Do the Middle Shout Loudest? Signs and (Counter)Signals of Trustworthiness and Toughness

Áron Székely

St Cross College
University of Oxford

Supervised by Professor Diego Gambetta

Thesis submitted in partial fulfilment of the requirements for the degree of
Doctor of Philosophy

Michaelmas 2014

Acknowledgements

Many people have helped me along the way to completing this endeavour. Their support and assistance, in ways big and small, has been invaluable. In particular, I would like to thank the following:

- My supervisor, Diego Gambetta. His intellectual creativity and originality, rigorous critiques, and generous support have been essential for the completion of this thesis. It has been a privilege to work together.
- The people at the Bologna Laboratory for Experiments in Social Sciences, Marco Casari, Riccardo Ghidoni, and Gianadrea Lanzara, and at the Centre for Experimental Social Sciences, Oxford, who helped to make my experiments possible.
- The administrators at the department and college who have been such help throughout the years.
- Colleagues, and friends, who have contributed in a variety of ways to improving this thesis, and specifically Wojtek Przepiorka, Dominic Burbidge, Daniel Honig, and Dingeman Wiertz.
- The Economic and Social Research Council for funding my doctoral research.
- Giorgia Gon, who has been a source of constant aid and an ever-compliant listener and critique throughout this work. She has heard 'countersignalling' more times than anyone ought to. Thank you.
- Finally, my family, who have been a source of unending support.

To my family.

Abstract

What do we do when we realise that others with whom we interact already know something about relevant qualities of ours? Do we relying on the information that we naturally emit to get our message across, or do we take actions to try to change the situation in our favour? And, does the information that we emit, generated intentionally or unintentionally, allow us to cooperate with others and conflict to be resolved peacefully? Drawing on signalling and 'countersignalling' theories, and a recent behavioural theory, I explore these questions using two experiments in which the relevant qualities are trustworthiness and toughness, and observational data from the 1997 Survey of Inmates in State and Federal Correctional Facilities concerning violence among prisoners. In the experiments, subjects, who do not know what will follow, initially create a natural indicator of their qualities, and subsequently, have the opportunity to send another indicator, this time fully informed. In one experiment, consistent with countersignalling theory, subjects whose generosity is clear deign to send further information, while those whose generosity can be called under question choose to update their initial action the most. In the other, the toughest subjects put the most effort into getting their message across. Consistent among both, senders' actions correspond to receivers' evaluations. Finally, I find that pertinent information can allow conflicts to be resolved successfully; it is linked with less violence in prison and in a laboratory-based contest.

Contents

1. Introduction

We all interact with others in situations in which their interests and ours, potentially, or actually, conflict. Neighbours consider whether to lend money, or other resources, to one another and face the danger that their resources are not returned; prisoners involved in competition over privacy, space, money, and cigarettes, to name a few, lose their access to them if they yield to others' demands but risk a costly fight if they resist. Communication about relevant qualities that the people involved possess—trustworthiness for borrowers and toughness for prisoners—can allow cooperation to occur and costly fighting to be avoided.

When, as is often the case, some *prior information* about those relevant qualities exists, what strategies do people use to communicate? Will a new neighbour preceded by a reputation for insincerity take obsequious displays to assure others of his trustworthiness? The effect that a particular kind of pre-existing information, which gives perceivers an inkling about emitters' qualities, has on communicative behaviour is a focus of this thesis.

Other times people lack prior information, or it is prevented from circulating, and have few opportunities to take communicative displays of their relevant qualities. How does this affect peoples' behaviour in situations of conflict? Will it make the breakdown of non-violent communication more likely, jeopardising the stability of the social order? Whether information about peoples' ability to use force prevents or promotes violence from erupting is the other main topic addressed.

Three chapters address these topics empirically. In the first, 'Signs and (Counter)Signals of Trustworthiness', I examine whether, and how, people use intentional messages about their trustworthiness to change the behaviour of others when they realise that relevant information that they unintentionally created in the past is also present in the interaction; and, conversely, how people perceiving the messages respond to them. In the second, 'Information and Violence in Prison', I turn my attention to prisoners serving time in US state and federal prisons and investigate how the information that they have about one another's ability to use violence changes the amount of fighting that occurs. In the third, 'Communication and Information in Interpersonal Conflict', I deal with both questions experimentally: how people use, and others respond to, intentionally and unintentionally created messages of toughness, and the effect that this information has on the amount of laboratory-simulated fighting.

Two chapters are based on data created using laboratory experiments. In such experiments, subjects—who are typically but not necessarily students—are called into the 'lab', a room containing a set of networked computers arranged in portioned cubicles that prevent the subjects from observing one another's decisions, to participate in an interactive experiment. Bringing subjects into this environment allows the researcher to tightly control many aspects of a subject's experience. General experimental methodology—randomly allocating subjects to treatment and control and exogenously manipulating independent variables—allows the 'fundamental problem of causal inference' to be overcome (Holland, 1986). But, running experiments in a laboratory setting provides further benefits in that the experimenter can precisely control the range that the treatment independent variable takes, and auxiliary variables—for instance whether the subjects are paying attention or not—that can allow a treatment effect to

emerge, or overshadow it, can be monitored and controlled. Internal validity is their major strength.

External validity is where laboratory experiments face difficulty. It is often argued that a treatment effect (or the lack of a treatment effect) observed in a lab experiment using a specific student sample only vaguely applies to real-world behaviour among populations composed on non-students. Part of this problem is the artificiality of the laboratory setting that can elicit unnatural behaviour not found elsewhere. One technique I employ to assuage this concern is to really use the quality that people are communicating about or interpreting. When I look at communicating trustworthiness, the message is based on *real* generosity—not an induced preference for generosity based on external manipulation of payoffs; when I look at toughness too, the message is based on real toughness. For the experiment on interpersonal conflict, subjects carry out a physical, competitive, and mildly pain-inducing exercise. Others are to mildly contextualise the experiments and employing intuitive, and engaging experimental interfaces. These factors may enable subjects to grasp the situation, and problem, at hand more easily and clearly, and tackle it using the heuristics and mental tools that they normally would do, leading to natural behaviour.

One chapter relies on a nationally representative sample of observational data concerning prisoner behaviour. While losing some certainty regarding internal validity, here behaviour is recorded *in situ*, providing assurance about the relevance of the relationships between dependent and independent variables to real world behaviour. Due to the particularity, and extremity, of the conditions that people experience in prison, capturing these in ethically acceptable experiment is difficult. The strength of the experimental chapters lies in their certainty regarding causality, while the observational chapter

provides information about the relevance to external behaviour. Each contributes its own piece of the puzzle to build an overall picture.

Sociology has a rich heritage of research on communication and information. Erving Goffman (1956, 1971), Pierre Bourdieu (1984), George Herbert Mead (1934), Thorstein Veblen (1899), Marcel Mauss (1954) have made some of the most important and influential contributions. Yet, this literature does not directly address the fundamental issue of reliability in communication when there is an actual, or potential, divergence in the interests of the parties involved.[1] Instead, I look to signalling theory, a theory that is widely used in economics, biology, and political science to understand communicative behaviour, but little so-far used in sociology, a fascinating and empirically near-unexamined extension of signalling theory known as countersignalling theory, and a behavioural theory regarding the effect of information on violence.[2]

Signalling and countersignalling theory both explicitly tackle the issue of reliability and provide an explicit and internally coherent—ensured by the use of game theory—individual-level account of communicative behaviour. They provide a firm foundation upon which the behavioural theory regarding violence and information rests. The additional advantage is that both are middle-range theories that have clearly delineated set of components and phenomena that they can explain (Hedström & Udehn, 2009).

Using these theories, I examine the following questions:

[1] Erving Goffman (1956) mentions it in passing (e.g. pp. 18, 19) and Thorstein Veblen may be considered as having formulated a prototype *verbal* model of signalling theory in the *Theory of the Leisure Class*.
[2] There are theories from criminology that deal with violence in prisons, and I describe these in the appropriate chapter.

- Can signs created in experiments reliably indicate the qualities of senders and do receivers respond to them?
- Can signals created in experiments reliably indicate the qualities of senders and do receivers respond to them?
- How do senders behave when they can emit both signs and signals? Do the middle shout loudest?
- How do receivers behave when they perceive both signs and signals?
- Does information availability affect fighting?

Using countersignalling theory to guide my hypothesis, I address whether generosity in the lab, measured by giving in a dictator game, can function as a reliable sign and signal of trustworthiness, and examine how senders use these two types of information, and how receivers interpret them, in the initial empirical chapter (Chapter 2). This experiment focuses on countersignalling and provides an initial test of it in the context of prosociality. The following chapter (Chapter 3) deals with the final question—how information affects fighting—and examines whether information availability is associated with a decrease in violence in the highly competitive and conflict ridden context of prison using data from the 1997 Survey of Inmates in State and Federal Correctional Facilities, a survey on prisoners in the U.S. system. The final empirical chapter (Chapter 4) address all of the questions in one experiment. In it, I look at how the informative content of messages is changed and interpreted and how the availability of information affects behaviour. This chapter has the broadest theoretical basis and brings together the preceding two, testing signalling and countersignalling theory regarding the quality of toughness and experimentally testing information availability. It tests another quality and complements the observational data based chapter.

In the rest of the introduction, I describe the theoretical foundations that unify my thesis and argue for testing and using countersignalling theory. I begin by outlining a basic, but powerful, taxonomy of communication. Then I discuss the issue of reliability that potential communicators face, the solutions provided by signalling theory, and the conditions under which the issue can be solved. Next, I describe countersignalling theory, drawing on the taxonomy of communication already introduced. I then present informal observations of countersignalling-like behaviours and a potential countersignalling mimicry case, both of which demonstrate the promise of countersignalling theory. Finally, I conclude with some relevant related explanations, and the links that they have to countersignalling.

Parts of what I write below are included in different forms in the empirical chapters, each of which has its own specific theoretical section with its own focus. An in-depth description of countersignalling theory is provided in Chapter 2 and the conjecture regarding information availability is described in Chapter 3. Chapter 4 skirts over the theories for it employs those described in the chapters before it. Here, I describe signalling theory in depth, provide a brief description of countersignalling theory, and discuss useful, and important, aspects of them or relevant related issues that are not covered later.

Signalling theory was first formalised by the economist Michael Spence in his highly influential paper 'Job Market Signaling' (1973) and expanded upon in his subsequent book (1974).[3] These works are heavily used in economics. At a similar time, in biology, Amotz Zahavi (1975) developed a verbal model of reliable signalling, and subsequently,

[3] See Riley (2001) for a discussion of the context in which Spence developed his model.

Grafen (1990) provided a formalised model demonstrating how signal reliability can be maintained under natural selection.

Signalling theory was introduced to sociology by Bacharach and Gambetta (2001) and Posner (1998, 2002), and expanded upon and refined in later works by Gambetta (2005, 2009b).[4] While signalling models have existed in economics since the 1970s, a precise articulation of (i) why signalling theory is useful for sociology, (ii) a semantic structure applicable to (un)intentional choice-based communication in humans, and (iii) empirical demonstrations of how it can be fruitfully applied to substantive sociological issues such as trust (Gambetta & Hamill, 2005) were lacking until provided by Gambetta and co-authors. My discussion of signalling and signalling theory relies heavily on this body of work, and broadly follows the structure employed by Gambetta (2009b).

1.1 Perceivable features

The perceivable features that people have may be called *manifesta* (Bacharach & Gambetta, 2001, p. 155). Manifesta are near infinite, and include the physical features that people possess as well as the marks that their existence creates in the external world around them. Eye colour, height, and facial structure are easily perceivable and commonly used manifesta that are physical aspects of people. The clothing that one wears, the reputation possessed, and expressions of language—whether vocalisations, writing, or signing—are manifesta too, marks, some more ephemeral than others, created in the world external to the person.

[4] For a description of signalling theory's predecessors in sociology, who never attempted any formalisation, see Gambetta (2009b, pp. 186–189).

Out of some manifesta, useful information about peoples' *qualities* (sometimes called krypta), which are important but that cannot be directly observed, can be extracted (Bacharach & Gambetta, 2001, p. 155). Trustworthiness, honesty, and intelligence, are qualities that are important to many of us in everyday life, however, we cannot directly perceive peoples' trustworthiness, honesty, or intelligence, and, instead have to rely on their manifesta to make such an inference.[5]

1.1.1 Signs and signals

Manifesta come in two distinct kinds: *signs* and *signals*.[6, 7] A sign is simply anything in 'the environment that is perceptible and by being perceived happens to modify our beliefs about something or someone' (Gambetta, 2009b, p. 170). They are objects, behaviours, or marks that are created and displayed by people *without the intent* to change the beliefs of others. Yet, out of the signs, others can extract information, for signs are statistically linked to qualities. Smoke and footfalls are linked to presence, houses and cars to inhabitation, and the sound of gunfire to shooting. People emitting such information do so unintentionally and are often negatively affected by its perception, and aware of potentially harmful information contained in signs, some savvy would-be senders strategically avoid emitting them. Hunter S. Thompson describes how 'Except in posed or prearranged situations, most [Hells] Angels are leery of being photographed or recorded and even of talking to a man with a notebook. Tapes and film are regarded as

[5] Not all qualities are equally hidden. Strength, for instance, is near observable, in the form of muscles, while trustworthiness is a quality entirely hidden until a person reveals it through behaviour.

[6] I introduce only the most basic taxonomy of manifesta. See Bacharach and Gambetta (2001) for an extensive taxonomy based on decision making. For signalling theory based on evolution, the most comprehensive taxonomy is provided by Hurd and Enquist (2005),

[7] Whether manifesta are signs or signals is dynamic and they often transform from one into the other (Gambetta, 2009b, pp. 170, 171). Smoke is usually a sign of presence but is transformed into a signal by people who are stranded and are looking for rescue. Conversely, a Yakuza gangster who cuts off his little finger in the 'Yubitsume' ritual, signals his sincerity and determination today to atone for an error or debt. Initially carried out as a signal, in future interactions it becomes a sign of (potentially past) Yakuza membership and error.

especially dangerous because they can't be denied. This is true even in peaceful situations, where a casual photograph can place a man at the scene of a crime not yet committed' (Thompson, 1967, p. 185). Anticipating the danger in signs, Hells Angels simply avoid creating them.

Signals, in contrast, are manifesta that people take, create, or display, *with the intention* to change the beliefs that others hold about qualities of theirs (Gambetta, 2009b, p. 170). Intentionality in the mind of the sender is the distinguishing feature of signals.[8] Whether or not the signal actually correlates with a quality and receivers can use it to update their beliefs does not matter whether a display, or action, is a sign or a signal.

Many signals are actions and come in myriad, sometimes remarkable, forms. American Border Patrol Agents, investigating Mexico-U.S. cross-border drug smuggling tunnels, recall how, upon approaching the end of a tunnel emerging into Mexico, smoke was deliberately blown down into it for them to smell, as signal of presence and a warning for the agents to stay away—which they sensibly do (Higginbotham, 2012, p. 61). Signals may also be remnants of past actions; gang members frequently use tattoos as a signal of affiliation.

[8] Intentionality may be 'occurrent', and simultaneously present with the signal, or it may only be 'triggerable' (Bacharach & Gambetta, 2001, p. 174). The latter case refers to situations in which a signal continues to be taken and occurrent intention has passed—there is no intent to communicate in the mind of the sender—yet, the intention remains there, in a latent, 'submerged' form. Such that events are able to make the occurrent intention re-emerge. Consider signalling with clothing. Initially, there may be occurrent intent to communicate, but, after time, dressing becomes automatic, and the occurrent intention disappears. Still, 'it remains an intention, for all that, on this simple test were one to be made aware of the issue by, for example, losing a suitcase containing all one's clothes, one would choose to re-outfit oneself with similar manifest—for example, with new clothes in the same style—with the explicit intention of signaling' (Bacharach & Gambetta, 2001, p. 174).

1.1.2 Signal reliability

Although a vast range of manifesta can be signals, not all of them are equally effective at transferring information. When the interests of people who are interacting potentially, or actually, diverge or collide, 'reliability'—whether the signal has truth-value—because of mimicry, becomes an issue.[9] If there is a sufficient benefit to be gained, and they can afford to do so, then all people, irrespective of whether or not they possess a quality, will use a certain signal. Both people with a certain quality, the 'high types', and people without a quality, or only a lower level of it, the 'low types', take the same signal, making it impossible for observers to separate the two based on the signalling action alone. Dressing smartly is one such signal, that is correctly, not considered a reliable signal of trustworthiness by Belfast taxi drivers because both untrustworthy and trustworthy would-be clients can easily dress in a suit (Gambetta & Hamill, 2005, p. 82).[10] When the use of a signal is distributed equally among low and high types, it is completely uninformative. Some signals, in potentially conflicting situations, become completely unreliable while others retain some, or all, of their informative content. Whether a signal is reliable or not is the question answered by signalling theory.

1.2 Signalling theory

In its most basic form, signalling theory involves three agents.[11] One agent is the receiver, one agent is a low type sender, and the other is a high type sender. High and low refer to the level of the unobservable quality that senders possess and that the receiver would like

[9] An extreme example of mimicry among non-human animals occurs in male mourning cuttlefish (*Sepia plangon*) during courtship. Upon finding a receptive female in the presence of a rival male, the male cuttlefish places himself in-between the female and the rival and changes the side of his colouration facing the other male to mimic the patterning of a female cuttlefish. This fools the rival male into courting the female mimicking side, distracting him from the real female. In turn, the mimicking male displays male courting patterns on the side facing the real female. These males use mixed-sex mimicry and genuine signals to simultaneously distract a rival and to court a female (Brown, Garwood, & Williamson, 2012).

[10] This is not to say that clothing is ignored by taxi drivers; rather clothing is considered in more subtle ways that are difficult to mimic and likely to be more reliable (Gambetta & Hamill, 2005).

[11] See, for instance, the beer-quiche game of Cho and Kreps (1987, p. 183).

to know. Senders and receivers are assumed to know the distribution of types in the population, and that they all known that they know the distribution of types, and so on *ad infinitum*.[12]

The senders each choose a signal that receivers observe. In turn, the receivers observe the signals of the senders and, using their knowledge of the prior distribution of types in the population, update, or not, their beliefs about the type of each sender. Subsequently, the receiver responds to the sender.

Senders are, in the typical case, all motivated to be treated as high types by receivers. Both the violent and the meek wish to become part of a prison gang and receive the benefits of power and relative safety. In contrast, the receiver gains the most from treating each type appropriately: the high type as a high type and a low type as a low type. In contrast to the wishes of the would-be sender-recruit, the Aryan Brotherhood only want to allow the violent to join (e.g. Grann, 2004, p. 159). Allowing the meek to join undermines their reputation, status, and power, and ultimately the benefits that its members obtain. As implied by the example, receivers may be individuals or groups.[13]

Crucially, there must be *asymmetry* of information between the senders and the receiver, such that the senders know their type and not the receivers. If senders and the receivers know which sender is which type receivers have no incentive to updated their beliefs following the perception of a signal—they already know the type of the sender—and senders have no incentive to signal. An asymmetric distribution of information is a key requirement for signalling to emerge.

[12] Known as common knowledge (Aumann, 1976), this is a standard assumption in game theory.
[13] In Michael Spence's initial works (1973, 1974), for instance, the receivers are groups of firms who compete with each other for employees.

Sometimes, the terms model and mimic are used, primarily in the biological literature, to describe senders. 'Model' refers to a sender that another type sender would like to be confused with. A mimic is a sender that takes steps, by creating or displaying signals, to deceive receivers into mistakenly believing that he or she is a type different to his or her true one (Gambetta, 2005, pp. 223, 224). In the two-type sender signalling theory, the model is the high type, and a low type that employs the same signal as the high type is a mimic. Not all low types are mimics; only those that try to match the signs and signals of high types.

Whether or not a signal is reliable, is partially determined by the cost structure of the signalling situation. In addition to the distribution of types in the population, the reliability of a signal depends on the link between sender type and differential costs of the signal, or sender type and the differential benefits that they receive from taking a certain signal.

1.2.1 Reliability based on differential costs

Spence models a job market containing sender-employees and receiver-employers. The senders are high or low productivity types, and they decide how much to invest into education, the signal. The receiver-employers observe the senders' education level and hire them for a certain wage. Employers are motivated, by competition for employees with other firms to offer high types high wages—offering a wage that is below employees' expected productivity allows other firms to offer more—and by cost minimising to offer low types low wages.

To arrive at an outcome in which the sender types take different signals and consequently, receivers can perfectly identify the type of the sender based on their signal, Spence assumes a negative correlation between the type of a sender and the amount of resources that it costs to invest in a certain signalling level. He refers to this as 'a *prerequisite* for an observable, alterable characteristic to be a persistently informative signal' (Spence, 1973, p. 359, italics in original). The cost of signalling has to be sufficiently *differentially costly* among low and high type senders—differentially costly meaning that the cost of each additional unit cost to signalling, the marginal cost, should cost less for an individual who is higher type sender.[14] Consequently, taking a certain signal costs less for a person with a high level of a quality than it does for a person with a low level of the same quality.

