demonstration of context-dependent valuations. *Experimental Economics*, 13(1):1–13.
- Tversky, A. and Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157):1124–1131.

-
- UK Sentencing Council (2017). *Reduction in sentence for a guilty plea*. United Kingdom Department of Justice.
- US Bureau of Justice Statistics (2003). *Sourcebook of criminal justice statistics*. United States Department of Justice. <https://www.ncjrs.gov/pdffiles1/Digitization/208756NCJRS.pdf>.
- US Sentencing Commission (2018). *Guidelines Manual*. United States Department of Justice.
- Vadera, A. K. and Pathki, C. S. (2021). Competition and cheating: Investigating the role of moral awareness, moral identity, and moral elevation. *Journal of Organizational Behavior*, 42(8):1060–1081.
- Verplanken, B. (2006). Beyond frequency: Habit as mental construct. *British Journal of Social Psychology*, 45(3):639–656.
- Verschuere, B. and Shalvi, S. (2014). The truth comes naturally! does it? *Journal of Language and Social Psychology*, 33(4):417–423.
- Vrij, A. (2008). *Detecting lies and deceit: Pitfalls and opportunities*. John Wiley & Sons, Chichester, United Kingdom.
- Vrij, A. (2018). Verbal lie detection tools from an applied perspective. In *Detecting concealed information and deception*, pages 297–327. Elsevier, London, United Kingdom.
- Vrij, A. (2019). Deception and truth detection when analyzing nonverbal and verbal cues. *Applied Cognitive Psychology*, 33(2):160–167.
- Vrij, A., Fisher, R. P., and Blank, H. (2017). A cognitive approach to lie detection: A meta-analysis. *Legal and Criminological Psychology*, 22(1):1–21.
- Vrij, A., Mann, S., Kristen, S., and Fisher, R. P. (2007). Cues to deception and ability to detect lies as a function of police interview styles. *Law and Human Behavior*, 31(5):499–518.
- Wakker, P. P. (2010). *Prospect theory: For risk and ambiguity*. Cambridge university press.
- Wallmeier, N. (2019). The hidden costs of whistleblower protection. Working paper, SSRN.
- Wang, J. T.-Y., Spezio, M., and Camerer, C. F. (2010). Pinocchio's pupil: using eye-tracking and pupil dilation to understand truth telling and deception in sender-receiver games. *American Economic Review*, 100(3):984–1007.

-
- Waytz, A., Dungan, J., and Young, L. (2013). The whistleblower's dilemma and the fairness–loyalty tradeoff. *Journal of Experimental Social Psychology*, 49(6):1027–1033.
- Welsh, M., Burns, N., and Delfabbro, P. (2013). The cognitive reflection test: How much more than numerical ability? In *Proceedings of the 35th Annual Meeting of the Cognitive Science Society*.
- Wood, W., Quinn, J. M., and Kashy, D. A. (2002). Habits in everyday life: Thought, emotion, and action. *Journal of personality and social psychology*, 83(6):1281.
- Wood, W. and Rünger, D. (2016). Psychology of habit. *Annual review of psychology*, 67(1):289–314.
- Yoon, S. and Fong, N. (2019). Uninformative anchors have persistent effects on valuation judgments. *Journal of Consumer Psychology*, 29(3):391–410.
- Yoon, S., Fong, N. M., and Dimoka, A. (2019). The robustness of anchoring effects on preferential judgments. *Judgment and Decision Making*, 14(4):470–487.
- Ziegler, A., Romagnoli, G., and Offerman, T. (2020). Morals in multi-unit markets. Discussion Paper No.2020-072/1, Tinbergen Institute.
- Zuckerman, M., DePaulo, B. M., and Rosenthal, R. (1981). Verbal and nonverbal communication of deception. *Advances in Experimental Social Psychology*, 14(1):1–59.

List of co-authors and contributions

Chapter 1 and Chapter 3 of this dissertation were co-authored, whereas Chapters 2 and Chapter 4 were single-authored. Below is a reference list including a list of co-authors for each Chapter and the contribution of each co-author.

Chapter 1 - Verifiability approach

The initial idea to provide a game theoretic analysis of the verifiability approach was provided by Theo Offerman. Konstantinos Ioannidis analysed the model and wrote the first draft for his Tinbergen Institute MPhil thesis. All authors revised the draft for submission to journals and eventual publication.


Chapter 3 - Anchoring and markets

The initial idea for the experiment was provided by Konstantinos Ioannidis. Konstantinos Ioannidis programmed and conducted the lab experiment, performed the initial analysis of the data, and wrote the first draft of the working paper. All authors collaborated on the design of the experiment and revised the draft for submission to journals and eventual publication.

The Tinbergen Institute is the Institute for Economic Research, which was founded in 1987 by the Faculties of Economics and Econometrics of the Erasmus University Rotterdam, University of Amsterdam, and Vrije Universiteit Amsterdam. The Institute is named after the late Professor Jan Tinbergen, Dutch Nobel Prize laureate in economics in 1969. The Tinbergen Institute is located in Amsterdam and Rotterdam. For a full list of PhD theses that appeared in the series we refer to List of PhD Theses – Tinbergen.nl. The following books recently appeared in the Tinbergen Institute Research Series:

799. Y. XIAO, *Fertility, parental investments and intergenerational mobility*
800. X. YU, *Decision Making under Different Circumstances: Uncertainty, Urgency, and Health Threat*
801. G. GIANLUCA, *Productivity and Strategies of Multiproduct Firms*
802. H. KWEON, *Biological Perspective of Socioeconomic Inequality*
803. D.K. DIMITROV, *Three Essays on the Optimal Allocation of Risk with Illiquidity, Intergenerational Sharing and Systemic Institutions*
804. J.B. BLOOMFIELD, *Essays on Early Childhood Interventions*
805. S. YU, *Trading and Clearing in Fast-Paced Markets*
806. M.G. GREGORI, *Advanced Measurement and Sampling for Marketing Research*
807. O.C. SOONS, *The Past, Present, and Future of the Euro Area*
808. D. GARCES URZAINQUI, *The Distribution of Development. Essays on Economic Mobility, Inequality and Social Change*
809. A.C. PEKER, *Guess What I Think: Essays on the Wisdom in Meta-predictions*
810. A. AKDENIZ, *On the Origins of Human Sociality*
811. K. BRÜTT, *Strategic interaction and social information: Essays in behavioural economics*
812. P. N. KUSUMAWARDHANI, *Learning Trends and Supply-side Education Policies in Indonesia*
813. F. CAPOZZA, *Essays of the behavioral economics of social inequalities*
814. D. MUSLIMOVA, *Complementarities in human capital production: The Importance of Gene-Environment Interactions*
815. J. DE JONG, *Coordination in market and bank run experiments*
816. Y. KIM, *Micro studies of macroprudential policies using loan-level data*
817. S.R. TER MEULEN, *Grade retention, ability tracking, and selection in education*
818. A. ZIEGLER, *The Strategic Role of Information in Markets and Games: Essays in Behavioral Economics*
819. I. VAN DE WERVE, *Panel data model for socioeconomic studies in crime and education*
820. Y. GU, *Roads, Institutions and the Primary Sector in West Africa*

-
821. A. Y. LI, *Share repurchases in the US: An extensive study on the data, drivers, and consequences*
822. R. DIAS PEREIRA, *What Makes us Unique? Genetic and Environmental Drivers of Health and Education Inequalities*
823. H. P. LETTERIE, *Essays on the regulation of long-term care in the Netherlands*
824. D. PACE, *Essays on the cognitive foundations of human behavior and on the behavioral economics of climate change*
825. J. N. VAN BRUMMELEN, *On the estimation of parameters in observation-driven time series models*
826. Z. CSAFORDI, *Essays on industry relatedness: Productivity spillovers through labor flows, diversification and agglomeration economies*
827. B. VAN OS, *On Dynamic Models: Optimization-Based Methods and Practical Applications*
828. D. Ó CEALLAIGH, *Self-control failures and physical inactivity: Measuring, understanding and intervening*
829. S. B. DONOVAN, *Ties that bind and fray: Agglomeration economies and location choice*
830. A. SOEBHAG, *Essays in Empirical Asset Pricing*
831. H. YUAN, *Essays in behavioral economics*
832. A. A. LENGYEL, *Essays on government bond markets and macroeconomic stabilization*
833. S. KÜTÜK, *Essays on Risk Creation in the Banking Sector*
834. E. VLADIMIROV, *Essays on the econometrics of option pricing*
835. R. E. C. PRUDON, *From the onset of illness to potential recovery: Empirical economic analysis of health, disability and work*
836. K. MOUSSA, *Signal Extraction by the Extremum Monte Carlo Method*
837. D. FAVOINO, *The Adaptation of Firms to Institutional Change*
838. B. WACHE, *On the estimation of parameters in observation-driven time series models*
839. A. FEHÉR, *Essays in law and economics*
840. Q. WIERSMA, *Dynamic models for multi-dimensional time series*
841. R. SILVESTRINI, *On the importance of firm heterogeneity, business dynamism, and market power dynamics in the macroeconomy*
842. E.S.R. DIJK, *Innovative start-ups and competition policy – how to reign in big*
843. T.D. DCHENK, *Essays in causal inference with panel data*
844. S. TYROS, *Workers' skills and (green) technology adoption*
845. C.J. GRASER, *Mechanisms for the evolution of prosociality*



This dissertation comprises four independent chapters, each exploring the impact of information on decision-making processes within economic contexts. The overarching theme across these chapters is the role of information in shaping human behaviour and economic outcomes.

The first two chapters investigate strategic interactions involving a sender possessing information desired by a receiver. The first chapter employs game theory to analyse these interactions, while the second chapter adopts an experimental approach to delve into how habits affect strategic communication. The third chapter investigates the influence of irrelevant information on individuals' valuation of goods, exploring whether market interactions mitigate this effect. The fourth chapter studies the decision-making process surrounding the disclosure of information pertaining to corporate fraud through whistleblowing mechanisms.

Through these diverse topics, this dissertation contributes to the broader understanding of how information intersects with economic decision-making processes.

Konstantinos Ioannidis holds a BSc in Mathematics from the Aristotle University of Thessaloniki (2010), a BSc in Statistics from the University of the Aegean (2013), a MSc in Mathematical Modelling in Natural Sciences and New Technologies from the University of the Aegean (2013), a MSc in Statistics and Actuarial-Financial Mathematics from the University of the Aegean (2015), and an MPhil in Economics from the Tinbergen Institute (2017). In 2017, he joined CREED at the Amsterdam School of Economics at the University of Amsterdam as a PhD student, supervised by Theo Offerman and Randolph Sloof. Konstantinos currently works as a postdoctoral researcher at the University of Cambridge funded by the Leverhulme International Professorship in Neuroeconomics awarded to Peter Bossaerts.