Consider high and low productivity individuals. Each additional year spent in education costs more for the low type than it does for the high type, arising from the psychological cost of studying, the opportunity cost of missing other activities, and the greater number of hours that the low type needs to invest in studying. Let the cost of staying in education for an extra year, after compulsory schooling has finished, cost a low type 10 utils and 0 utils to a high type. If a potential employer is willing to pay 9 utils worth of money to hire the high type, and 0 to hire the low type, then it is not worthwhile for the low type to gain by staying in education an extra year. Even if the employer believes that all people who are able to obtain an extra year's education are high types, the most possible benefit that the low type can gain, it still costs the low type more to obtain the extra year education (10) than the maximum possible gains (9), keeping the signal of the extra year of education reliable.

[14] This is known as the 'single crossing condition' in economics because the indifference curves of high and low types cross only once.

A illustration of the differentially costly signalling principle is recounted by Robert Merton, who, lacking the tools, never explicitly casts the event in signalling terms, but, nevertheless, identifies the key elements (1968, pp. 163–165).[15] During the Second World War, Kate Smith, a radio broadcaster, the sender, carried out an 18-hour, near non-stop, radio campaign, the differentially costly signal, to convince sceptical New Yorkers, the receivers, to purchase war-bonds. Her appearances on the broadcast, and display of dedication, led to New Yorkers making 39 million dollars worth of bond pledges. The reason for this, in Merton's words,

> was not so much what Smith *said* as what she *did* which served to validate her sincerity. It was the presumed stress and strain of an eighteen-hour series of broadcasts, it was the deed not the word which furnished the indubitable proof. Listeners might question whether she were not unduly dramatizing herself, but they could not escape the incontrovertible evidence that she was devoting the entire day to the task… The currency of talk was accepted because it was back by the gold of conduct (p. 165, italics in original).

Speech, a typically unreliable signal, is transformed into a reliable one because of the strenuous effort that she incurs to promote bond purchasing that a less committed broadcaster would be unwilling to do.

Typically, initiation rituals present opportunities for the person being initiated to display qualities using differentially costly signals. The notorious white-only prison gang, the Aryan Brotherhood, requires would be members to 'make their bones' and kill another prisoner (Grann, 2004, p. 159). Boys seeking to be accepted as men in the Satere-Mawe, a tribe of Amazonian Indians, must display their toughness and commitment by wearing a glove filled with brutally painful poisonous ants for up to 30 minutes. To improve and

[15] For the original study and analysis see Merton, Fiske, & Curtis (1946).

increase their position within the tribe, they repeat the ritual multiple times (Haddad Junior, Cardoso, & Moraes, 2005, p. 235).

Conversely, an unreliable signalling act, that fails because it is not differentially costly, emerges from the 2012 Venezuelan presidential elections, where, in the run-up to the elections and following allegations of poor health, the late Hugo Chávez released an exercise video to signal his vigour and physical capability of running the country to the electorate. However, as many commentators recognised, the exercises in the video are near-costless and not differentially costly—walking and gentle stretching—proving nothing about his health and making it a useless signal.

In this version of the theory, a signal does not have to be costly to both types to ensure its reliability (Gambetta, 2009b, pp. 181, 182).[16] Signalling may be costless for a high type, or even directly beneficial, so long as signalling is costly enough for a low type. A productive student may find joy in studying, thinking, and research, and directly benefit from education. So long as education is costlier for unproductive students, it can serve as a reliable signal.

1.2.2 Reliability based on differential benefits

The other version of signalling theory ensures signal reliability through *differential benefits*. In this version of the theory, it is the benefits that different agents obtain that correlate, positively, with their type while the cost of signalling does not.

[16] For signalling theory based purely on differential benefits, the signal must have some costs to both low and high types.

Under this differential benefits assumption, a higher type sender can afford to spend more on signalling than a lower type because he or she obtains greater benefit from that signal. Were lower types to mimic the signal sent, by definition, the lower type will obtain fewer benefits, despite being treated in the same way by the receiver. Sufficiently differentially beneficial signals can ensure that high and low type senders emit different signals and, thus, the reliability of signalling.[17]

Consider the following event from Tolstoy's *War and Peace*. In present day Lithuania, near Kaunas, Napoleon orders a regiment of his Polish uhlans, cavalry, to cross the nearby river Neris at a suitable ford. Instead of complying, the colonel of the regiment begs 'to be allowed to swim across the river before the Emperor's eyes'. Upon permission, the colonel, along with hundreds of his mounted troops who join him, swim the river at its deepest part—where the water is cold and the current is so fast that around forty of his men and horses drown trying to get across and many turn back—to display their devotion to the observing Napoleon (Tolstoy, 1952, pp. 345, 346). Although the cost of this signal is substantial, for there is peril in crossing the river, the cost is equal for the cavalrymen irrespective of devotion, and it is not this that ensures the act's reliability as a signal. What ensures this signal's reliability is that the pleasure that truly devout senders feel, knowing that Napoleon, the receiver, is watching their dangerous river crossing, is much greater than the pleasure that the non-devout obtain. Recognising the reliable display of the colonel, Napoleon enrols him in the *Legion d'honneur* (p. 346).

One influential use of this version of signalling theory in the social sciences is by Eric Posner in *Law and Social Norms* (2002). He proposes that discount rates—which may arise

[17] See the Sir Philip Sidney Game (Smith, 1991) for a model that explicitly demonstrates the stability of reliable signalling based on differential benefits.

from multiple sources not only from preferences—are the crucial quality that senders signal about (p. 18, 19). Low discount rate individuals (the high types here) gain more from cooperative enterprises, valuing tomorrow's cooperation highly or expecting to participate in extended cooperation, than high discount rate individuals (low types) who value cooperation less or know that they will break-off cooperation early. The former obtain more benefits from being recognised as cooperative types, and, as a result, low discount rate people can afford to invest more in signalling their type than high discount rate individuals.

1.2.3 Signalling equilibria

The equilibria of signalling games—combinations of actions by actors that are best responses given their beliefs—which determine whether a signal has the potential to be reliable, or not, can be summarised as 'separating' (also called sorting), 'semi-sorting', and 'pooling' equilibria.

If a signal exists that a high type can afford to take but a low type cannot mimic, the existence of a separating equilibrium is guaranteed (Bacharach & Gambetta, 2001, p. 160). In a separating equilibrium, the low and the high type take different actions, and consequently, they are perfectly separable by the receiver based on their signals. Gustav Eiffel famously included a small apartment for himself at the top of the Eiffel tower, serving as an unquestionable indicator of his confidence in the robustness of the flimsy-looking structure.

In semi-sorting equilibria, all high types take a signal, and some proportion of low types, less than one, also take the signal. For this to be possible, low types must just be able to afford the signal that a high type can also afford (Bacharach & Gambetta, 2001, p. 160).

Observing a signal here only partially guarantees that the sender is a high type, because some low types also emit the identical signal. Many signals are of this type: signatures, passports, debit cards, tattoos, scars, are all, to a more or less degree, vulnerable to mimicry by people who are determined enough. Even self-immolation or suicide bombing—which initially appear to be truly separating signals of conviction, determination, and sacrifice—are not entirely separating for the signal's reliability can be called into question. People taking such actions may be mentally ill or unstable—induced into a temporarily clouded mental state—be under the influence of drugs, preventing them from fully appreciating, or caring, about the cost of the act. They may also be tricked: one may be told that others will rush in to put out the fire after he sets himself ablaze, when in fact they do not, and leave him to burn. Nevertheless, such extreme signalling acts elicit strong reactions from observers precisely because they are near-mimic proof: Mohamed Bouazizi's self-immolation in 2011 started the wave of protests that spread across North Africa and the Middle East.

Completely unconvincing signals, of which there are also many, occur in pooling equilibria in which both high types and low types can afford the signal (Bacharach & Gambetta, 2001, p. 160). When pooling occurs, high types take the same action as low types, and observing such a signal gives no clue about the type of the sender.

1.2.4 The sign-signal distinction

Signalling theory employed in economics does not typically distinguish between signs and signals. Whether or not the sender intended to change the beliefs of receivers does not matter. Why separate the two in sociology?

One answer is that the two have to be distinguished when using signalling theory to explain human behaviour. In most circumstances, the actions of a sender who generates a sign cannot be explained by the information content of that sign. Only if there is a feedback loop, through which a sender incurs consequences, and learning by the sender occurs, is it plausible to explain that sender's behaviour in terms of the unintended consequences that arise from their sign.[18] On the other side, the actions of receivers can be explained irrespective of whether they interpret a sign or a signal.

A second answer is that when a receiver is confident that a sender's manifestum is a sign, and not a signal, there is no need to use signalling theory to assess its reliability. The statistical link between the sign and the quality can be taken at face value. Footfalls, or the sound of a car pulling up a driveway are perfect signs of someone's approach when the actions creating them are not intended as communication. Usually, this ideal condition does not exist and there is nearly always a shred of doubt in the mind of the receiver that a signal sender is actually mimicking a sign-emitter. Could the footfalls of the approaching person, who is meant to be arriving for an illegal weapons deal, be faked by an associate to serve as misdirection, allowing the 'genuine' seller to pull a double-cross? This may introduce a meta-game, one in which a mimic attempts to convince a receiver that an action is a sign, and not a signal, to avoid having to fulfil the conditions of reliability of signalling theory.[19] The separation of signs and signals delimits the application of signalling theory.

---

[18] Consider a person who takes a non-signalling action *a* when confronted with a decision, bringing him utility u(*a*). As a by-product, it produces some information that observers use and that brings the person some additional benefit u(a_s). Confronted again with the same situation, that same person takes action *b* that leads to the utility u(*b*), where u(*b*)>u(*a*) but u(*b*)<u(*a*)+u(a_s). In absence of the sign's benefits, upon repetition the person would prefer to take the action *b*. In contrast, if *a* also functions as a sign, the person prefers to take the action *a* so long as there is learning. Here the sign affects the person's decision and is part of the explanation for that person's behaviour.

[19] Think of feinting in a football game to mislead an opposition player. From the sender's perspective, this is a signal—an intentional message for the opponent to go one way. From the receivers perspective it is a

1.3 Countersignalling theory

People sometimes extract information out of signs, and interpret it, at an explicit, or conscious, level. Experienced muggers may carefully consider the signs of their potential victim before deciding to 'move in' or not. Often, however, extraction and interpretation occurs near-instantly and unconsciously. 'Thin slices' of others' behaviour, short samples of their actions, are enough for people to form surprisingly accurate beliefs about those people (Ambady & Rosenthal, 1992; Oltmanns, Friedman, Fiedler, & Turkheimer, 2004). Displaying silent videos, as short as 2s, of teachers' behaviour, to subject-judges is enough for them to form beliefs about the teachers that match well with the beliefs of different judges who have had much longer interactions with those same teachers (Ambady & Rosenthal, 1993). Fundamentally, people use the signs of others, consciously or subconsciously, to update their beliefs about those others. Signs in their varied forms pervade our social environment.

Standard signalling theory assumes that receivers have no information about specific senders apart from the knowledge of the prior distribution of types and the signal that those senders choose to take. It does not consider the presence of signs that senders possess, and out of which receivers can derive useful information about the sender, and by doing so removes a potentially important feature of many real-life interactions. In almost all interactions, we come to know something about the person we are interacting.

Countersignalling theory, a game theoretic model developed by Feltovich, Harbough, and To (2002), enriches signalling theory by including two realistic additions:

---

sign. Indeed from the receivers perspective *it must be a sign*. For if it wasn't a sign, then the receiver would have no reason to believe it. To make the signal work, the sender has to *disguise it as a sign*.

- noisy information that is exogenously emitted by senders and observed by receivers; and
- a third 'medium type' sender—in addition to the low and high types—who has an intermediate level of the quality.

A characteristic of the first addition, the noisy information, is that there is uncertainty in its link with the property of the sender. Observing only the noisy information, a receiver cannot be certain that the sender possesses a quality or not. Signs are typically like this. We cannot be sure that a scar on someone's face comes from a bar-fight or accidentally falling down a flight of stairs.

Its other feature is that it is outside of the sender's control so that the sender has no influence over what is emitted to receivers. This requirement of exogeneity may be fulfilled when a sender is not aware of signs emitting information. Potential victims of muggings are not aware that the way they walk can make them more or less attractive targets. According to Grayson and Stein the 'prime difference between perceived victim and non-victim groups, therefore, seems to revolved around a "wholeness" or consistency of movement' (1981, p. 74). Were they to become aware of this, and have enough incentive to consistently put effort into the behavioural change, they could modify the way that they walked. Still, so long as they are unaware, or unconcerned with it, the sign is exogenous to their influence.

Even if senders become aware of a sign, and have enough incentive to manipulate it, they may be unable to do so because some signs are physically outside of senders' control. A naturally taken deed, good or bad, that is recorded and sealed, say on video and uploaded to the Internet, is a sign that is close to being completely exogenous to the

sender. Given enough servers that host the video, it is near impossible to remove it, leaving others to observe the sign and infer something about that person. Reputation can qualify too. Once it is created, reputation is largely outside the control of the subject of it, and often, actions that are intended to change it for the better backfire.[20] The information contained within these signs is 'frozen' and difficult to later manipulate. Other manifesta that are naturally difficult to change may be used as signs by observers: race, ethnicity, gender, height, accent, and vocabulary are some examples.[21]

The second addition is that of a medium type sender, who possesses an intermediate level of the quality. Including a third, intermediate type sender realistically captures how people often categorise their own and others' subjective qualities. We perceive and differentiate between people who have qualities at the two extremes, we distinguish between the brilliant and dull, the tough and the weak, and the erudite and the ignorant. However, it is difficult to make fine gradations in qualities that lie in the large grey area between the two. Even if possible, we often find it unnecessary, difficult, or too energy consuming, to differentiate between more than three categories. Instead, we group the in-between together into a single broad category.[22]

Including these two components into signalling theory brings about a starkly contrasting result of a certain kind of signalling equilibrium known as a 'countersignalling'

[20] Both signs and signals can be sources upon which a reputation is created. However, not all reputation is derived from signs and signals. Reputation can arise purely from another person spreading rumours and the truth-value of the reputation can be distorted from unintentional misrepresentation or intentional manipulation.

[21] This is not to say that such signs are always exogenous. In some contexts they may be displayed intentionally, turning them into signals: a potential employee could display a video of his past good deed to a future employer. However, the *content* of the sign remains largely exogenous.

[22] Based on a similar reasoning, Rapaport argues that grading should be done using a triage system (2011). 'The fundamental insight is that, whereas the extremes are clear (it is easy to identify clearly good work and clearly bad work), it not worth making fine distinctions among the work that is neither clearly good nor clearly bad.' (p. 351).

equilibrium.[23] Feltovich and co-authors find that senders who possess an intermediate level of the quality are the ones who invest the most into signalling; the intermediate types 'shout the loudest' and the 'most'. Both low and high types invest *less* into signalling than medium types. Specifically, high types send the same signal that low types send—high types mimic the lows in order to separate themselves from the medium types. That low types invest less into signalling than those in the middle is unsurprising, however, that high types also invest less into signalling is a counterintuitive and novel result.

For low types, their noisy information already marks them out as low types, making it unprofitable for them to signal. For those of a high quality, not investing in signalling is a way of separating themselves from the senders in the middle, who are afraid of being confused with the low types and, therefore, signal.

One factor favouring the emergence of countersignalling is when a signal is affordable to both medium and high types. When this is the case, high types cannot separate themselves by investing more in signalling, for medium types can afford the same signal, and so have to signal less than medium types in order to remain identifiable.

References to the idea of countersignalling can be found in numerous sources. The Islamic scholar, al-Jahiz, who lived in the 8th and 9th Century, noted that 'A man who is noble does not pretend to be noble, any more than an eloquent man feigns eloquence. When a man exaggerates his qualities it is because of something lacking in himself; the bully gives himself airs because he is conscious of his weakness' (Hourani, 2005, p. 52).

[23] I only discuss this briefly here for it is addressed in Chapter 2.

This alludes to one interpretation of countersignalling. Here, countersignalling is simply behaving naturally, ignoring the informative content that one's action emits. An educated person speaks with a natural accent, or an old money individual unthinkingly displays his antique Jaguar in the driveway. Neither make an effort to downplay, or destroy their signs. They simply behave naturally, in a way that is costless to them. They *avoid* the entire act of signalling and emit no signal. The avoidance of a signalling act is the action that can allow receivers to update their beliefs. The non-signal, or countersignal, is what informs; not doing anything when something is expected is the signal here.

The other way of interpreting countersignalling is that high types take a signalling action, but take a signalling act that matches with low types: buying a cheap, shabby-looking car or deliberately wearing tatty clothing as a wealthy person. Both are signalling acts, but the high types take the same signal as a low type. So long the action of high types pool with low types, it helps keep countersignalling reliable. The difference between the two interpretations is analogous to the difference between taking a test and deliberately doing poorly, or not even doing the test in the first place.

In addition to extending signalling theory to include important features, countersignaling helps us to explain some puzzling behaviours. Why do the nouveau riche waste money on extravagant displays of wealth while those with old money spurn gauche affectations (Aldrich, 1988; Duncan, 1973)? Why do some of the world's wealthiest people lack college degrees? Why do 'plastic gangsters' act so aggressively in prison?

It may also give us insight into the dynamics and change that social systems undergo. Consider luxury spending in China as a potential case that signalling and countersignalling can help us to understand.

Due to the Communist Party of China's policies of collectivisation in the 1950s, there was little wealth in modern China before the re-introduction of free markets and free enterprise under Deng Xiaoping in the 1980s (Cameron & Neal, 2003, pp. 344, 383). Thus, wealth and its signals were severely constrained. Following the reforms, a surge of spending on luxury goods, which may be understood as reliable signals of wealth, occurred (The Economist, 2011). By 2010, the overall spending on luxury goods in China was estimated to be $9 billion (Wwd Staff, 2010), or—according to Barclays Capital—about 12% of the world's spending on luxury goods (Marston, 2011). It is possible that wealthy people in China in the post-reform years faced a standard signalling game in which they could only communicate their wealth to others using costly signals that were affordable to relatively few. Since the wealth that senders possessed was so recent that they had not had the time to build up signs of it, and, on the other side, receivers had not had the experience of interpreting signs of wealth, senders were restricted to using costly signals. Consequently, people may have been spending on luxury goods to signal wealth.

Over time however, wealth built up so that the cost of luxury good displays become affordable to a greater part of the population, signs of wealth were created, and receivers may have become perceptive to signs. All of these are factors that assist the emergence of countersignalling: wealth displays are less separating and exogenous noisy information is built. Recently, there are some indications that are consistent with a countersignalling equilibrium. Some Chinese consumers are seeking 'different, subtler luxuries', avoiding

the 'standard' luxury brands and that there are at least some 'elites in the cities, who want less flashy brands' (The Economist, 2014a). Others are turning away from 'pure social status products' and making 'lifestyle purchases' instead (Zakkour, 2014). If these provisional data are found to hold, it could be that some wealthy consumers are beginning to move into a countersignalling system. Luxury consumption in China may then be a system that has moved from a no signalling equilibrium to a signalling equilibrium, and now is in the process of transitioning to a countersignalling equilibrium.

I do not discuss countersignalling theory in depth here. I leave this for the following chapter (Section 2.2) in the context of the trustworthiness experiment. Instead, I now turn to the sparse empirical research that has been carried out on countersignalling, and then to some examples, and domains, that I have found during the course of my doctorate to be germane for observing countersignalling.

1.3.1 Empirical research on countersignalling

Apart from the experiment published in Feltovich et al.'s initial work (which I discuss in Section 2.2), almost no empirical research has tested countersignalling theory. Out of the 40 sources reported on the Web of Science that cite Feltovich et al's paper, two set out to test countersignalling. The first paper to do so forms Chapter 2 of my thesis and is published in the *Journal of Economic Behavior & Organisation* (Gambetta & Székely, 2014).

One other paper, published towards the end of my thesis, tests the curvilinear prediction of countersignalling theory for advertising spending, the signal, by Taiwanese hotels using observational data (Chiu & Chen, 2014). As expected, they find that hotels of intermediate quality spend the most on advertising (p. 1022). All of the rest are theoretical papers or ones that mention countersignalling in passing.

1.3.2 Anecdotes of countersignalling

It is important to keep in mind three points when reading the following. First, the descriptions only correspond to parts of countersignalling systems. In some cases, they only recount the experience of a lone high type, and how he or she behaves to stand out. In others, both a medium and a high type are discussed while an implied low type is left unmentioned. Second, the illustrations come from a mix of perspectives. Most are written from the perspective of receivers and describe how the behaviour of senders influences their evaluation of senders' types. Others come from the perspective of senders and their views on how they are evaluated by receivers. Rarely do I have the luxury of observations from both sides simultaneously. Third, countersignalling is a mechanism, which implies that not all people who are a certain type can be expected to take the same behaviour; we can expect that there is variation in the behaviours of people with the same quality-level.[24]

Perhaps *wealth and social status* are the qualities regarding which most instances of countersignalling-like behaviour can be found. An obvious place to begin concerns the displays by new and old money. People from Old Money backgrounds are known for actively being discreet with their wealth. Nelson W. Aldrich Jr.—who comes from an Old Money background himself—writes that it is the lack of wealth displays by old money that creates the high-status belief in others, 'once the wealth has been there, for perception, it needn't go on being there. Indeed, *it must not go on being there*; it must retire discreetly behind the veil of time and disappear like the Cheshire cat, leaving a smile.' (Aldrich, 1988, p. 31, italics in original). Not only is it unnecessary for them to display

[24] Despite my attempts to find clear examples of animal behaviours corresponding to countersignalling, I have been unable to identify any.

wealth, but they should actively make the displays of wealth disappear. Describing the houses of old money and new money in an English village, Duncan highlights how old money actively de-emphasise the value that their houses appear to have—they down-play this signal of wealth (1973, p. 353). In an ethnographic study of old and new money people in a rural north-western Connecticut town, Satow describes both the actively downplayed signal and reliance on exogenous information by old money who 'emphasized their superior breeding and nonconspicuous consumer style' (1993, p. 236). In contrast, New Money (often disparagingly called the *nouveau riche* or *parvenu*) can be brazen and gauche in displaying their wealth.

Sergei Brin—co-founder of Google—and Mark Zuckerberg—Facebook's founder—both billionaires, are known to dress plainly. Warren Buffet, Ingvar Kamprad, Jim Walton, Azim Premji, Frederik Meijer, Richard Kinder, David Cheriton, and John Caudwell are all billionaires who have been noted for their frugal, at least in public, habits (Hawkins, 2007).

Countersignalling-like behaviour is mentioned by Thorstein Veblen in his classic book *The Theory of the Leisure Class* (1899). He describes the transition from signalling to countersignalling-like displays, and a condition, the increase in affordability that matches countersignalling theory and provides fertile soil in which countersignalling thrives. In the first case regarding the use of corsets, and in the second instance the displays of aesthetic in landscape architecture.

> [T]he corset was until within a generation or two nearly indispensable to a socially blameless standing for all women, including the wealthiest and most reputable. This rule held so long as there still was no large class of people wealthy enough to be above the imputation of any necessity for manual labour and at the same time large enough to form a self-sufficient, isolated social body whose mass would afford a foundation for special rules of

> conduct within the class, enforced by the current opinion of the class alone. But now there has grown up a large enough leisure class possessed of such wealth that any aspersion on the score of enforced manual employment would be idle and harmless calumny; and the corset has therefore in large measure fallen into disuse within this class (Veblen, 1899, pp. 184–185).
>
> To-day a divergence in ideals is beginning to be apparent. The portion of the leisure class that has been consistently exempt from work and from pecuniary cares for a generation or more is now large enough to form and sustain an opinion in matters of taste. Increased mobility of the members has also added to the facility with which a "social confirmation" can be attained within the class. Within this select class the exemption from thrift is a matter so commonplace as to have lost much of its utility as a basis of pecuniary decency. Therefore the latter-day upper-class canons of taste do not so consistently insist on an unremitting demonstration of expensiveness and a strict exclusion of the appearance of thrift. So, a predilection for the rustic and the "natural" in parks and grounds makes its appearance on these higher social and intellectual levels. (Veblen, 1899, p. 136)

Potlatches, sumptuous competitive feasting ceremonies, are frequently mentioned signalling acts taken by the indigenous North American Indians. However, in a study on the Makah Indians that she carried out in the early 1940s, Elisabeth Colson describes a twist to the standard story (1953). Before white settlers arrived in the area inhabited by the Makah, status and power within the group were obtained by using costly displays of resources, for '[b]oasting and the display and assertion of privileges were validated by the distribution of goods. The Makah observed the old maxim, "[p]ut up or shut up."' (p. 215). Those who could afford to 'put up' were the Makah with large inherited wealth, indicating their aristocratic lineage, or those who had been able to build wealth proving that they had the support of group members. Following white arrival, 'it became possible for men and women to obtain large supplies of goods by their own efforts', destroying the reliable link between wealth-displays and lineage or the number of supporting group members, and making the signal affordable to a greater proportion of the population (p. 215). Consequently, people holding potlatches were disparaged for their callous attempts to obtain status, 'upstarts who are trying without success to force an entry into the upper

class' (p. 222). Or, as interpreted by Max Gluckman, as 'a *nouveau riche* trying to cover his low-class and that the real high-class people do not need to do this' (1963, p. 310).

I now turn to *expertise and knowledge*. Scientists of epic stature highlight how little they know. Galileo, Newton, Laplace all declared, in one way or another, that what they know is small (Merton, 1987, p. 7). The mathematician Eric T. Bell writes that 'A common and engaging trait of the truly eminent scientist is his frequent confession of how little he knows' (Bell, 1931, p. 204).

C.W. Mills notes that complex writing in sociology is used by sociologists who are afraid of being labelled as a 'mere literary man' or a 'mere journalist', as opposed to a scientist. It is the mediocre who are uncertain and resort to 'socspeak' (Mills, 1959, p. 220). Sperber describes how obscurity in expression, typically a sign of poor thinking, is interpreted as profundity when uttered, or written, by gurus, '[o]bscurity inspires awe, a fact I have been only too aware of, living as I have been in the Paris of Sartre, Lacan, Derrida and other famously hard to interpret *maîtres à penser*' (Sperber, 2010, p. 583).

According to Pierre Bourdieu, it is the 'most knowing avant-garde producers' who can afford 'the audacious imposture of refusing all refusals the recuperating, in parody or sublimation, the very objects refused by the lower-degree aestheticism. The "rehabilitation" of "vulgar" objects is more risky, but also more "profitable"' (1984, p. 54). Referring to aesthetic in general he writes that 'It is well known that all dominant aesthetics set a high value on the virtues of sobriety, simplicity, economy of means, which are as much opposed to first-degree poverty and simplicity as to the pomposity or affectation of the "half-educated"' (1984, p. 224).

Erving Goffman describes how Roger Stevens, a highly successful real-estate agent who engineered the sale of the Empire State Building highlights himself by avoiding displays that are expected of him. He describes 'the startling fact' that Roger 'has a small house, a meagre office, and no letterhead stationery' (1956, p. 32).

Businesses also exhibit countersignalling-like behaviour. Fischer Travel Enterprise an exclusive travel agency avoids advertising, does not list its phone numbers, and has business cards with limited information (Fombrun, 1996, p. 1). The Connaught Hotel in London 'thought by many to be the world's most exclusive' 'never advertises for customers and its managers go out of their way to avoid the press' (Fombrun, 1996, p. 2).

Companies that sell luxury goods to the wealthy seem to cater to a preference for *not* showing off. Expensive Gucci handbags have smaller logos than the cheaper ones, and more expensive Mercedes vehicles have a smaller Mercedes symbol (Han, Nunes, & Drèze, 2010). A cursory look at the Ralph Lauren clothing ranges demonstrates a similar trend. The cheapest 'Polo Ralph Lauren' polo shirts have large logos emblazoned on their fronts, the more expensive 'Black Label' ones have discreet logos, and in the most exclusive 'Purple Label' range, the logos are often absent. Berger and Ward (2010) find that intermediate expensive sunglasses are the likeliest to have visual branding on them: both cheap and expensive ones are less likely to do so. These informal observations nicely match the findings presented by Chiu and Chen regarding the concave relationship between advertising spending and quality among hotels in Taiwan (2014)

Finally, and fascinatingly, some 'monumental buildings' in *the old world*, which can be considered as displays of power and resources, follow the pattern predicted by

countersignalling. The biggest pyramids in Egypt, Khufu's Great Pyramid, and Khafre's pyramid, were erected during the 'Old Kingdom', before the state, and its rulers, reached the apex of its power. The Pyramid of the Sun and the Pyramid of the Moon in Teotihuacan, Mexico, were built early in that state's development, and while that state became more powerful later, no public buildings greater than these were ever built by it. In Peru, an early state built the largest pyramids although more powerful ones came after it (Marcus, 2003, pp. 115–134). Jared Diamond adds another instance, 'the emerging Japanese state had built huge keyhole-shaped earthwork mounds called kofun', the largest of which were built early in the rise of Japan's first, intermediate powerful, state (Diamond, 2003, p. 891).

None of the observations that I describe above conclusively prove that countersignalling occurs. There may be other explanations for the behaviours (some of which I discuss in section 1.3.4), and the full countersignalling system, three sender types, receivers, noisy information, the signal, and associated costs and benefits, is not clear. A subtle point cannot be addressed by the anecdotal observations: whether the lack of signalling by high types is informative or not. In other words, does the countersignal by high types help receivers to identify the sender as a high type? Alternatively, and in which case it is not countersignalling, is the non-signal by high types simply a way for them to save on the costs of signalling?

It is not surprising that the observations do not conclusively isolate countersignalling as an explanation. Countersignalling is subtle, and researchers are only ever likely to record the precise information that is needed when one goes searching for them with the specific theory in mind. The lack of signalling by high types is a key component of

countersignalling theory, yet few researchers will record an absence of an action without a specific reason, as provided by countersignalling theory, for doing so.

Why use countersignalling? One good reason is that the components of the theory are plausible and widespread. Another is that actions that could be countersignalling can be found across a variety of domains, suggesting that it has explanatory scope. It may also be useful for understanding dynamics of signalling systems, particularly how systems change when signs are built up over time and when signals become more accessible. There is also an omission in the literature regarding empirical tests of countersignalling. Indeed, a motivation of this thesis is to take seriously the theory of countersignalling.

1.3.3 Countersignalling mimicry

A potential case of countersignalling mimicry can illustrate how not signalling changes receivers' beliefs in such a way that they evaluate the non-signaller as a high type. In 2009, Bernard Madoff was convicted of starting, and running, one of the biggest pyramid schemes in the history of finance. By skilfully manipulating his manifesta, he convinced thousands of investors—people highly motivated to discern between genuine and fakes—that he was trustworthy, extracting at least $65 billion dollars in investments from them. His strategy was three-fold, (i) advertising his fund as returning only modest profits, (ii) deterring potential investors by making it 'seem impossibly difficult to invest with him', and (iii) even after accepting investor's money he emphasized his reluctance about having accepted the money (Chernow, 2009). Indeed, 'Madoff posed as a man beleaguered by his own generosity, who took on new clients as a favour to friends' (Chernow, 2009). Instead of signalling his trustworthiness and ability, he actively downplayed them in order to highlight the signs of his reliability; relying on word-of-

mouth to spread his reputation and his previous achievements as a 'pioneering marketmaker' (Chernow, 2009; The Economist, 2008).

By avoiding the typical signals that arouse investor suspicion, claims of high yields and actively courting them, and relying on his exogenous signs, he was able to become highly trusted by usually cautious individuals; to such an extent that typical questions, or raised suspicions, were brushed aside (The Economist, 2008). His exploitation of the system, by using methods predicted by countersignalling theory, is suggestive of the theory's applicability.

1.3.4 Related explanations

There are two related conjectures that may be competing with countersignalling as explanations for real world phenomena. One may be called 'subtle signalling' and the other is known as 'middle status conformity'. I explain how they differ to countersignalling, and exclude them by design in the experimental chapters. Nevertheless, because they should be kept in mind when considering real-world phenomena it is worthwhile to mention them.

First, senders may use subtle signals that only receivers who are 'in the know' can detect. These senders only signal to a subset of sophisticated receivers who can recognise the cryptic signals that they emit. Subtle branding on luxury handbags may be one example of this.

Consider the following setup with three types of senders, high, medium, and low, as in countersignalling theory, who choose between three different handbags as signals: a simple handbag with no branding s_0, a luxury handbag with obvious branding s_1, and a

luxury handbag with subtle branding s_2. High and medium types can afford s_0, s_1, and s_2 handbags, while low types are restricted to buying only s_0. There are two types of receivers, sophisticated and unsophisticated. The former accurately distinguish between all three signals, s_0, s_1, and s_2. In contrast, the latter cannot distinguish between s_0 and s_2, and may only distinguish between s_0 and s_1. Finally, high types only care about signalling their type to sophisticated receivers, while medium and low types care only about signalling their type to unsophisticated receivers.

In this scenario, low types are restricted to emitting s_0, so they buy and display the non-branded handbag. Medium types avoid s_0 so that receivers can distinguish them from low types, leaving them to choose between s_1 and s_2. They would like to send the same signal that the high types emit, whichever that is. High types want to distinguish themselves from low types so do not buy the non-branded handbag (s_0), also leaving them with s_1 and s_2. In addition, they want to distinguish themselves from medium types. Were they to emit s_1, it would leave open the possibility that medium types mimic them and the sophisticated receivers treat them as such—the signal is affordable to both and medium types are happy to be identified as high types. Instead, they send the signal s_2 because medium types cannot mimic this signal. Why can medium types not mimic s_2? The receivers to whom medium types want to signal, cannot distinguish between s_0 and s_2, so if they were to take the signal s_2, they would be confused with low types who are emitting s_0. By emitting s_2, high types separate themselves from both medium and low types and are correctly recognised by sophisticated receivers. As they do not care about unsophisticated receivers, it does not matter what they believe about the high types. This leaves medium types to send the signal s_1.

The reliability of the signal by high types, like in countersignalling theory, is ensured by its resemblance to the signal emitted by the low type. High types here too use the signal of a low type as a 'wall' behind which they can take cover and avoid being mimicked by medium types. Were no low types to signal s_0, high types could not use s_2 to separate themselves as medium types would be free to mimic them without fear of confusion with low types. In this sense, this explanation is very strongly related to countersignalling, although it is still separable. For, under countersignalling, both high and low types are predicted to invest in s_0.

Second, individuals whose group memberships are not secure and may be called into question conform the most to their group's ideal behaviours. They fear being excluded from the group and desire to be accepted as full members. People outside of the group, and those who are firmly established in the group, are both less conforming in their behaviours. The former gain nothing from conforming and the latter are already secure in their group membership. This is known as the middle-status conformity conjecture (Phillips & W. Zuckerman, 2001).

The explanation of middle-status conformity shares some strong similarities to countersignalling. Countersignalling theory's three types, low, medium, and high are analogous to middle-status conformity's outsider, borderline, and insider status individuals. The actions of the non-conforming insiders and outsiders can be the same under middle-status conformity, matching the predictions of countersignalling. The conforming actions by those on the borderline can be interpreted as signals, displays of commitment or desire to be a group member.

Still, some important differences allow the two to be separated. Countersignalling theory posits the presence of signs in addition to the signalling possibility. Middle-status conformity does not require their presence. Next, middle-status conformity makes no prediction about the type of non-conformity that should happen. It specifies the set of actions that insiders and outsiders do not need to conform to, but not the actions that they should take. Countersignalling theory, in contrast, specifies both the actions low and high types should avoid, and the actions that they should take: high types should pool with low types in their signalling behaviour. Finally, high type non-conformity is not done to signal, to separate from the medium type—as in countersignalling, although, of course, it may still function as a sign to observers.

1.4 Multiple equilibria

Signalling games, including countersignalling theory, typically have multiple equilibria. Even when the relevant conditions are met, such that high types can afford to signal, and low types cannot afford to do so, pooling equilibria may exist. This is because senders' and receivers' beliefs play a crucial role in determining their expected utilities and therefore their actions. Their beliefs on the equilibrium path (the combination of actions that are played in the equilibrium)—based on their beliefs about the prior distribution of types and updated using Bayes' theorem—and their beliefs about out of equilibrium behaviour (actions that are not played in equilibrium). Equilibrium refinements have been proposed, to whittle down the number of plausible equilibria, but so far, no solution has been found to always predict a unique equilibrium solution. The practical implication of multiple equilibria is that pooling, partial pooling, and separating outcomes may be compatible with signalling and countersignalling theory, making it difficult—outside of demonstrating inconsistency between beliefs and actions—to

definitively disprove the theories. Therefore, a lack of support for the separating and partial pooling predictions do not disprove the respective theories themselves.

Nevertheless, if signalling and countersignalling theory are to have useful explanatory value, separating and partial pooling equilibria should be observed in plausible situations. For if they are not, it makes it less likely that these equilibria are the ones that actually are played by people in the real world. Throughout the thesis, I use the separating prediction of signalling theory and partial pooling prediction of countersignalling, and setup experiments accordingly.

1.5 Information availability and conflict

Signalling and countersignalling theory both concern the *content* that information takes if it is present. The former predicts the signal that senders should send when that is the only source of information and the latter predicts the signal when there is also another channel, noisy information, for information transmission.

Another perspective is to consider the *availability* of information, irrespective of whether the information shows the person emitting it to be low or high type. How does the presence, or absence, of information affect conflict when there is asymmetric information?

When there is conflict in interests over resources—all involved parties prefer to have that resource—and no allocation mechanisms apply, information about the ability of the people involved to use violence could peacefully distribute them. Knowing who is likely to win, were a fight to take place, removes the need to actually use violence. The high possible costs from fighting produce a strong motivation to avoid it. In contrast, when

no signs and signals of fighting ability can be observed or emitted, physical force has to be used. I discuss this theory in Chapter 3.

1.6 Summary

I began by introducing the topics of the thesis, the chapters and methodologies that I use, and the research questions that I address. Then I turned to the theoretical framework for communication, and the two specific theories of signalling and countersignalling, describing the available evidence on countersignalling theory, and using multiple qualitative anecdotes and a potential case of countersignalling mimicry, to argue that it is a germane theory to test. Finally, I addressed two explanations related to countersignalling, the issue of multiple equilibria, and the conjecture on information availability and fighting. I now turn to the empirical chapters.

1.7 Contributions

'Signs and (Counter)Signals of Trustworthiness' is co-authored with Diego Gambetta. Both Aron Szekely and Diego Gambetta conceived and designed the experiment, analysed the data, and wrote the paper. It appeared in the *Journal of Economic Behavior & Organisation* in 2014 and I include it in its published form with some marginal changes. Diego Gambetta assisted in the design of the experiment used in the chapter 'Communication and Information in Interpersonal Conflict'.

2. Signs and (Counter)Signals of Trustworthiness

2.1 Introduction

We investigate a common case in which people realise that some past action of theirs, which they performed for motives other than informing others of their qualities, becomes known to others in situations in which this could have consequences for them.

Going about their lives, people act in myriad ways, both good and bad. Many of these actions are trivial and go unnoticed, forever lost. But some actions are perceived by other people, other actions are recorded on media whether agents are aware of it or not, others still leave enduring marks on them, perceivable well after the event that produced them occurred. The information potential contained in these past actions can remain dormant. However, out of this baggage with which everyone travels, people often extract *signs* of other people's qualities, good or bad, and use this information in deciding how to deal with them.

Our question is: what do people do when they realise that this is happening? Under what conditions do they take further action or do nothing about it? If the sign is of a quality of theirs that can help them in the situation in which they find themselves, should they still invest resources in signalling to further strengthen the receiver's belief that they truly have that quality? If an inmate enters jail preceded by the information that he has received a 20-year sentence for armed robbery, does he need to, as it were, flex his muscles publicly and vigorously to ensure that other inmates will leave him alone, or can

he afford to sit back and just bank on that indirect sign of his toughness?[25] And if the sign is of a quality that could be detrimental, should people engage in some new deed to make up for it? Can an inmate rumoured to be a 'snitch' do anything to persuade other inmates that it is all a big misunderstanding, and that he really is a most loyal fellow, or would further actions make his situation even worse?

2.2 Theoretical background

Observable features of an agent that are displayed *intentionally* by that agent for the purpose of changing the beliefs of others about a quality of that agent are called signals. Like signals, signs carry information about people's qualities, but unlike signals, signs are not produced with the intention of informing anyone. Signs, in our context, are perceivable features of an agent which *unintentionally* convey information about qualities of that agent (Gambetta, 2009b, p. 170). They are by-products of our actions rather than being produced by us in order to send a particular message. On the one hand, the lack of intention to inform makes signs a source of reliable information—for instance, small gestures of generosity when it is clear that they are not carried out instrumentally are signs of genuine generosity (Gambetta & Przepiorka, 2014). On the other hand, the information that signs convey tends to be noisy: while raising the probability that an agent has or lacks a certain quality, they often fall short of providing full reassurance. This is either because some 'noise' in the transmission affects their reception or because the actions that created the signs are imperfectly linked to the quality under scrutiny.

[25] What, exactly, committing armed robbery conveys depends to some extent on the context. In America, it may be less convincing as a sign of toughness than in Britain. Anticipating that civilians will be armed, robbers in America arm themselves too, making armed robbery the standard form of robbery. In Britain, armed robbery counts as an extreme action that only those strongly committed to the possibility of using violence take.

Signalling theory—arguably the best instrument we have to understand intentional communication when agents' interests do not fully overlap and information is asymmetric—is silent about signs.[26] In the theory-stylised landscape, the receiver can only use the signalling action taken (or avoided) by the sender to update his beliefs about that sender. Until the sender takes an action, the receiver has to rely on prior beliefs about the base-rate distribution of types to identify the sender.[27, 28] In real life, however, this simplification—especially outside of anonymous markets—is almost always violated: before receiving the signal, receivers already know something relevant about the qualities of the sender they are interacting with.

An extension of signalling theory, known as 'countersignalling theory' (CS) (Feltovich et al., 2002), can help us to deal with these richer situations. This theory discards the assumption of a single source of information and admits, realistically, three conditions:

- the existence of prior information independent of signalling but dependent on a sender's quality—sometimes called exogenous information—which the receiver has about the sender and which the sender knows the receiver has;
- the noisiness of the exogenous information[29]; and

[26] We refer here only to the signalling literature developed in economics. The theory in biology differentiates between signals and signs—which are called cues (Diggle, Gardner, West, & Griffin, 2007, p. 1242; Scott-Phillips, 2008)

[27] In some economic renditions of the theory, agents are conceived as strategic to the hilt, ready to always take into account the signalling value of anything they do before they do it, so nothing is ever a sign. This may be true in some extreme cases in which the peril of mutual aggression is permanently present as among Mafiosi or inmates. But this seems hardly the case for most ordinary people in most daily activities.

[28] In his seminal paper (1973) and subsequent book (1974) Michael Spence touches on this issue. In his paper, his second model pp. 368–374 (and Model 3a in his 1974 book), addresses the impact that an uninformative—evenly distributed according to quality—non-alterable characteristic (which he calls an *index*) has on signalling equilibria. Also, in Model 3b in his book (pp. 38–46), he looks at the effect that an index has when it is linked to signalling costs. However, Spence's indexes are different to what we are concerned with here for they are not correlated with quality.

[29] Specifically, four features typify the exogenous information included in Feltovich et al.'s model. First, senders know the probability distribution that their exogenous information can take, but not its specific realised outcome that receivers observe. Second, receivers observe the specific point outcome of the exogenous information that is drawn from the sender's probability distribution. Third, the probability of

- the existence of three levels of sender quality, or types, instead of the customary two types—countersignalling theory includes not only those who have or do not have the quality of concern, known in the theory jargon as 'high types' and 'low types' respectively, but also those who fall somewhere in between, the 'medium types'.

When these features are introduced, the main finding[30] of CS theory is that in certain conditions medium types will signal more intensely than both low types and, more surprisingly, high types too.[31]

This is how, intuitively, the reasoning goes. Low types carry prior information that makes their type partially apparent to receivers. The best they can reasonably hope for, given the noise of their exogenous information, is to be mistaken for medium types, while they know that they are very unlikely to be mistaken for high types. But what for low types is a dream for medium types is a nightmare, so the prediction is that

- *medium types will signal vigorously (s*) investing resources that ensure that they will distinguish themselves from the low types.*[32]

Expecting this

- *low types would then find it too costly to signal with such an intensity as to be thought of as medium types, hence signalling for them is not worth it (s_0).*

---

sending a higher level of exogenous information correlates positively with quality. Fourth, senders have no control over it: it is exogenous to sender choice.

[30] Although there are multiple equilibria as with other signalling games, the countersignalling equilibrium is robust to common out of equilibrium belief refinements, namely the Intuitive Criterion, D1, and D2.

[31] Like signalling theory, CS theory assumes that there is a negative relationship between quality and signalling cost, and that as the signalling level increases the signalling cost increases at an increasing rate. In other words, cost is decreasing in quality and convex in signalling level. The cost of signalling is assumed to satisfy the single-crossing property, thus the marginal cost to signalling is lower for higher type senders.

[32] s represents a signal in the set $S \subset \mathbb{R}_+$ and $s^* > s_0$.

The decision of whether to signal or not to signal for low types is a consequence of two contrasting pressures: low types who emit the signal s^*, would be sometimes mistaken, to their advantage, for medium types. However, taking into account the exogenous information that already allows receivers to gauge them as low types, a substantial proportion of the time, their gains from signalling s^* would not be enough to offset their costs.

Medium types, in contrast, by sending a signal s^*, distinguish themselves from those below, bringing themselves benefits. Yet, signalling has its downside for medium types: by sending s^* they also distinguish themselves from those high types who choose not to signal (s_0). The consequences of signalling are double-edged for mediums types.

At the other extreme, high types display exogenous information that gives receivers a strong hunch that they are in fact high types. The worst that could reasonably happen to them is to be mistaken for medium types, while they can be confident that they will not be mistaken for low types. For high types too, signalling vigorously may seem like the obvious course of action to pursue but, under certain conditions, it would be an unwise one: if the high types did that they would behave as, and be mistaken for, medium types. Enter countersignalling: a provocative result of the model is that

- *high types separate themselves from medium types by not signalling and pooling with low types (s_0).*

The lack of investment in signalling itself has an informative effect for receivers and is not just a saving of resources achieved by not signalling. Doing nothing, being 'laid back', turns out to be informative: high types fully separate themselves from medium types for, given what is already known about them, they can afford to signal by not signalling. Their silence confirms their type.

But why, one may wonder, do high types not raise their signal intensity to separate themselves from medium types given that they could afford it? This question, addressed by Feltovich et al. (2002) in their founding paper, can be summarised as follows. First, there are situations in which increasing the signal intensity is simply impossible and CS becomes the only separating course of action open to high types: this occurs when the signal intensity is capped upward (p. 640).[33] Next, there is a more complex reason. When high types countersignal, they send the same signal as low types—they 'pool' in their actions—but are still separated from low types by their noisy exogenous information, which allows them to be identified correctly most of the times. If the exogenous information sufficiently differentiates between high and low types, the gain in payoffs that could potentially be obtained by high types who signal is too small as they are already correctly recognised often enough. With a certain payoff distribution the cost of the minimum signal that high types would have to send to make it unaffordable for any medium type to mimic would offset the additional benefit (pp. 636–639). This means that high types prefer pooling with low types, instead of trying to out-signal medium types.

CS theory can make sense of two facts suggested by casual observation and some evidence. First, that people whose quality is open to question go out of their way to convince others that they truly possess the relevant quality—*nouveau riche* display opulent

[33] For example, if one can only pass or fail an exam. Medium types can make an effort and pass, and so can high types of course. But if high types passed, they would not be distinguishable from medium types. Failing for them can be the optimal (counter)signalling option for, coupled with the exogenous information, it ensures that they are not confused with low types, and that therefore, *a fortiori*, they can only be high types (remember that medium types are forced to try and pass the exam for their exogenous information is not strong enough to sufficiently distinguish them from low types).

appurtenances,[34] inmates or school children with feeble reputations behave more aggressively to signal their toughness.[35] Second, the converse observation, that high types who have what it takes seem blasé about it, and do nothing to persuade others of their qualities, even playing them down—as people often like to repeat, Steve Jobs was a college drop-out.[36] One could conjecture that high types do not try hard to persuade anyone that they are skilful, clever, wealthy, tough or trustworthy simply because it is perfectly clear that this is how they are. Their past actions speak for them loud and clear. High types' abstention from signalling could thus be motivated by a desire to economise on resources rather than to send information through a non-signal.[37] But countersignalling theory makes a subtler claim: its counter-intuitive results speak to those situations in which the exogenous information is not rock solid and could be thrown into doubt: then, the theory predicts, it is the very lack of signalling that fully reassures receivers.

CS theory to our knowledge has been investigated experimentally before only by Feltovich et al. (2002). Student-subjects played the role of senders and a computer played the role of employer-receivers. Senders were, at the start of each experimental interaction, randomly assigned a productivity type (high, medium or low). They were then asked to choose whether to invest some money into obtaining a good grade or not

[34] In *La Distinction*, Pierre Bourdieu writes that 'Where the petit bourgeois or nouveau riche "overdoes it", betraying his own insecurity, bourgeois discretion signals its presence by a sort of ostentatious discretion, sobriety and understatement, a refusal of everything which is "showy", "flashy" and pretentious, and which devalues itself by the very intention of distinction.' (1984, p. 249).

[35] For inmates see James (2003, p. 17); Sykes (1958, pp. 104–105); for high school children see the longitudinal study by Faris and Felmlee (2011, p. 60).

[36] Oddly, Steve Jobs' alma mater—Reed College—out of which he dropped, also behaves nonchalantly about its quality; it is an intensely academic liberal arts college that became the first U.S. university to refuse to participate in the college ranking system.

[37] This explanation rests on the assumption that the exogenous information is noisier for low and medium types than it is for high types. That is, low types may be mistaken for medium types and vice versa, but high types are unlikely to be mistaken for anything but high types. Then medium types may have an incentive to separate themselves from low types by signalling and high types would not have an incentive to signal, saving resources and accounting for the inverted U-shaped curve pattern that CS can explain.

investing money and getting a bad grade; here 'grade' formed the binary signal. Obtaining a bad grade in a test was costless for all types, while the cost of a good grade correlated inversely with type. Additionally, and outside the control of senders, they each automatically sent a 'test score' which could be a pass or fail, and which functioned as their exogenous information. The computer-employer was programmed to employ all senders and pay them an amount equal to their expected productivity, conditional on both grades (signal) and scores (exogenous information). The subjects were not told about the countersignalling equilibrium, although they knew that they were being judged according to their expected productivity. Feltovich et al. allocated subjects to one of two groups in a between-subjects design. One was a control group in which the test scores (the exogenous information) of senders did not depend on type (they assigned a 0.5 probability of emitting a pass or fail to all types)—making the exogenous information useless to evaluate productivity potential. The other was the countersignalling treatment group in which the exogenous information was informative, for the probability of emitting a pass correlated positively with type: high types always emitted a pass, medium types sent a pass half of the time and a fail the other half of the time, and low types always produced a fail.

Their results proved partially satisfactory: they found that (i) high types chose to obtain good grades less often in the countersignalling treatment group than in the other group—as predicted by theory; however, they also found that (ii) medium types too chose to invest in getting good grades less often in the countersignalling treatment group, meaning that a comparison between high type and medium type signalling

behaviour (a key part of the theory) in the countersignalling treatment led to a non-significant difference.[38]

2.3 Design and hypotheses

Instead of academic skills and productivity, for our experiment we chose generosity and trustworthiness as the qualities of concern. We exploit the fact that generosity is correlated with trustworthiness and can thus be a sign and a signal of the latter (Ashraf, Bohnet, & Piankov, 2006, p. 202; Gambetta & Przepiorka, 2014).

Our experiment consists of three parts.[39] First, subjects participate in a *veiled* dictator game (DG) through which we expect them to reveal the level of their generosity, and which will be used as their noisy sign—'veiled' because they are not informed of what will follow in parts two and three. Second, they are informed that they will take part in a trust game (TG) in which what they chose in the veiled DG will be revealed, albeit somewhat 'noisily' (uncertainty is added to the information) to trusters. Third, before playing the TG they are given a chance to revise what they gave in the veiled DG. Here they can choose to send a signal to trusters to strengthen or override what they did previously. Both their veiled and unveiled giving is transmitted to trusters in the TG. Those who choose to revise participate in a repeat of the DG, this time unveiled; those who choose not to revise, whether because they are content or resigned, move directly to the TG.

[38] They report a *p*-value of 'about 0.12' for a χ^2 test and a *p*-value of 0.0643 for a robust rank-order test.
[39] Subjects were told at the start of our experiment that it consists of three parts. What we did not tell them at the time was what parts two and three consisted of. A number of other papers use a similar design. Schniter et al. (2013) reveal an end trust game to their subjects only after these make their initial choices; Ellingsen et al. (2010) first elicit their subjects' beliefs and only after tell them that their beliefs will be sent to others, and Binmore et al. (1985) reveal the existence of a second round only after their subjects finish a first round of bargaining (reported in Thaler, 1988, p. 199). See Ellingsen et al.'s (2010, p. 96) discussion about why such a design arguably does not violate the norm of non-deception in experiments.

We employ a DG with a non-binary discrete choice-space to allow subjects to generate their sign, on the basis of which we then categorise their type, as high, medium or low.[40] We use instead a dichotomous TG, which is complex enough to capture the information that we need but simple enough to be readily interpreted.

With a few exceptions (Fehrler & Przepiorka, 2013; Gambetta & Przepiorka, 2014), in studies of signalling (including Feltovich et al.'s experiment), experimenters randomly assign participants their type and corresponding payoffs. This exogenous attribution of type relies on an unrealistic ability to switch type at will, on a mental plasticity which is not in everyone's gift: subjects who in the real-world belong to one type may be assigned a different type in the lab, and while some subjects may adapt to the incentive structure provided by the experimenter, in other cases their imprinted habits may determine their lab behaviour. In order to avoid this confounder, we sort participants out according to their *real* generosity as manifested in the veiled DG; in other words, we allocate subjects their 'true' type.[41]

2.3.1 Veiled dictator game

Subjects are given an endowment of £8 and asked to decide how much of it to allocate to an anonymous passive subject, who is given no endowment, and how much of it to keep for themselves (for screenshots of all the stages see the experimental instructions

[40] We could have adopted a binary sign structure as Feltovich et al. (2002) did. In this case, we would *a priori* decide that a given amount would be shown to receivers as 'generous' all of the times, a lower amount would be shown as generous half of the times, and a still lower amount would never be shown as generous. So the sign the receivers observe would not be the actual amount sent in the DG, but an indicator of the probability that a sender is of a given type. We decided against using this structure because we preferred to preserve a closer link between the action of the sender and the content of the sign in order to keep the experiment as simple as possible.

[41] We did not tell subjects what proportion of the participants were high, medium, or low type. In this respect, in our experiment subjects do not have common priors, contravening an assumption of signalling theory.

provided in Appendix B). Passive subjects were randomly allocated participants whose only role was to receive money given to them by trustees in the unveiled and veiled dictator game. After the two stages, passive subjects left the experiment and did not play the Trust Game. This ensured that the decisions of trusters was only affected by the sign and signal the trustees generate.[42, 43]

2.3.2 Unveiling

When making their initial DG choice, subjects are 'veiled', that is not informed of the further stages of the experiment. Thus, their decision can be regarded as a natural sign of their generosity. Subsequently, subjects are unveiled and given instructions for the rest of the experiment.

They are then given the option to revise their initial giving. If they choose to revise, they repeat the DG, but this time knowing what is to come later. If they choose not to revise they move directly to the final stage that is the TG. Trustees who choose to revise (*revisionists*) are randomly paid for *either* their initial or revised giving and trustees who choose not to revise (*non-revisionists*) are paid only for their initial choice.[44]

---

[42] Subjects are shown eight one-pound coins on their part of the screen that they can allocate to a passive subject in £1 increments from a minimum of £0 to a maximum of £8. To give money to a passive subject, active subjects click on a coin on their part of the screen that transfers it to the passive subject's side of the screen. To reverse this, subjects click on a coin on the passive subject's screen section. They can change their minds as many times as they wish before clicking on the 'Continue' button.

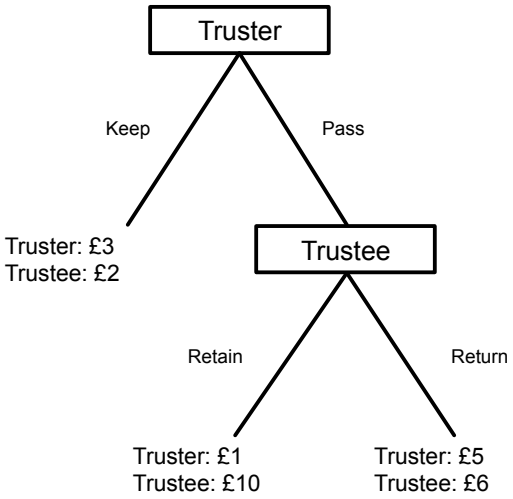
[43] We used two sets of control questions to ensure that all subjects understood the experiment. One set of control questions was asked before subjects participated in the veiled dictator game. The other set was asked after the unveiling and before subjects decided whether to revise or not. After each of these, the experimenter verbally went over the correct answers to the questions and gave a brief explanation.

[44] This corresponds to reality: being naturally generous incurs a monetary cost to the individual, then not revising—continuing on with one's natural behaviour and being generous to others—also re-incurs the same monetary cost.

2.3.3 Trust game

The final stage of the experiment is a dichotomous TG. Here, each subject first plays as a trustee and then as a truster. Trusters decide whether to keep their endowment or pass it to an anonymous trustee, and trustees choose either to retain or return any money sent to them by a truster (Figure 1). Only if trusters believe that trustees will return does it make sense for them to pass, otherwise they should keep, making it important for trusters to discriminate between trustees who will return and trustees who will retain. Trustees have a strong incentive to be trusted because this gets them an additional £4 if they are trustworthy in the TG or £8 if they are not, relative to not being trusted.

Figure 1. Our trust game. A truster has to believe that the probability of the trustee playing return is greater than or equal to 0.5 before choosing to pass (Appendix B).



When making their choice, trusters are given noisy information about the trustees' initial giving in the form of six overturned blue cards. They are told that only four of the cards hide the true amount the trustees gave, while one card covers a value that is £1 higher

than the true value and one card covers a value that is £1 lower.[45] Trusters are allowed to turn over only one card and see the value underneath. This injects some noise into the sign, which affects both sides, for trustees too do not know for sure what card trusters drew. Noisiness is a key part of the countersignalling mechanism as well as being a feature commonly present in signs.[46]

Trusters are then shown whether trustees chose to revise or not, and if they did a single red card appears displaying the exact value a revisionist trustee decided to give. If a trustee chose not to revise then the red card is absent.

We elicited trusters' choices by asking them how they would respond to different combinations of trustees' veiled and unveiled giving that were displayed to them sequentially.[47] Some of these combinations were chosen by trustees; others were never picked by trustees in the experiment and remained hypothetical combinations. Not all theoretically possible veiled–unveiled combinations were shown to trusters, but, to avoid a list that could prove tiresome for subjects, only a sub-set that seemed more interesting for us (for details on the combinations that we chose see Table 2 in Section 2.4). Trusters were not told which combinations were hypothetical and which ones were actually chosen by trustees. Trusters were not given feedback between their responses to the

[45] For those who give £0 naturally, one of the cards covers £1 and the rest cover £0. Consequently, there is less noise around the sign of someone who gave £0 than around the sign of someone who gave more. The upper limit of the sign (£8) was never reached, so the asymmetry is irrelevant at the top.

[46] As the correlation between generosity and trustworthiness is imperfect, some noise is already objectively present. However, we cannot assume that subjects know about this or that, if they do, they know how strong the link really is; for instance, trusters may believe that the link between meanness and untrustworthiness is near-perfect, and trustees may believe that that is what trusters believe. So, we introduced this additional uncertainty, which aims to make trustees more insecure of their position with respect to trusters, and trusters less confident about drawing conclusions from observing intermediate natural generosity.

[47] The order of veiled-unveiled combinations was randomised across sessions. Trusters who took part in the same session observed the veiled-unveiled combinations in the same sequence, and trusters who took part in different sessions saw the combinations in a different order.

different veiled-unveiled combinations, making our implementation a partial version of the strategy method. Trustees were asked to specify whether they would, in the event of being trusted, choose to return an equitable amount or keep most of the resources. In one randomly chosen round each truster was matched with a trustee and shown that trustee's veiled–unveiled combination. Payment for the TG was decided for both based on their decisions in this round (neither was aware which round would determine their payments).

The giving in the unveiled DG is designed to offer trustees an opportunity to send an exact signal of their generosity. For this signal to be reliable, it must be more affordable for a trustworthy trustee than for an untrustworthy trustee. There are two reasons why this can be the case. First, the expected cost of giving in the unveiled condition is higher for trustees who were less generous in the veiled DG—and are likelier to be untrustworthy—because trusters are less likely to bestow trust upon them, precluding the possibility for them to obtain either of the two higher payoffs of £6 and £10. Second, for truly generous individuals (who are more likely to be trustworthy) giving money, up to a point at least, is not a cost; it is only a cost for the selfish. There is however an opposing argument, in that an untrustworthy trustee may find it more affordable to give because he anticipates a higher gain to being trusted than a trustworthy trustee: the former will keep most of the money, £10, while the latter will split it, keeping only £6.

If we vary only the probability that a trustee expects to be trusted, then a trustee who intends to return in the TG has to believe that the probability that a truster sends him money is more than twice as high as the probability that a truster sends him money if he were ungenerous, before he can expect to gain by spending more on signalling. For instance, a trustee intending to be untrustworthy and who believes that he will be trusted

with a probability of 0.2 if he signals, and never otherwise, can spend up to £1.6 on signalling (10×0.2+2×0.8−2). If he intends to be trustworthy, he can spend the same amount on signalling if he believes that he will be trusted with a probability of 0.4 (6×0.4+2×0.6−2). By itself, this condition is fulfilled in a number of cases. Including unequal signalling costs into the calculation makes the condition easier to fulfil, as it decreases the difference in the probabilities necessary to allow a trustworthy trustee to signal more: e.g. if every £1 given by an untrustworthy trustee is twice as costly as for a trustworthy trustee, then a generous trustee only has to believe that he is slightly more likely than an ungenerous trustee to be sent money before he can afford more in signalling.

At the end of the experiment, we also administered a short questionnaire on the motivation of subjects for making their choices.[48]

We decided against having a control group with no communication between participants. Other studies of the binary-choice trust game (Bracht & Feltovich, 2009; Ermisch, Gambetta, Laurie, Siedler, & Noah Uhrig, 2009) reveal that about 40–60% of trusters trust strangers about whom they have no information. These studies include one carried out only six months before ours in the same laboratory, drawing subjects from the same pool and offering similar payoffs (Gambetta & Przepiorka, 2014), and we will take the level they found, 46%, as our benchmark. This spared us from having to run a control group. We programmed and ran the experiment using z-Tree (Fischbacher, 2007).

[48] The questionnaires were carried out at the end of the experiment while subjects' payments were being prepared—by which point they may have been tired and losing concentration—and they were not incentivised monetarily to complete them. Both of these points imply that the answers subjects wrote down regarding their decisions will be a 'lower bound' estimate of their reasoning and decision process.

2.3.4 Hypotheses

We expect trustees' giving in the veiled DG to be positively correlated with their trustworthiness in the TG (H1a); and we expect trusters to realise that this is the case, and thus more of them to trust those who gave more in the veiled DG (H1b).

The other hypotheses are inspired by the countersignalling model described in Section 2.2.[49] We expect trustees who gave intermediate amounts in the veiled DG to be the most uneasy about their generosity not being recognised by trusters, so they should be more likely to revise to make sure that trusters see their generosity precisely. By contrast, naturally low givers and naturally high givers will be less likely to revise: the former resign themselves to not being trusted and the latter rely on their generosity showing (H2).

Trustees who give an intermediate amount in the veiled DG and choose to revise should do so to persuade trusters to trust them—by revising and either increasing their giving or re-giving the same amount they gave naturally (H3a). More trustees who gave an intermediate amount in the veiled DG choose to revise in order to persuade trusters to trust them than initially generous or miserly trustees (H3b).

More trusters will send money to naturally low generosity trustees who revise and increase their giving than to naturally low generosity trustees who do not revise (by signalling they look like medium trustees some of the time). Trusters respond to signals from low generosity trustees (H4a). However, low types, 'marked' by their sign, will have to invest more in signalling than medium types to achieve the same level of trust (H4b).

[49] This holds assuming that our subjects' behaviours conform to the model's countersignalling equilibrium predictions. Similar to Powell and Wilson's experiment on Hobbesian jungles (2008) we do not solve our experiment for equilibria.

More trusters will choose to trust medium generosity trustees who revise and increase their giving than medium types who do not revise (H5).

Trusters will *not* treat high type trustees who revise and re-give the same amount differently to high type trustees who do not revise; revising and re-giving the same amount will be irrelevant. If the sign generated naturally by high types suffices for them to be separated often enough from those below them, then supplementing it by re-giving the same amount in the form of a signal—which is a more precise channel of information—will not improve their probability of being trusted as they are already recognised as high types. Eliminating noise that we injected in our experiment from the information will not gain high types benefits (H6a). If the sign of high generosity trustees is sufficiently separating, then trusters will also treat high type trustees who revise and *increase* their giving the same as high types trustees who do not revise because they are already recognised as generous, making signalling wasteful for them (H6b).

2.4 Results

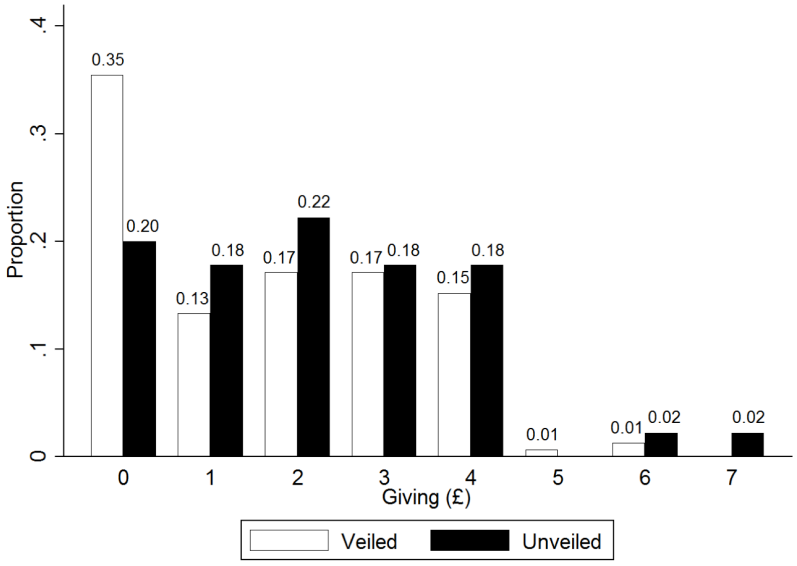
We collected a total of 158 independent decisions by trustees in the veiled DG.[50] The modal giving was £0, although the majority of trustees (63%) gave between £1 and £4 (Figure 2). Our distribution is broadly similar in both mean (21.3%) and spread to that

[50] Only three subjects gave more than 50% of their endowment in the veiled DG—one gave £5 and two gave £6. We remove them from our analysis for a combination of reasons. While those who give up to 50% (£4 in our case) are common in dictator games, those who give more are outliers (Engel, 2011, p. 589). Some authors consider this a sufficient reason to remove subjects from their analysis of DGs as it suggests that they failed to understand the experiment (Pradel, Euler, & Fetchenhauer, 2009; Schotter, Weiss, & Zapater, 1996). Our three outliers display additional indicators in support of this inference. One subject who gave £6 in the veiled DG and revised upwards went on to exhibit incoherent behaviour in the TG, and the other two outliers wrote incorrect or contradictory answers in their ex-post questionnaire. We could have alternatively combined these outliers into a group with trustees who gave £4. Our results are only slightly affected if we do this.

typically found in experimental dictator games (Belot, Duch, & Miller, 2010; Engel, 2011, p. 589).[51, 52, 53, 54]

Comparing the distribution of unveiled giving by revisionists, we find that, relative to veiled decisions, a smaller proportion of revisionists choose to give £0 (20% vs. 35%) and larger proportions choose to give between £1 and £4 (76% vs. 63%). Here too, however, few trustees give £4 (18%).

Figure 2. The amounts given by trustees in the veiled and unveiled dictator games.



H1a. We expect trustees' giving in the veiled DG to be positively correlated with their trustworthiness in the TG.

[51] Our experiment was carried out at the Centre for Experimental Social Science, Nuffield College, Oxford.

[52] We collected data on 158 active subjects and had 35 passive subjects from a total of 7 experimental sessions run over three days from 31/10/12 to 2/11/12. Subjects were graduate and undergraduate students who had not taken part in Gambetta and Przepiorka's (2014) experiment.

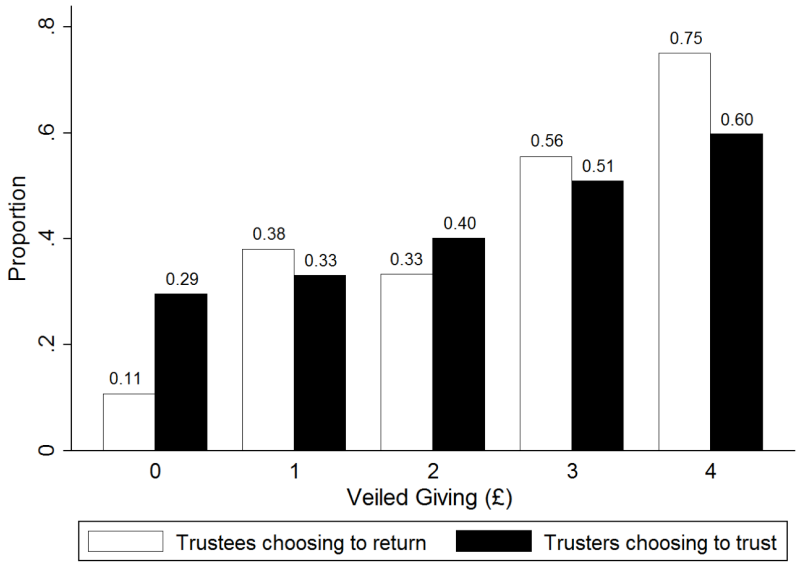
[53] Our trustees had a mean giving of 21.3%, which was slightly lower than that reported by Engel, at 28.4%, in a recent meta-review of 129 papers using dictator games published between 1992 and 2010. We also find more people give £1–3 than that reported by Engel, and no spike in the distribution at 50%.

[54] Belot et al. look at the differences between student and non-student subjects at CESS and finds that students are less other regarding than non-students. As we used students for our experiment, it makes sense that we find a lower mean giving.

As expected, trustees who are naturally more generous are also more trustworthy (r=0.47)[55] (Figure 3). For every one-pound increase in the amount given naturally (relative to £0), the odds of being trustworthy more than double (exp(b)=2.22, χ=5.22, p<0.001).[56]

H1b. We expect trusters to realise that H1a is the case, and thus more of them to trust those who gave more in the veiled DG.

Figure 3. The proportion of trusting trusters and trustworthy trustees by trustees' veiled giving.



Trusters recognise that natural giving is a strong sign of trustworthiness and act accordingly, more of them bestowing trust on naturally generous trustees: the odds of

[55] This result includes both revisionist and non-revisionist natural generosity, removing the former group strengthens the association.

[56] All logistic regressions include control variables for age, sex, and experimental session. The full regression outputs can be found in Appendix A.

being trusted increases by 40% for every pound increase trustees naturally give (Figure 3) ($\exp(b)=1.40$, $z=9.04$, $p<0.001$).[57]

H2. We expect trustees who gave intermediate amounts in the veiled DG to be more likely to revise.

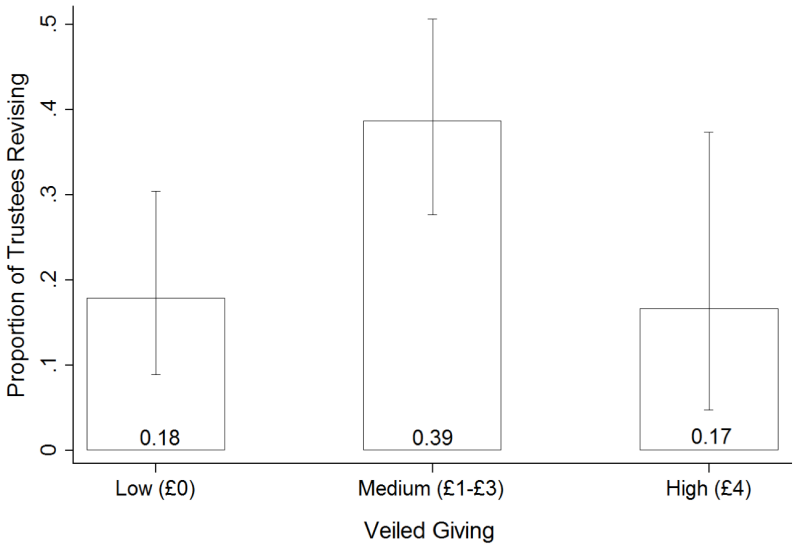
Although only 28% of trustees chose to revise their giving (43 out of 155), those among them who gave an intermediate amount in the veiled DG were the most likely to revise as our hypothesis predicts (Figure 4). Among trustees initially giving £1, £2 or £3, respectively 38%, 41% and 37%, chose to revise. In contrast only 18% and 17% of trustees who initially gave a low (£0) or a high (£4) amount respectively revised. In a logistic regression containing both a linear and a quadratic term for veiled generosity, we find a positive and significant linear coefficient ($p=0.002$) and a significant and negative quadratic coefficient ($p=0.002$). This implies an inverted U-shaped relationship between the amount given when veiled and the probability of revising.

These results reinforce a way of conceptualising subjects' types which rests on clear cut behavioural foundations, and which accurately capture our three types: those perfectly self-interested who give nothing (*low types*); those somewhat other-regarding who give something but less than half of their endowment (*medium types*); and the altruists who give half (*high types*). If we re-analyse the data using these three groups we find, on the basis of a logistic regression, that medium types revise more than low types ($p=0.007$) and more than high types ($p=0.037$). Thus, we find broad support for Hypothesis 2.[58, 59]

[57] The logistic regressions used to test Hypotheses 1b, 4a, 5, 6a and 6b have S.E.s clustered according to subject to account for the multiple observations per subject.

[58] Schniter et al. (2013) report experimental results that mesh well with ours. Their subjects took part in a trust game. Before trusters sent the trustees any money, the trustees had the opportunity to make a non-

Figure 4. The proportion of subjects revising by their veiled giving. Error bars indicate 95% confidence intervals.



H3a. Trustees who gave an intermediate amount in the veiled DG and choose to revise should do so to persuade trusters to trust them.

A slight majority, 52% (15/29) of intermediate type revisionists revise and re-give the same amount again or revise and increase their giving. A decision to revise and re-give the same amount as naturally given we consider as a signal because it removes noise

binding promise to trusters about how much they will return if they are trusted. Once trusters made their investment decision, and the trustees decided how much to return (and whether to honour their promise), it was revealed to participants that they will take part in a second trust game still matched with the same partner. However, before this new trust game began, trustees had another chance to send a non-binding promise. Schniter et al. find that trustees who broke their initial promise subsequently increase how much they promise to send in the second trust game by a greater amount than those who had initially kept their promise (p. 249) and promise-breakers chose to send longer messages than promise-keepers (p. 250); both actions are interpreted as an effort to assuage the concerns of trusters who could see trustees' previous untrustworthy behaviour.

[59] As the trusting behaviour of trusters may act as a proxy for their beliefs about how they will be treated as trustees, in one version of this regression we included a proxy for this: mean trusting behaviour. This had no effect on our findings. In addition, we find a strong link between the trustworthiness of trustees and their trusting behaviour as trusters (r=0.31, p=0.001).

around the giving of trustees; it is a costless signal that purifies the naturally given sign.[60] We also interpret the action of revising and increasing as a signal, this time a costly one. The other 48% of intermediate type revisionists decreased the amount they gave upon revision implying an unexpected—unexpected for they became less other-regarding after unveiling—short-term selfish motivation to revise.

H3b. More trustees who gave an intermediate amount in the veiled DG choose to revise and signal than initially generous or miserly trustees.

When we compare the actions of intermediate types to the actions of low and high types, we find that 100% of low types (10/10) who revised and 25% (1/4) of the high types who revised did so with a signalling motive. From this angle, relative to low types, medium types do not seem to be more likely to revise for a signalling purpose—although they are still more likely to revise than high types. However, this interpretation of our results does not consider the different choice that low and medium types face: the former cannot decrease their investment, while the latter can. This makes a straight comparison between revisionists' decisions problematic.

A better way to view our results is to take into account both the proportion of types who choose to revise and the proportion within each type who revise in order to signal. When we do this, we find that 20% (15/75) of the intermediate types revise and increase their giving or re-give the same amount again relative to what they gave in the veiled DG. Of the low generosity trustees, 18% (10/56) revise and increase or re-give the same amount again; and of the high generosity trustees just one individual revises and gives the same

[60] The questionnaire responses of the revisionists who re-gave the same amount support this interpretation e.g. 'It meant that people could actually see what I had choosen [sic]', 'More opportunity for person to see how much I sent'.

amount (1/24, 4%) while none of them increase. Using one-sided Fisher's test for equality of proportions, we find that medium types do not significantly differ from low types (p=0.47), but significantly differ from high types at the 10% level (p=0.056). Intermediate generosity individuals appear to signal at similar levels to low types, but more than high types (Figure 5 and Table 1).

Figure 5. Proportion of revisionist and non-revisionist trustees by their natural generosity. Error bars indicate 95% confidence intervals.

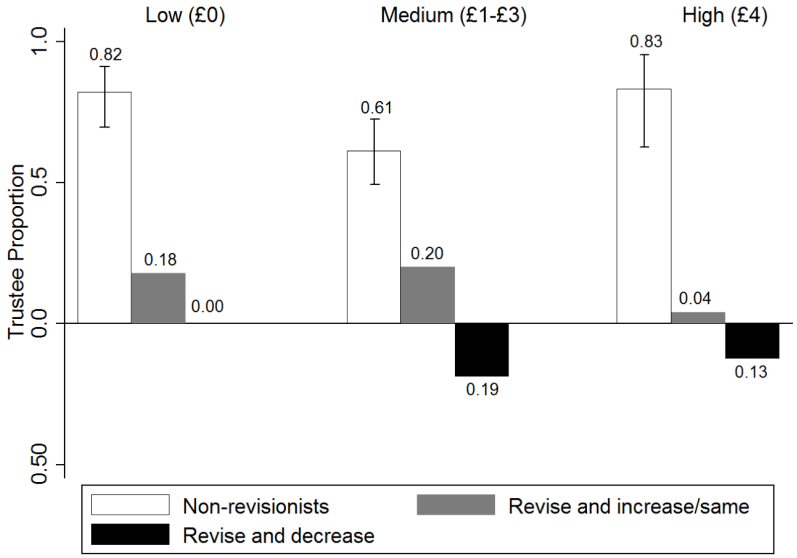


Table 1. Subjects' actions according to their type. The number of subjects is enclosed in parentheses, and row proportions are shown above.

Type	Action				Total
	Not revise	Increase	Same	Decrease	
Low	0.82	0.18	0.00	0.00	1.00
	(46)	(10)	(0)	(0)	(56)
Medium	0.61	0.16	0.04	0.19	1.00
	(46)	(12)	(3)	(14)	(75)
High	0.83	0.00	0.04	0.13	1.00
	(20)	(0)	(1)	(3)	(24)
Total	0.72	0.14	0.03	0.11	1.00
	(112)	(22)	(4)	(17)	(155)

We analyse 2450 decisions by trusters, which gives us an average of 77 observations for each veiled-unveiled combination that we are interested in (Table 2).[61] We obtained more results for trusters than trustees because when eliciting truster behaviour we used a partial variant of the strategy method, in which 19 different veiled–unveiled giving combinations were asked from each of our truster subjects.

When we report or discuss truster behaviour (the rest of this section and Section 2.5.2) and we say that someone gave £X in the veiled DG, we mean that this is the sum the truster sees on the card, regardless of whether someone gave £X, or the noise we introduced makes it seem with 1/3 probability as if they gave £X, while in reality they gave £X±1.

Table 2. The veiled and unveiled giving combinations that we showed to trusters and subsequently analyse. A dash '–' indicates that we did not collect data for this combination.

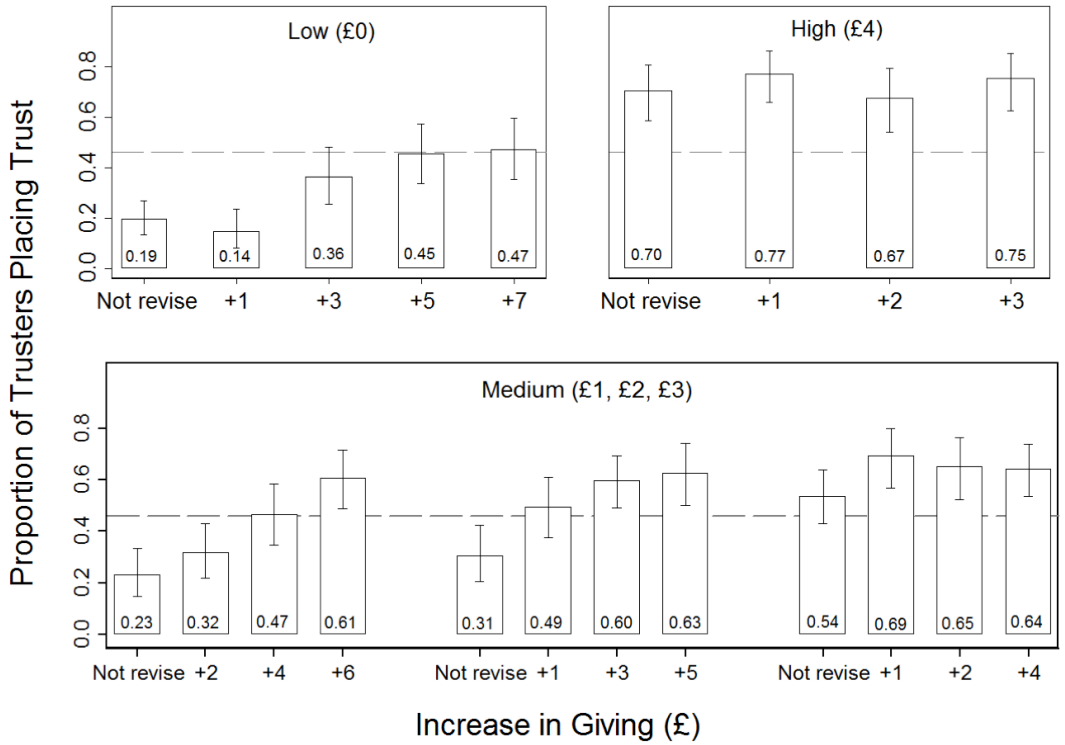
Unveiled giving (£)	Veiled giving (£)					Total
	0	1	2	3	4	
Not revise	140	86	75	99	74	474
0	62	85	83	66	48	344
1	90	80	83	-	-	253
2	-	-	-	92	64	156
3	72	82	77	-	50	281
4	-	-	-	68	57	125
5	73	73	99	66	74	385
6	-	-	-	-	58	58
7	72	76	67	95	64	374
Total	509	482	484	486	489	2,450

[61] We do not analyse all trusters' decisions that we collected. Some trustees chose combinations which were not included in the partial strategy method set of veiled-unveiled combinations which we had planned to elicit, and we do not use trusters' decisions in response to those (even though we had to show them to trusters to work out their payments). We also cap our analysis of trusters' decisions in response to combinations including only a maximum giving of £4. This is because a giving of £5 or more is so rare and would have complicated the analysis with no gain. Merging the responses that trusters give in response to veiled giving of £4 and £5 has no meaningful impact on our results.

H4a. More trusters will send money to naturally low generosity trustees who revise and increase their giving than to naturally low generosity trustees who do not revise.

We find evidence in support of this hypothesis. Using a logistic regression, we find that for every £1 of extra giving upon revision relative to a naturally miserly non-revisionist, the odds of being trusted for those who initially gave £0 increases by 27% (p<0.001). In other words, trusters pay attention to a trustee's strategic generosity even when the trustees initially revealed themselves to be miserly (Figure 6).

Figure 6. The proportion of trusters placing trust according to veiled and unveiled giving combinations presented to them using a partial strategy method. The relative increase in generosity refers to the amount extra given by a revisionist in the unveiled dictator game relative to the amount that revisionist naturally gave in the veiled dictator game. The dashed reference line corresponds to 46%, the proportion trusting those about whom they have no information in Gambetta and Przepiorka (2014). Error bars indicate 95% confidence intervals.



Statements made by low type revisionists (all of whom revised upwards) suggest that they recognise the effect that increasing generosity can have: 8 out of 10 wrote that they revised upwards to give the impression of trustworthiness to trusters (although only 3 out of the 10 were then actually trustworthy!).

H4b. Low types will have to invest more in signalling than medium types to achieve the same level of trust.

As we expected, low types have to invest more in signalling than intermediate types to achieve the same level of trust. For low types, no amount of giving makes those who gave £0 at first more trusted than a stranger—that is someone about whom there is no information (as we explained above, our measure for trust in strangers is taken from Gambetta and Przepiorka (2014)). In contrast, all medium types can surpass the trust a stranger receives (Figure 6).

H5. More trusters will choose to trust medium generosity trustees who revise and increase their giving than medium types who do not revise.

For trustees who naturally gave £1, every one pound increase upon revising corresponded to a 31% increase in the odds of being trusted ($p<0.001$) and for those who gave £2 naturally every £1 increase corresponded to a 29% in the odds of being trusted ($p<0.001$); all relative to a non-revisionist. There was a non-significant gain of revised upward giving for those who gave £3 in the veiled DG ($p=0.161$). Thus, we find support for this hypothesis for trustees whose veiled giving was £1 and £2, but not for those whose giving was £3. Most medium types could increase their probability of being trusted by revising (Figure 6).

Almost half of the medium types who revised upwards explicitly stated that increasing their giving upon revision could make them more trusted (5/12, 42%).

H6a. Trusters will not treat high type trustees who revise and re-give the same amount differently to high type trustee who do not revise.

We find no difference in the probability of being trusted between naturally generous trustees who do not revise and naturally generous trustees who revise and give again the same amount they gave in the veiled DG (p=0.304).

H6b. Trusters will also treat high type trustees who revise and increase their giving the same as high type trustees who do not revise.

Increasing giving once unveiled has no effect on the probability of being trusted for naturally generous trustees (p=0.868). Both results support H6a and H6b, showing that the initially generous have no incentive to revise (Figure 6).

When asked why they chose not to revise when given the possibility, many stated that the reason was because they were 'happy' or 'confident' with their natural decision, and some added that their natural choice should be enough to make them trusted. Two trustees gave reasons of principle '…a choice is a choice is a choice…', while two others made a similar point but did so by denying that the situation had changed for them: 'I could not see any reason for doing so [revising]—nothing had changed since the first decision [veiled giving] was made (in my mind anyway).' This would suggest that

generous trustees expected their decision not to cause them harm, and trusters' responses proved them to be correct.

2.5 Discussion

Generosity turns out to be a formidable indicator of trustworthiness as trustees who naturally give more are substantially more trustworthy: 75% of trustees who gave a high amount returned the money in the TG, while 89% of trustees who gave £0 did not return the money. Trusters recognise that (lack of) natural generosity is a reliable sign of (un)trustworthiness and respond appropriately by rewarding trustees who were more generous and shunning those who were not generous: those who gave nothing naturally were only trusted by 29% of trusters, while those who gave £4 were trusted by 60% of them (Figure 3).[62, 63]

As argued by Gambetta and Przepiorka (2014), whose results match ours, this finding could have far-reaching implications. Given the strong link between generosity and trustworthiness (and meanness and untrustworthiness), acts that reveal one's generosity should be an important source of information about people's pro-social dispositions, and a source with marked advantages: they are easy to observe, easy to produce in everyday life; and, by benefitting others, acts of sharing are, all else equal, more efficient as signals of trustworthiness than wasteful displays in which people 'burn' resources. Spontaneous gestures of sharing (or not) valuable resources could have been functioning as the first

[62] The value of 60% refers to the proportion of the time that trusters choose to pass in encounters with trustees who display a veiled giving of £4 and any level of unveiled giving. Trusters pass more of the time to trustees who have a veiled giving of £4 but who did not revise, or who revised and increased (as can be seen in Figure 6).

[63] It is worth noting that, as can be seen from Figure 6, a staunch 30% of trusters never trusted, irrespective of trustees' veiled and unveiled giving, and that trusters who themselves were untrustworthy when they played as trustees were less likely to place trust—i.e. they were more sceptical—than trusters who had been trustworthy as trustees.

signs of (un)reliability in cooperation in the course of human evolution, and could have 'spread through positive assortment'. The information value of generosity might also explain the ubiquity, across a wide variety of cultures, of ritualised gift-exchange and sharing ceremonies.

2.5.1 Trustees' decisions

On discovering that they will play a TG and that the DG choice they made under the veil of ignorance will be revealed to trusters, despite the noisiness we injected, most trustees (72%) felt no need to revise the split they chose naturally. Many of them seem to believe that trusters would not be convinced by a belated signal of generosity; as one participant said, 'to increase the amount given would not be a guarantee that this would prompt the receiver in Part 3 [the TG] to give more back. A revision might make it obvious that the sender is only giving in order to receive—a cynical attitude which might not inspire generosity from the recipient.'

We found however marked differences in the proportion of trustees who choose to revise. As we expected, fewer of those at the extremes revised, appearing either content or resigned. Those who gave £4 seem confident that their natural action should suffice to persuade trusters to trust them, confirming our expectation that naturally generous trustees have no incentive to increase giving.[64] Those who gave nothing seem to believe either that it would be pointless to be more generous, for savvy trusters would doubt their motives, or expect, rightly, that the cost necessary to convince a truster is simply too high to be worth it.

[64] Very few naturally generous trustees chose to revise (4/24), and almost all who did so chose to decrease their giving in the unveiled DG (3/4). Only a lone subject chose an action that suggests a motive of communication, by re-giving the same amount he or she had naturally given—a costless signal that removes the noise.

As expected, it is medium types who revise the most. Once we take into account both the proportion of each type who revise *and* the proportion who take a communicative action, we find that medium types, while signalling in similar proportions to low types, signal more than high types who hardly bother to revise at all. The first choice of medium types does not conclusively show them to be generous and the possibility of looking even worse, given the uncertainty we introduce in the information, could prevent trusters from trusting them.

We had two unexpected results. First, just under a fifth of low types (18%) were not at all resigned and chose to revise. They all increased their giving of course,[65] 8 out of 10 mentioning that they did so to induce trusters to trust them in the TG (more on them below). Second, a considerable proportion of medium type revisionists (48%) *decreased* their investment—as Feltovich et al. find, 'too few' individuals of our medium types signal. It is as if discovering that they are in a strategic situation destroys their initial modest generosity, encouraging them to go for immediate gain. Perhaps they are pessimistic about the effect that a belated increase would have on trusters and decide to make the most of it while they can. Several of them say that this was the reason why they revised downward—one remarking: 'I was not sure A person [the truster] would give me any money even if I increase (as the stakes for A are very different) while this [giving in

[65] Given the design the low type revisionists could not decrease. But none of them re-gave what they naturally chose either: removing noise around their miserliness would not inspire more trust, and besides, it would obviate the possibility that they are confused with medium types, which can happen when they rely simply on their signs.

the revised DG] was a sure amount to keep.'[66] Together, these go against a prediction of countersignalling theory; here the middle do not shout louder than the lows.

Whether revising in order to lower their giving and go for short-term gain, or to restate or increase their giving to influence trusters' beliefs, medium types appear to be the group most 'uneasy' about their experimental past. On the one hand, they appear as relatively more fretful to have their qualities recognised by trusters—which may correspond to the lack of calm that, casual observation suggests, is displayed in real life by 'intermediate' agents. On the other hand they seem a little conflicted normatively—first they show some generosity, though not going for the 'fair' even split, but then seem to repent and revert to more greedy behaviour.

2.5.2 Trusters' decisions

As we move to discuss trusters' behaviour, there are two points to bear in mind:

- As mentioned before, the reference point which we use to evaluate the level of trust bestowed on trustees about whom trusters have no information is that of Gambetta and Przepiorka (2014), in which 46% of trusters trusted trustees about whom they knew nothing.
- We used the strategy method hence trusters' responses are hypothetical and the combinations to which they responded were not always chosen in reality by trustees. Using this method we collected at least 48 observations for each of our veiled–unveiled combinations.

[66] Others made no mention of the impact their action would have on trusters' decisions, suggesting some failed to realise that by sending a more generous signal they could induce trusters to trust them and increase their gain in the TG.

Trusters overall show a remarkable appreciation of the information they have on trustees.

They take a dim view of trustees who give 0 in the veiled DG and do not revise, trusting only 19% of them. Trusters respond somewhat positively but very prudently to those 18% in the low type group who revise and increase their giving. While giving £1 is a waste as it gains trustees no more trust (14%), giving £3 or £5 in the second DG makes them a little more trusted at 36% and 45% respectively. Trusters therefore seem to interpret unveiled DG giving as a signal of relative generosity even when the natural DG giving was zero. Yet, the additional expenditure does not push trust up by much and seems hardly worth it.[67] Trust in them is 'capped' as at most they can expect to be trusted in 47% of the interactions with trusters, and this only if they pour 7 additional pounds into their signal; this level of trust is nearly identical to the reference level of 46% bestowed upon strangers, the most redemption low types can hope for as their past actions brands them. Miserly trustees who revised and increased their giving are, it seems, overly optimistic about their chances of being trusted and do not appreciate the 'stigma' that the sign they naturally emitted placed on them. It may be a trait of the self-interested to underestimate the negative impact their behaviour has on others, and to think that their image can be quickly brightened up again.

Trusters respond more trustingly to the signals of those who gave an intermediate amount in the veiled DG and chose to revise upward. For £1 and £2 natural givers each

[67] Establishing precisely whether a choice paid off is difficult with our experiment because we did not formally model subjects' costs and benefits. Still we can propose some tentative conclusions about the rationality of trustees' behaviour using the changes in the probability of being trusted they receive according to deviations away from their natural level of generosity. While we cannot say which strategy was best for all trustees in an absolute sense using this logic, what we can say is that a certain strategy made more sense for one group of trustees than it did for others.

additional pound makes a substantial difference in their probability of being trusted; this is most apparent for the £2 givers, as, relative to not revising, a single pound increase pushes the probability that they will be trusted up from 31% to 49%—for them a £1 increase gains more in probability of being trusted than a £3 increase gains for those who naturally gave £0. The gain does not materialise, however, for those who gave £3 first and then revised upward later. One interpretation is that this group see themselves as falling in the intermediate group, and thus in need of sending a stronger message of generosity and hence trustworthiness. Yet, trusters already seem to perceive £3 givers as being on the borderline of the generous side—not as starkly as the £4 givers (who are trusted 60% of the times), but enough to be deserving of a modicum of trust (51%).

Although overall medium type trustees seemed to have gained significant additional trust by revising and increasing, most individuals in this group chose not to revise (61%). This is puzzling for not revising for these trustees would seem a dangerous strategy: if they rely on their sign as the sole source of information, 1/6th of the times they are perceived, due to the noise we introduced, as having given only £2 naturally. This corresponds to a large decrease in the probability of a truster bestowing trust: only 31% of trusters bestow trust on a trustee who gave £2 naturally and did not revise, while 54% bestow trust in a trustee who gave £3 naturally and did not revise. One could imagine that those at the top of the medium type group who naturally gave £3 anticipated (correctly as it turns out) the futility for them of revising and increasing. This does not seem to be the case however as the revision rate of this sub-group is very close to the rate of the rest of medium types.

But there is a subtler reason for not revising: consider that 1/6th of the time these trustees who gave £3 naturally would be mistaken for trustees who gave £4. By

mimicking the strategy of the truly generous and doing nothing, they can be treated as one. At least one trustee, who gave £3 naturally, followed this strategic reasoning in an attempt to persuade trusters that he was a high type:

> In part 1 [veiled dictator game], I chose to give the Recipient £3. I assumed that if I didn't choose to revise in part 2 [unveiled dictator game], then the A person [truster] would never know how much exactly did I give to the Recipient. Even if he clicked on a card with £3 underneath, he would probably assume that he had chosen to reveal a card that was showing £1 less than the real value; the fact that I hadn't chosen to revise would have made an overall good impression, having person A to think that I have probably had made the fairest possible choice, handing down £4. Thus it would make an impression of giving out £1 more than I actually did, and would be able to save myself £1 what I might have lost, had I chosen to revise my decision in part 2.

Trusters reward all trustees who display a natural giving of £4 in the same way, regardless of whether they signal or not. Most of the naturally generous trustees chose not to revise (83%), as our sophisticated participant above expected, and they appear to have got it right. Using an additional precise source of information did not produce benefits for the naturally generous. Trustees here seem to clash against the ceiling of unconditionally distrustful trusters, which leaves little room for improvement. Decreasing though is certainly not appreciated by trusters: a £2 reduction results in trust sinking to a measly 23%, well below the level achieved with no communication.

To summarise, we find that:

a) generosity functions as a sign of trustworthiness and trusters recognise this;

b) trustees who naturally give an intermediate amount revise the most frequently;

c) a majority of intermediate type revisionists increase their giving, and while more intermediate types signal than high types they signal at similar levels to low types;

d) naturally miserly trustees can increase their probability of being trusted by using a costly signal but only to the level in which strangers are trusted;
e) intermediately generous trustees can make substantial gains by signalling and increasing their giving; and
f) naturally generous trustees cannot gain by utilising a signal of the same intensity as their sign, or even from increasing their signal.

How do our findings fit with countersignalling theory? We found that low type trustees decided to revise more than predicted, and medium types revised less than predicted. Still, more medium types revised than both low and high types, although medium types signalled at similar levels to low types—in contrast to the theory's expectations. For both low and medium types, signalling by revising and increasing giving was effective, as countersignalling predicted, but for the former it was too expensive to be affordable, for the latter it is likely to have been worthwhile to signal. Regarding the behaviour of high types, most conformed to the predictions of the model and did not revise—had they revised and signalled they would not have gained further trust.

We set out to answer the question, when people realise that a past behaviour of theirs provides information about them to others, and that this may have consequences for them, good or bad, what do they do? The answer for a majority of subjects is simply 'nothing'. Others do decide to take action: to either remedy negative information receivers have about them, or to cut their losses, decrease their generosity and gain more in the short term. Yet what people do varies by their conditions. Individuals who naturally behave in a generous way, and consequently emit the 'right kind' of sign primarily rely on their sign to get their message across. They eschew the use of a strategic signal. Rightly so, as giving the same as their preferred level but through signalling is

wasteful in terms of time and energy, while giving more than what they prefer is wasteful also in monetary terms. Others, who behave in a naturally miserly fashion also mainly avoid the use of a signal. For them, it is costly—they prefer to give nothing—and their gains to doing so are capped at the level of a stranger. For those who give an intermediate amount naturally, signalling seems to be a profitable strategy as they can be perceived as either trustworthy or untrustworthy, and so a proportion of these revise in an effort to be trusted.

3. Information and Violence in Prison

Why do prisoners fight? A recent theory proposes that prisoners mainly fight because they *lack information* about one another's abilities to employ violence (Gambetta, 2009a, pp. 78–110). Using data from the Survey of Inmates in State and Federal Correctional Facilities 1997 (SISFCF) on nearly ten thousand prisoners incarcerated in state facilities and three thousand prisoners incarcerated in federal facilities, I test two hypotheses derived from the informational theory of prison violence.

Prisoners' misconduct behaviours are usually understood using deprivation theory or importation theory. Both hold that inmates are assimilated into prison culture, known as prisonization (Clemmer, 1940), and follow the behavioural rules—such as the norms, mores, and conventions—that are prescribed or proscribed within it (Dobbs & Waid, 2004). Where prison culture originates from is the main disagreement between the two. Deprivation theory posits that prison culture, and inmate behaviour, are primarily functions of the prison environment: prisoners adapt to, and generate a culture, in response to the constrained conditions, the 'pains of imprisonment', that they face inside the facilities (Sykes, 1958). In contrast, importation theory argues that it is not the prison environment, and the deprivation associated with it, that are the main factors affecting prisoner behaviour, rather, it is the behavioural patterns, and culture, acquired by individuals outside of prison, and that they bring with them upon conviction, that determines the culture of prisons (Irwin & Cressey, 1962).

These two traditional theories de-emphasise prisoners' agency and decision-making and focus on culture and values as explanation for violence. Yet, qualitative studies describe

that prisoners carefully consider their actions and the outcomes that they would receive from them. Marek Kaminski, who spent time imprisoned, writes, '[t]hat inmates constantly calculate and re-calculate payoffs in numerous gaming contexts is common knowledge in prison' (2003, p. 190). He meticulously records and analyses, using simple game theoretic models, the decisions that prisoners there undertake and the incentives that they face (2003, 2004).

Moreover, deprivation theory has little to say about which inmates' characteristics influence their involvement in violence. And, by itself, importation theory has little predictive clout, for it has to be supplemented by explicit or implicit auxiliary theories that specify how the pre-incarceration characteristics affect prisoners' values and subsequently violence.

Gambetta recently proposed a theory that places the decisions of prisoners at its centre and incisively specifies the conditions under which prisoners can be expected to fight. He combines specific features of prison life, the scarcity of resources and uncertainty among prisoners, and prisoners' preferences, to derive predictions regarding the links between prison violence, communication, and information (2009a, pp. 78–110).

Like Sykes's famous 'pains of imprisonment', Gambetta identifies the loss of resources as a crucial structural factor: inmates can only obtain many of the resources that they value—such as food, money, space, and privacy—in very limited amounts.[68] As there is

[68] One inmate interview by Hans Toch (1977, p. 149), says about the absence of privacy and space in prison: 'Sometimes there is [sic] twenty or thirty people in the showers, and they're always making remarks to you, and you don't feel free… Even though the physical pressure is there for a short time, the mental pressure is there permanently. And if you're on the toilet and everybody is just walking by, then it's really an intense thing. I've been to the bathroom in front of people on the streets, and it's just nothing at all like

incomplete control over the allocation of these resources—the prison authorities for instance cannot always enforce property rights—they will be distributed, at least some of the time, according to the ability of individual inmates to employ violence. The other important structural factor is that prisoners cannot choose the people with whom they live and associate, and are thus forced to interact with others about whom they know very little, creating uncertainty in those interactions.

The preferences of prisoners, which are simple, and 'uniformly recorded by all research' (Gambetta, 2009a, p. 80) are that they prefer to:

(1) have more of the limited resources rather than fewer;

(2) 'have their property rights and prerogatives respected by other inmates,' (p. 80) but;

(3) violate the rights of other prisoners and take their resources instead of going without them; and

(4) avoid fighting. For weaker inmates, who expect to lose in a fight, this is clear, but even for stronger inmates expecting to win, serious injury may occur in any struggle.

From these preferences and features of prison life, it follows that whenever it is possible for them to do so, prisoners try to form a hierarchy of violence potential—their ability to employ violence—using the *information* that they possess about one another's prowess. When possible, they should take communicative acts that are less costly than violence, but are reliable indicators of it, to display violence potential rather than resorting to fighting. Whether a would-be attacker is likelier to win, or a potential defender is likely to do so, one of the interactants has an incentive to capitulate and avoid a fight. In the case of the former, the potential defender backs down and gives in, while in the latter it is the

---

it is in here. You just want to say, "Jesus, leave me alone." But you can't close the door, and it's a cell, and people are looking at you…'

reverse. In this way, the cost of violence is avoided, while the available resources are apportioned according to peoples' ability to employ violence such that no one can be better off by unilaterally changing their actions. Conversely, when relevant information is absent, violence prevails. When it is unclear who would win and who would lose, prisoners can have incentives to find out by fighting; if they estimate a sufficiently high probability of success they may prefer to risk the cost of fighting to obtain resources.

Information about violence potential may be lacking because people do not possess relevant manifesta that they can use to signal with, and out of which observers can extract information and infer whether they are 'tough guys' or meek. The lack of raw materials prevents prisoners from communicating using reliable signals, and, inhibits the ability of other prisoners to acquire information from their signs. Such prisoners, without signs and signals of their violence potential, lack *violence capital*. The less violence capital that a prisoner possesses, whether it indicates toughness or meekness, the harder it is for others to identify that prisoner's place in the prison hierarchy, increasing the likelihood of violence. Violence capital is a resource that allows prisoners to avoid violence.

Prisoners who possess myriad indicators of their past violent behaviour—an extensive criminal record, gang tattoos, and scarring, to name a few—and those who walk effeminately and have a reputation for meekness are less likely to fight than those with few manifesta to display. Irrespective of whether they are considered tough or weak, their position in the hierarchy is clear. Tough guys are left alone because the cost to challenging them, and getting into a fight are high. As a prisoner in Fort Leavenworth puts it, 'When guys like a Bucklew or a Bowles get into a fight with another inmate, it only lasts a few seconds, 'cause they are going to rip out a guy's windpipe with their first blow and rip off his nuts with their second' (Earley, 1992, p. 294). Weaker prisoners are

exploited, but, once dominance has been established over them, there is no need to use violence. These prisoners too will avoid fighting—but potentially not exploitation—if their position is clear. In contrast, prisoners with few indicators of their position in the hierarchy, about whom there is greater uncertainty, will be challenged, and some of these will defend themselves, leading to a fight.

Thus, I expect that prisoners with greater violence capital will be involved in fewer fights, for their position in the hierarchy is clearer, than those who possess less violence capital (H1).

Violence capital is distinct from the ability of prisoners to use violence, their violence potential. A prisoner who is meek and low in fighting ability can build up high violence capital by generating signs and signals that identify his low fighting capability—the violence capital of this prisoner indicates that he is near the bottom of the hierarchy. Conversely, a fearsome inmate may be low in violence capital, waiting to be uncovered. While it may seem unintuitive to call people who possess signs and signals indicating their lack of toughness as high in violence capital, the important point is that they possess manifesta that allows others to identify their violence potential. Both the meek and the fearsome may be low or high in their relevant displayable features that influence their likelihood of fighting.

Information may also be lacking because the display, transmission, or perception of one's own and others' violence potentials is inhibited. Prisoners may be made to wear baggy uniforms with high collars and long sleeves that mask scars, tattoos, and physical build. They may be placed into single cells—or share only with one another inmate—they may have severe restricts, typically in higher security category facilities, on the range

and time that they can move within a facility, and prisoners' media access may be limited, with few available newspapers and books, or no television. In these cases, the *spread of information* is hampered.

Individuals who receive more information, and about whom more information is spread, end up in fewer fights for there is less ambiguity about their rank (H2).

Child molesters are an exception to the information theory. If a prisoner is found out, or even suspected, to be a child molester, he or she may be killed, or severely beaten (e.g. Earley, 1992). A greater availability of information in this case leads to more violence. Still, prison authorities often separate these inmates, and in my analyses, I test a control variable for prisoners who are convicted for such crimes.

Fighting from this perspective can be understood as either (or simultaneously both) the breakdown of non-violent communication—a failure to resolve hierarchy without force—or as a communicative act that generates reliable information about violence potential. In this latter interpretation, fighting builds information. Both interpretations are consistent with the informational theory and derived hypotheses: if fighting is used for communication, it is reasonable to expect that it is a last-resort display used once cheaper signs and signals have failed or are unavailable.

Nevertheless, one may wonder why, if fighting builds highly reliable information about violence potential should prisoners be involved in violence more than once? A single fight builds information about one's willingness to use violence and to a degree, one's ability to employ it. In a two-person fight, call the participants inmate *A* and inmate *B*, it reveals, with some noise, which inmate out of the pair is better in using it; whether *A* or

B is higher in violence potential. It reveals with much less precision how *A* or *B* would fare against another, uninvolved inmate, *C*. Even if the ranking between *A* and *B* is clear, *A* and *C* (or *B* and *C*) may still have reason to fight to establish their relative positions. The more information that *A* and *B* possess, built from fighting or previous signs, the more easily the outcome between *A* and *C*, or *B* and *C*, is decided without violence. So, one fight, at least for newcomers with low violence capital, is unlikely to be enough to definitively establish their positions in prison and multiple fights may occur before enough information is built up to assuage its use.

Quantitative studies on prison violence usually employ one or more of the traditional theories (e.g. Dhami, Ayton, & Loewenstein, 2007; Gover, Pérez, & Jennings, 2008; Jiang & Fisher-Giorlando, 2002). Indeed, the only empirical research that has set out with the informational theory in mind is Gambetta's chapter *Why Prisoners Fight (and Signal)* in which he also constructs and proposes the theory. In it, he compiles qualitative and quantitative data from a range sources to test predictions that are derived from his theory and finds broad support for the effect of information on prison violence. While the evidence he compiles is very promising, the results rest on qualitative data and simple bivariate analyses, and multivariate regression techniques to control for confounders are not used, making the conclusions, in a sense, provisional.

Apart from Gambetta, the only author to theorise about and explore the links between strategic prisoner behaviour and information is Kaminski (2003, p. 192). He focuses his attention on organised tests and initiation rituals that experienced prisoners set up to sort rookies according to their toughness and cleverness and uses screening models to explain such behaviour (Kaminski, 2003, 2004). These are a different but closely related class of models that capture the converse situation to signalling, in which receivers move before

senders, taking an action that will separate between the different types of senders (Rothschild & Stiglitz, 1976). Gambetta's framework is more general, and I focus on it for this reason.

3.1 Data and measures

My data come from the 1997 Survey of Inmates in State and Federal Correctional Facilities.[69] This contains information on a nationally representative sample of U.S. inmates held in state and federal prisons that were collected using a two-stage sampling process such that prisons were first randomly selected, and then prisoners from the selected prisons were randomly chosen for interviews. Data collectors, using 'computer-assisted personal interviewing', interviewed prisoners and each interview took about one hour to complete (United States Department of Justice, 1997).

I restrict my analyses to male prisoners in male only facilities, and thus remove facilities that house female prisoners and a combination of male and female prisoners from the sample. Female only prisons may be different to male only prisons in a multitude of ways that are difficult to capture and control for statistically and should be given their full attention in separate analyses. Female prisoners, for example, are known to form 'pseudo-families' with one another, a social pattern typically not found among male prisoners (Huggins, Capeheart, & Newman, 2006). Similarly, the potential for interaction between men and women may lead to complex differences, relative to male only institutions, that are difficult to capture and control. Separating male from female and

[69] Although there is a more recent, 2004, version of the SISFCF, I do not use it. This is because a number of crucial variables (security category of facility, region that the facility is located in) are not present or only available by requesting access and subsequently adhering to stringent, and for me impractical, conditions that are set out under a Restricted Data User Agreement.

mixed facilities is a common strategy in the literature (Berg & DeLisi, 2006, p. 636; Huebner, 2003, p. 109; Steiner & Wooldredge, 2009b, pp. 467, 468, 2013, p. 1238).

Out of the 1,278 facilities that comprise the male state prison universe, 220 (17.2%) were selected for sampling, all of which participated in the survey. From the 220 sampled male prisons, 12,269 prisoners were chosen for sampling, and 11,344 actually participated, leading to a near-perfect participation rate of 92.4% for male prisoners. The federal facility universe encompasses 113 male prisons, 32 (28.3%) of which were selected for sampling and subsequently participated. 3,525 male prisoners were chosen for sampling and 3,173 participated in the interviews—a high 90%. Following the removal of male inmates housed in mixed-sex facilities, I am left with 10,768 inmates in state and 3,173 in federal prisons (no male inmates in the federal sample are housed in mixed-sex facilities). Due to the impressive participation rate of prisoners housed in state and federal facilities, 92.4% and 90% respectively, differential selection into the sample is unlikely to be a major problem.

I further restrict the sample to include only prisoners who have information on the dependent variable and all of the independent variables in the final model specification. Unlike many other data sources regarding prisons, the SISFCF are astonishingly complete, and consequently, following restriction I am left with 9,911 prisoners in state facilities—92% of the male only state sample (9911/10768)—and 2,721 prisoners in federal facilities—or 86% of the male only federal sample (2721/3173).[70] Except one variable, indicating the amount of time left to serve for inmates in federal facilities, none of the variables included in the final models had missing values that comprise a

[70] The prisoners in the sample are incarcerated and not on parole or probation at the time of the interview.

substantive percentage (more than about 3-4%) of the sample and are thus not considered as substantial sources of bias. The amount of time left to serve, had 7.6% missing values, however, cross-tabulations indicate that these prisoners are similar on the independent variables to the ones that remain for analysis.[71]

I analyse data on prisoners in both state and federal facilities, however as there may be substantial and complex differences between these two, I analyse them separately. The basic division between prisoners in the two is that prisoners housed in the federal facilities are guilty of breaking federal law, and those housed in state facilities are guilty of violating state law.[72] In practice, a smaller proportion of the federal population is comprised of violent criminals than the state population (23.8% vs. 47.9%), more of the federal prisoners are sentenced for drug related offences (37.4% vs. 20.6%), and overall, state facilities house prisoners with a wider array of offences of conviction. In terms of institutional structure, federal institutions have a greater number of security categories, minimum, low, medium, high, than state facilities which are classified into maximum, medium, and low security, which may allow finer-grained allocation of prisoners in federal facilities than in state ones.

3.1.1 Fighting as the dependent variable

My dependent variable is the incidence of fighting. This is extracted from two questions, one that asks each prisoner whether, 'Since your admission, have you been in any fights or been hit or punched?' and the other that asks 'How many times?' if the answer is

[71] Inspection reveals that the data I am left with are similar to comparable data used by other researchers analysing the same survey—both same and different waves (Huebner, 2003, p. 115; Steiner & Wooldredge, 2009a, p. 446).

[72] As there is much overlap between state and federal law, defendants can often be tried for breaching either, and decisions regarding which avenue to pursue are often made by the investigators and the prosecutors of a case (Wright, 2006).

'yes'.[73] By combining the answers to these two questions, I obtain a count variable that indicates the number of times that an inmate acknowledges that he has been in a fight or been hit or punched since incarceration. Someone to whom this has never occurred receives a score of zero, someone to whom it has happened to once receives a score of one, and so on.

About a quarter of inmates (27.3%; 2,709/9,911) in state facilities report having been in a fight at some point during incarceration, and, on average, they engage in 1.27 fights. Excluding prisoners who were never in a fight, the 2,709 prisoners who were in at least one fight, get into a substantial number of them at 4.66 fights on average. The fights that occur often have substantial consequences, as many prisoners who were in a fight also report receiving an injury from it (31.4%; 851/2,708).[74, 75] Fewer prisoners report fighting in federal facilities than in state ones: only 11.5% (314/2,721) have ever been in a fight, with an average of 0.33 fights, and even among those who have fought at least once, the average number of fights is lower at 2.89. Moreover, a slightly smaller percentage of prisoners who fight also end up with injuries 27.4% (86/314). It seems that incarceration in federal facilities is safer than in state ones.

[73] The 'admission' mentioned in this question refers to the prison system as a whole and not only their current facility (confirmed in personal communication from the Bureau of Justice Statistics). Thus, it is conceivable that the fights that prisoners report here occurred in prisons other than the one they are interview in, and the only one that the survey has data on. This could lead to the problem of misclassification; a problem also noted by Kuanliang and Sorensen (2008, p. 34) and Steiner and Wooldredge (2009a, pp. 456–457). Unfortunately, this issue is not possible to solve or estimate, because, outside of the limited instances of cross-state transfers that are recorded, no data are kept about the transfer of prisoners that occur between facilities that are within the same state. Still, multiple papers use this dataset to investigate misconduct (e.g. Huebner, 2003; Jiang & Winfree, 2006; Meade, 2012; Steiner & Wooldredge, 2008, 2009a; Sturgis, 2010).

[74] There are slight discrepancies in the numbers between the different comparisons as I do not restrict my sample to exclude prisoners who do not provide a useable response to the questions asking about (i) their injuries and (ii) whether they were found guilty of assault or not. Excluding these would unnecessarily restrict the sample.

[75] People who are injured in a fight may be more likely to remember that fight and subsequently report it, so it is likely that this proportion is a high-end estimate.

This dependent variable does not discriminate between cases in which one prisoner attacks another and those in which a prisoner is attacked. Both attackers and defenders will be given a number greater than zero in the variable, as, irrespective of whether they are attacked or do the attacking, they were in a fight, or have been hit or punched. This does not present a problem for the hypotheses under test here because prisoners with higher violence potentials, and those who are in places in which information spread is fast and broad, can be expected to attack and be attacked less frequently than prisoners with low violence potentials and in limited information spread institutions. When more information is available, the outcome of a fight is clearer, so one of the potential combatants is likelier to back down. When a target has more information about the likelihood of a potential aggressor winning, either he backs down to the advances, or the potential aggressor avoids challenging in the first place. When a potential attacker has more information about a defender, either the attacker backs down, or the defender accedes to his advances. If dominance over someone is known, there is no need to use violence.

I chose not to use an alternative dependent variable concerning the number of institutionally punished instances of violence. This could be extracted from the two questions, 'Since your admission, were you found guilty of physical assault on another inmate?' and if 'yes', 'How many times?' in the survey. In contrast to the measure that I use, which corresponds to self-reported *self-recognised* instances of violence, this variable would indicate self-reported *officially punished* instances of violence.

There are a number of reasons why I decided not to use officially punished assault. One fundamental issue that permeates such data is that disciplinary action for assault is under the discretion of prison guards and the prison administration, potentially introducing

biases into it. Most obviously, black and Hispanic inmates may be more likely to be punished than white non-Hispanic inmates for the same offence. Consistent with this conjecture, Poole and Regoli (1980) find that black and white inmates report equal levels of rule-breaking but that black inmates are more likely to be cited for misconduct. In addition, highly educated inmates may be able to argue their case for innocence more eloquently than their less educated fellows leading to fewer punished assaults among the former. Prisoners who share sleeping quarters with more others have less privacy, thus violence could have a greater possibility of being observed by other prisoners who report them. Physically more imposing inmates may be assumed to be the aggressors of interactions and more likely to be punished than smaller inmates based on identical evidence. The problem with using officially punished violence is that the data are refracted through the personal and institutional lens.

Another factor that could substantially bias the official data on assault is that different inmates are more or less willing than others to report victimisation to officials. Some prisoners will be ready to report that they have been assaulted, and others less so, differentially affecting the levels of officially reported assault. As Bottoms writes, these are the two main filters—the willingness of inmates to report violence and the willingness of staff to address violence—through which information about violence travels before it becomes officially recognised (1999, p. 222).

A further advantage of using self-recognised cases is that they are likely to be more comprehensive records of violence than those reported in official statistics, allowing greater statistical power. The latter should generally be a subset of the former, so, I expect that self-reported and self-recognised instances of violence are greater in number than institutionally recognised instances. Only 14.2% (1,408/9,895) of inmates in state

institutions report being punished for assaulting another inmate and most of the inmates who were in a fight were not institutionally punished (56.6%; 1,531/2,704). A paltry 5.6% (151/2,711) of inmates in federal facilities were written up for or found guilty of assault, and most prisoners, 61.3% (192/313) who were in a fight were never written up for it.[76]

A potential disadvantage to using my dependent variable is that it does not specifically ask whether the violence that the prisoner has been involved in was with another prisoner or a staff member. Although it is conceivable that some prisoners who report being hit or punched are reporting being hit or punched by staff, I do not think that it is a fundamental issue.

One reason for this is that the amount of inmate-on-inmate violence is likely to eclipse the amount staff-on-inmate assaults. Many more interactions occur among prisoners than do so between guards and prisoners, for there are more prisoners than staff and prisoners spend more time together in close proximity, so there is a greater window of opportunity for violence between inmates to happen. In addition, prisoners should be more likely to report assaults by guards than they are to report assaults by other inmates: there is no, or little, stigma attached to reporting staff. Expecting this, prison guards avoid violence; or at least hitting, punching, or fighting that can easily leave marks and that are captured by the dependent variable.

[76] There are two possible reasons for this. One is that prison authorities miss much of the violence that occurs—an inevitable aspect of running such an institution. The other is based on selective punishment by prison authorities. They try to punish aggressors and violent retaliators and avoid punishing victims who do not fight back. In this case, it could be that the 56.6% and 61.3% who are not institutionally punished are the passive victims of violence.

Another reason is that the staff-inmate assault would have to be differentially distributed in the direction predicted by my hypotheses to bias the data away from the null hypotheses. For instance, black and Hispanic prisoners should be assaulted less by staff than white inmates or the less educated than the more educated, both counterintuitive suppositions. A meaningful empirical investigation of the differences between self-recognised and officially punished instances of violence would require a substantial detour away from the aim of this chapter and I leave this for future work to pursue.

3.1.2 Violence capital variables

To test the violence capital hypothesis (inmates with lower violence capital fight more) I use six variables: age, prior incarcerations, education, race and ethnicity, incarceration for murder, and body mass index. I check an additional explanatory variable that is prior incarcerations interacted with time served (Table 3).

Age, in years, is a continuous variable. In the SISFCF, information about the number of prior incarcerations of prisoners is broken down into the categories of 'no previous incarcerations', 'one prior incarceration', 'two prior incarcerations', 'three to five incarcerations', 'six to ten incarcerations', and 'eleven or more incarcerations'. A negative binomial regression with number of fights as the outcome and the number of prior incarcerations as independent dummy variables shows that each has a similar effect on the outcome, implying a linear relationship. Based on this, I recode and use prior incarcerations as a continuous variable. Education is a continuous variable indicating the number of years spent in education. Race and ethnicity is a categorical variable with the groups White non-Hispanic, Black non-Hispanic, other non-Hispanic, and Hispanic. Being incarcerated for murder is a dichotomous variable. Body mass index is calculated

from information that prisoners report on their height and weight, and log natural transformed.

Table 3. The descriptive statistics for the variables used to test the violence capital hypothesis. Means or proportions and standard deviations in parentheses displayed. Potential explanatory variable is enclosed in square brackets.

Variables	State facilities	Federal facilities
Age	33.33 (9.75)	37.62 (10.76)
Number of prior incarcerations	1.29 (1.46)	1.00 (1.36)
[Number of prior incarcerations * time served]	4.32 (5.24)	3.65 (5.21)
Years in education	10.67 (2.35)	11.55 (3.06)
Race/ethnicity		
White	0.33 (0.47)	0.35 (0.48)
Black non-Hispanic	0.48 (0.50)	0.37 (0.48)
Hispanic	0.17 (0.37)	0.23 (0.42)
Other non-Hispanic	0.03 (0.17)	0.06 (0.24)
Murder	0.12 (0.32)	0.02 (0.12)
BMI (log)	3.27 (0.15)	3.29 (0.14)
Observations (facilities)	207	32
Observations (inmates)	9911	2721

In the following, I outline the reasons for using these variables to test the violence capital hypothesis, and where available, I describe relevant findings in the literature concerning these variables and prison violence.

Age provides more opportunity for peoples' violence potentials to come to the surface. Whether this indicates violence prowess, or a lack of it, the behavioural tendency of people becomes more perceivable with their age, as their habits and actions leave marks

on them that others can observe. Thus, *older prisoners should end up in fewer fights, for their violence capital is higher*.

Consistent with the violence capital hypothesis, there is clear consensus in the literature about the effect that age has on violence: older inmates are less likely to be violent (e.g. Harer & Langan, 2001, p. 524; Reidy, Cunningham, & Sorensen, 2001, p. 74; Sorensen & Cunningham, 2007, p. 551; Sorensen & Pilgrim, 2000, p. 1265).

Part of this effect may well be due to hormonal levels, for younger males have higher levels of testosterone than older males (Harman, Metter, Tobin, Pearson, & Blackman, 2001), promoting aggression and risky behaviour among them (Archer, 2006). However, more younger inmates appear to be victims of physical assault than older inmates (Fuller & Orsagh, 1977, pp. 41, 46; Wolff, Shi, & Siegel, 2009, p. 459). Were testosterone levels to be the only component in increased violence, younger inmates should be assaulted *less* than older prisoners, for, knowing that younger inmates are more aggressive, riskier, and ultimately likelier to fight back, other prisoners should be likelier to leave them alone. In contrast, the higher rates of victimisation are consistent with the violence capital hypothesis for young prisoners have fewer signs and credible signals to display and are thus subject to more challenges, inevitably leading to more reported victimisation.

The more number of times that a person has been previously incarcerated, the more opportunity he has had to build up the marks that make it easier for others to identify his violence potential. Thus, following this line of reasoning *prior incarcerations should be negatively linked to fighting*. Whether this violence prowess is high or low, the previous prison experiences crystallise the information that a prisoner emits and that others perceive. At the same time, there is likely to be a counteracting effect in that people who

have been incarcerated a greater number of time are more risk prone, aggressive, or have lower time discounting factors and are consequently likely to fight more. *A priori*, it is unclear which of these two effects is stronger.

To try to separate these two effects, I test, in addition to the prior incarcerations variable, an interaction between prior incarcerations and time already served. Aggressive individuals are more likely to respond to challenges by 'upping the ante' than backing down, so the amount of fighting that they initially report, when they are likely to be challenged, should be higher than deferential individuals with fewer prior incarcerations. However, over time, once their signs and signals are perceived in the prison population, they are left alone quicker—consequently reducing the number of fights that they enter into over time—than inmates who are less aggressive and who lack sufficient violence capital.

Evidence from the literature shows that prior incarceration is typically associated with increased violence (e.g. Gaes, Wallace, Gilman, Klein-Saffran, & Suppa, 2002, p. 375; Sorensen & Pilgrim, 2000, p. 1265; Wooldredge & Steiner, 2009, p. 810). Yet, this is not always the case, as some papers report no relationship (e.g. Griffin & Hepburn, 2006, p. 437; Worrall & Morris, 2011, p. 143), and one reports a negative relationship (Morris, Longmire, Buffington-Vollum, & Vollum, 2010, p. 426). This suggests the importance that the interaction variable could have for separating the effect of aggression and riskiness from information.

Prisoners who spend more time in the education system before incarceration have fewer opportunities to build up their violence capital. Typically, people who spend time obtaining education, lack signs and signals of their violence ability and the signs of

education, if they are interpreted at all, would be considered as weak indicators of low violence potential. Moreover, people with more education are likely to have had fewer experiences of fighting in the past, making them less sure about their own fighting ability (Gambetta, 2009a, p. 81). Consequently, *more highly educated inmates should end up fighting more*.

Education appears to have a reasonably stable effect in that it is linked to less violence (Berg & DeLisi, 2006, p. 637; Harer & Langan, 2001, p. 524; Kuanliang, Sorensen, & Cunningham, 2008, p. 1196; Sorensen & Davis, 2011, p. 155), although there are two exceptions that find no link (Morris et al., 2010, pp. 426, 427, 429; Steiner & Wooldredge, 2008, p. 447). Still, the evidence concerning education is not definitive because only two of the studies finding a negative effect are based on substantive samples and analysed with multivariate techniques that include sufficient control variables. In addition, Shields and Simourd (1991, p. 186), among young offenders, find that aggressors have more education than their victims. This contradicts the typical cultural explanation posited for the lower level of violence among the higher educated. If highly educated people were more adjusted to mainstream values, including passivity, then why would they be more likely to be aggressors?

Race and ethnicity may be associated with perceived violence. Particularly, black and Hispanic inmates could have higher violence capital than white non-Hispanic prisoners. Both minority groups are overrepresented in socioeconomically deprived neighbourhoods in which there are higher levels of latent and actual violence. This provides more opportunity—albeit a strange and unwelcome one—to people from impoverished backgrounds to build up signs and signals of their violence potential that subsequently decrease fighting in prison.

In addition, the perceivable cluster of characteristics that are associated with 'blackness' and 'Hispanic-ness' may themselves serve as a source of violence capital. Like the tattoos and beards of the Hells Angels, blackness and Hispanic-ness may trigger beliefs that their possessors are tougher or more aggressive. Trawalter and colleagues report a series of experiments demonstrating that both white and black subjects perceive that black people feel less pain than white people (Trawalter, Hoffman, & Waytz, 2012). In a subtle, and qualitative, account, Brent Staple describes in his famous essay *Just Walk on By: A Black Man Ponders His Ability to Alter Public Space* how his presence, and the 'fearsomeness mistakenly attributed to me', as a young black male creates uneasiness and fear in others (Staples, 1986). To dispel this perception, he employs an astute signal, whistling melodies from Beethoven and Vivaldi 'and the more popular classical composers' on his evening walks, for 'everybody seems to sense that a mugger wouldn't be warbling bright, sunny selections from Vivaldi's *Four Seasons*'. Therefore, the violence capital hypothesis implies that *black and Hispanic inmates fight less than white inmates*.

Race and ethnicity have mixed associations with prison violence in the literature. Some authors report that black inmates are involved in higher levels of violence than whites (e.g. Huebner, 2003, p. 112; Reidy et al., 2001, p. 74; Steiner & Wooldredge, 2008, p. 447). Others report no relationship (e.g. Camp, Gaes, Langan, & Saylor, 2003, p. 517; Sorensen & Cunningham, 2007, p. 551; Worrall & Morris, 2011, p. 143), and one reports finding lower levels (Griffin & Hepburn, 2006, p. 437). For Hispanic inmates, some papers report a positive association (e.g. Berg & DeLisi, 2006, p. 637; Gaes et al., 2002, p. 375; Steiner & Wooldredge, 2009b, p. 474) and others report no link (Sorensen & Cunningham, 2007, p. 551; Sorensen & Davis, 2011, p. 155; Steiner & Wooldredge, 2013, pp. 1245, 1247).

Yet, consistent with this interpretation of race and ethnicity as signs, it has been found that white inmates are victimised more than black or Hispanic inmates (Fuller & Orsagh, 1977, p. 42; Maitland & Sluder, 1998; Wolff et al., 2009, p. 459). Whiteness could be uninformative, or a slightly negative sign, leading to more challenges and higher levels of victimisation.

Prisoners incarcerated for murder bring with them their offence of conviction that often other prisoners get to know about. Knowing that someone has committed murder is a good indicator that that person is willing to use violence to an extreme degree and so *prisoners who are incarcerated for murder should be fights less*. It is a clear sign of violence potential, which should lead to fewer fights. Other crimes are typically less clear indicators of one's type, committing fraud, for instance, does not imply that the prisoner is meek only that the prisoner could be either meek or aggressive. No study that I know of has looked at the level of inmate on inmate fighting, or assault, among convicted murderers relative to the rest of the population. While a small number of studies have looked at the behaviour of murderers (e.g. Cunningham, Reidy, & Sorensen, 2005; Marquart, Ekland-Olson, & Sorensen, 1989; Sorensen & Cunningham, 2010), the outcome is never restricted to only inmate on inmate assaults.

Size, and its rough correlate of strength, may be taken as indicators of one's ability to inflict damage in case of a fight. Erwin James writes that 'a key factor in making sure you keep safe on the prison landings is how much you can bench press. "If you can manage 100kg you'll be OK in here,"' (James, 2003, p. viii). It also increases the unpredictability of a fight, as a single punch is enough to end it. *Larger inmates can be expected to get into fewer fights*. This is an easily perceived sign that increases the violence capital of a larger inmate.

Smaller inmates can still inflict a lot of damage, although their ability to do it is less apparent.

Prisoners' size has rarely been considered as a predictor. When it has, as either height or weight, it is found to have no link with violence (Archer & Southall, 2009, p. 36; Connell & Farrington, 1996, p. 86; Maitland & Sluder, 1998, p. 67; Shields & Simourd, 1991, p. 186; Worrall & Morris, 2011, p. 143).

3.1.3 Information spread variables

To test the second hypothesis (prisoners who receive or emit less information fight more) I include five variables in the model. These are, the number of prisoners in a cell, working inside a facility, the security level of a facility, TV availability, and newspapers, magazines, and books availability. I also test an additional potential explanatory variable that is the number of prisoners in a sleeping area interacted with time served (Table 4).

The number of inmates who share a sleeping area is included as a continuous log natural transformed variable. It is originally a highly right-skewed variable; log transforming it makes the distribution more normal and reduces the influence of the extreme values on the coefficient estimate. Working within a facility, TV access, newspaper, magazines, and books access are included as dichotomous variables. Security category is used as a set of dichotomous dummy variables.

Sharing a cell with more people increases the number of sources from which a prisoner can obtain information on others, and, it increases the number of channels through which information about that prisoner is distributed. Both factors reduce the need for

violence, for the information is already available. In this case, the *number of people sharing a cell will be negatively related to fighting.*

Table 4. Descriptive statistics for the variables used to test the information availability hypothesis. Means and standard deviations, in parentheses, displayed. Potential explanatory variable is enclosed in square brackets.

Variables	State facilities	Federal facilities
Number of inmates in cell (log)	1.71 (1.68)	1.04 (1.04)
[Number of inmates in cell (log) * time served (log)]	5.47 (5.84)	3.51 (3.67)
Working within the facility	0.56 (0.50)	0.85 (0.36)
TV available to watch	0.88 (0.33)	0.94 (0.23)
Newspapers, magazines, books, available to read	0.93 (0.25)	0.97 (0.18)
Security		
Maximum	0.38 (0.48)	-
High	-	0.13 (0.33)
Medium	0.50 (0.50)	0.34 (0.47)
Low	-	0.27 (0.45)
Minimum	0.13 (0.33)	0.19 (0.39)
Administrative	-	0.07 (0.25)
Observations (facilities)	207	32
Observations (inmates)	9911	2721

However, the greater number of people increases the potential for conflicts to emerge. To get around this issue, I test the variable for the number of people sharing sleeping quarters interacted with time served. Directly after entry, there could be a positive association between the number of people and violence, for an inmate is encountering more unfamiliar people. After time, the increase in information flow decreases this effect, so that people who share a cell with more people face a quicker decrease in

violence incidence. If both a main effect and an interaction term are included in the model, then I expect that the *number of people sharing a cell is positively linked to fighting but the interaction term between the number of people and amount of time incarcerated to be negatively related to fighting.*

The effect of crowding on violence has been empirically examined a number of times. However, crowding is operationalised as the number of inmates relative to a prison's design capacity, floor space, or the absolute number of prisoners in a prison (e.g. Ekland-Olson, 1986) and no study that I know of has looked at how the number of people that a prisoner shares a cell with affects violence.

Working inside a facility provides an inmate with the opportunity to build contacts with people from around the prison, and who may have different information to their own. This allows prisoners working inside the facility to gather more information about others than prisoners who are not working in the facility. *Prisoners who work inside the facility can be expected to fight less.*

Only a handful of papers have included working inside a facility as a predictor. Those that do, universally find that it decreases violence (Huebner, 2003, p. 113; Jiang & Fisher-Giorlando, 2002, p. 348; Steiner & Wooldredge, 2008, p. 447, 2009b, p. 475; Wooldredge & Steiner, 2009, p. 810).

In higher security prisons, information may be less able to circulate about prisoners as prisoners are more restricted in their movements, and more restricted in how much time they are allowed to mingle with other inmates (e.g. Bench & Allen, 2003, p. 375). Of course, sorting is occurring in that higher risk inmates are placed, or moved, and this

could be responsible for the relationship. Therefore, *more fights take place in higher security facilities.*

Most of the studies on security categorisation find that there is more violence in higher security level facilities (Gaes et al., 2002, p. 375; Huebner, 2003, p. 113; Sorensen & Cunningham, 2010, p. 117; Sorensen & Davis, 2011, p. 155; Steiner & Wooldredge, 2008, p. 448, 2009b, p. 476, 2013, pp. 1246, 1248; Worrall & Morris, 2011, p. 143). A few papers, however, surprisingly report no association (Griffin & Hepburn, 2006, p. 437; Morris et al., 2010, p. 429).[77]

Access to TV, books, and newspapers may decrease violence because they facilitate the spread of information about newcomers. Information about a prisoner's crime may be discovered from newspaper clippings or reported on in television programs. Erwin describes how some prisoners, known as 'cutting merchants' are sought out by other prisoners when they need information and 'wield power by collecting newspaper cuttings of reported crimes, sometimes going back years' (James, 2003, p. 8). TV rooms may also facilitate the spread of information for they present common areas in which prisoners can mingle and talk. Based on this, I expect that *TV availability and reading material availability are negatively correlated with fighting.*

The extent to which information from newspapers and TVs influences prison violence is not known; we simply do not know how many of the prisoners obtain information about others from newspapers and TV. It is plausible that this happens very rarely, but it is also

[77] There is a small and fascinating experimental literature on security categories' effects on prisoner behaviour (Bench & Allen, 2003; Berk, Ladd, Graziano, & Baek, 2003; Camp & Gaes, 2005). Unfortunately, none of these studies have inmate on inmate violence alone as the outcome and instead use a pooled measure of misconduct.

plausible that it happens often enough to substantially affect violence. What we do know is that there are cases in which information obtained from these sources is important. Take Carl Bowles, an inmate at the then maximum security prison Fort Leavenworth. When he wanted to know about the crimes of another inmate, he went looking, and subsequently found the information that he was searching for, in true crime magazines (Earley, 1992, pp. 6–8). No papers that I could find include access to media as a predictor of inmate violence. For these reasons, it is worthwhile to examine the variables of TV and reading material availability.

Table 11 (p. 120) provides a summary of the expected relationships between fighting and each explanatory independent variable.

For a small number of my explanatory variables the available literature points to a relationship that is opposite to the ones predicted by the hypotheses (e.g. less educated inmates, according to the majority of the findings, are more likely to be involved in violence). This begs the question: despite some existing evidence contrary to my expectations, why do I test these variables? One reason is that I employ a theory that specifies a particular combination of variables that should be included in analyses as explanatory ones. The set of variables that I use here have so far not been simultaneously used in analyses. Combining the relevant variables in one analysis may reveal relationships that have until now been left covered under the mud of confounding variables. Another reasons is that almost all of the papers that I cite above, except Lahm (2008) and Reisig (2002), obtain information on violence either from the official records, or from questions that ask prisoners to report their official violence record. These papers, therefore, address only officially recognised instances of violence in which both

violence and official sanctioning have occurred. Using self-reported, self-recognised cases of violence may provide another perspective on prison violence.

3.1.4 Control variables

My selection of potential control variables is guided by two concerns. First, I test the basic control variables that are typically used in quantitative studies on prison misconduct. Second, additional controls must be plausible confounders: either theory or empirical evidence links the control to both an explanatory variable and the outcome. This leaves a large set of potential control variables, 24 in total, all of which I test during model building.

At the individual level these are time already served, incarcerated for a violent crime, time left to serve, taken part in religious activity in the past week, ever stayed overnight in a mental hospital, used drugs in the month before arrest, working outside the facility, currently has children, grew up with both parents, participates in some prison activity,[78] sentenced for lewd acts with children, is currently married, was employed before arrest, ever suffered physical abuse, and involvement with an antisocial peer group while growing up.

At the prison level these are facility location, mean age of inmates, average number of prisoners sharing a sleeping area, proportion working, mean number of prior incarcerations, mean club involvement, proportion using drugs before arrest, proportion of prisoners sentenced for a violent crime, and proportion of white inmates. Prison level

[78] This variable is composed of participation in vocational or job training program, education program, classes in life skills, pre-release program, outside community activities, classes or groups doing arts crafts, bible club or other religious study group, other religious activities, prisoner assistance groups, prisoner self-help and personal improvement groups, drug or alcohol groups, such as Alcoholics Anonymous, or an ethnic or racial organisation.

variables are based on individual prisoners' attributes and are calculated for the restricted sample used in the analyses. The descriptive statistics for the control variables that are used in the final models are displayed in Table 5.

Table 5. Control variables included in the analyses of state and federal prisons. Means or proportions and standard deviations, in parentheses, displayed.

Variables	State facilities	Federal facilities
Time served (log)	3.42 (1.11)	3.49 (1.00)
Mental health issues	0.10 (0.30)	0.05 (0.22)
Involved with antisocial peer group	0.67 (0.47)	0.49 (0.50)
Work outside facility	0.10 (0.30)	-
Drug use	0.57 (0.50)	-
Abused	0.13 (0.34)	-
Violent crime	-	0.24 (0.43)
Time left (months)	-	104.57 (90.87)
Married	-	0.30 (0.46)
Employed	-	0.73 (0.44)
Region		
North East	0.15 (0.36)	0.10 (0.30)
Mid-West	0.19 (0.39)	0.15 (0.36)
South	0.46 (0.50)	0.54 (0.50)
West	0.20 (0.40)	0.21 (0.41)
Mean prisoner age	33.33 (3.43)	37.62 (2.48)
Proportion used drugs	0.57 (0.09)	-
Proportion in activity	-	0.48 (0.10)
Observations (facilities)	207	32
Observations (inmates)	9911	2721

3.2 Analysis

Since my outcome is a count variable, OLS regression is not an appropriate statistical tool to use. Poisson regression is an often-used alternative, however, it assumes that the conditional mean of the dependent variable is equal to its conditional variance. Initial inspection of the unconditional mean and variance of the dependent variable reveals that its variance is substantially greater than its mean, known as 'overdispersion' ($mean_{State}$=1.27 and $variance_{State}$=28.37; $mean_{Federal}$=0.33 and $variance_{Federal}$=3.36). In such a case, negative binomial regression is a more appropriate model to use. This generalisation of the Poisson regression allows the variance to be greater than the mean. Statistics concerning 'alpha', a parameter indicating the overdispersion, in the final models show that negative binomial regression is, indeed, more appropriate than Poisson regression (LR $\boldsymbol{X}^2_{State}$=1558.64, p<0.001; LR $\boldsymbol{X}^2_{Federal}$=1511.46, p<0.001). Consequently, I employ negative binomial regressions for analyses.

Throughout the analyses, I use cluster robust standard errors at the prison level to adjust the standard errors to account for the within prison correlations among prisoners. In addition, the final survey weights, provided in the SISFCF, are used to correct for their sampling procedure.[79]

Before model building, I further reduce my sample and remove observations that have missing data for any of my *potential* independent variables.[80] Without this, the number of observations in my models, at different stages in their builds, could differ. When estimating the final model, I use the sample already described. Thus, the sample on

[79] V2067 in the dataset.
[80] Leaving n_{State}= 8848 and $n_{Federal}$=2686 for the model building sample.

which model building takes places is slightly different to the sample on which the final model estimation is carried out.

As I test both of the hypotheses in the same model, I begin analysis by adding the basic explanatory variables (Table 3 and Table 4). Then I test whether the two interaction explanatory variables (prior incarcerations interacted with time served and number of prisoners in a sleeping area interacted with time served) are significantly associated with the dependent variable. If significant, I leave them in otherwise I exclude them from the analysis.

Then, I test, one by one, theoretically plausible and empirically based control variables. After adding a control variable, and estimating the model, I test whether the newly added control variable is linked to the outcome. If the control variable is significant at the 5%-level, I keep the new variable, and move on to testing another variable. If the new variable is not significant, I leave it out, and move on to trying another one.

A problem of this incremental way of model building is path-dependency: one may incorrectly exclude, or include, variables due to the order that they are tested in. To assuage this issue I do the following. Once I try all of the potential control variables, and have a temporary end model, I re-check whether variables that I initially exclude using the above procedure, including the two interaction explanatory variables, are significantly linked to fighting. In case they still are not, I leave them out. Otherwise, I include them in the model. Finally, I reduce the model by excluding non-significant control variables. The full model that I use is shown in Appendix C.

3.3 Results

Table 6. Reduced output for final models including controls on the number of fights inmates have been in in state and federal facilities. Coefficients, standard errors, and incidence rate ratios displayed. The full regression output is displayed in Appendix C.

Variables	State facilities		Federal facilities	
	Coefficient (S.E.)	IRR	Coefficient (S.E.)	IRR
Age	-0.07*** (0.01)	0.93	-0.04*** (0.01)	0.96
Number of prior incarcerations	0.09*** (0.03)	1.10	0.04 (0.07)	1.04
Years in education	0.05** (0.02)	1.06	-0.03 (0.03)	0.97
Race/ethnicity				
White	Reference category		Reference category	
Black non-Hispanic	-0.78*** (0.09)	0.46	-0.02 (0.19)	0.98
Hispanic	-0.59*** (0.11)	0.56	-0.61* (0.27)	0.54*
Other non-Hispanic	-0.54** (0.17)	0.58	1.10*[a] (0.48)	2.99*
Murder	0.04 (0.08)	1.04	0.17 (0.30)	1.19
BMI (log)	-0.29 (0.22)	0.75	-0.65 (0.70)	0.52
Number of inmates in cell (log)	-0.04† (0.03)	0.96	0.69* (0.29)	1.99
Number of inmates in cell (log) * time served (log)		-	-0.18* (0.08)	0.83
Working within the facility	-0.20** (0.07)	0.82	-0.45† (0.27)	0.64
TV available to watch	-0.20† (0.11)	0.82	-0.26 (0.37)	0.77
Newspapers, magazines, books, available to read	0.09 (0.11)	1.09	-0.58 (0.66)	0.56
Security				
Maximum	Reference category		-	
High	-		Reference category	
Medium	-0.17* (0.08)	0.85	-0.16 (0.19)	0.85
Low	-		-1.25*** (0.28)	0.29
Minimum	-0.50* (0.21)	0.61	-1.73*** (0.45)	0.18

Administrative		-	-1.29*** (0.34)	0.28
Constant	-0.01 (0.86)	0.99	-4.72 (3.08)	0.01
n_2	207		32	
n_1	9911		2721	

† p<0.1, * p<0.05, ** p<0.01, *** p<0.001

[a] Positive relationship based entirely on a single observation: one prisoner who is other non-Hispanic and who reports that he was involved in fights 45 times. Removing this single observation removes the significance of this coefficient.

3.3.1 Violence capital

In state prisons, the coefficients for age, time spent in education, and race and ethnicity, are consistent with the violence capital hypothesis (Table 6). Older inmates fight less than younger inmates (β=-0.07), inmates with more education fight more than less educated inmates (β=0.05), and black and Hispanic inmates fight less than white prisoners (β=-0.78 and β=-0.59).

These coefficients have substantial impacts on the number of fights that an inmate is expected to be involved in (Table 7).[81] A 16 year old prisoner, the youngest in the sample, is expected to fight 1.94 times, while an 89 year old prisoner, the oldest in the sample, is expect to fight 0.01 times.[82] That is a staggering 99.5% [(1.94-0.01)/1.94]*100 reduction in the number of fights. Even considering a smaller step and moving from the minimum age to the mean age of 33 years, there is still a dramatic decrease of 71.6% [(1.94-0.55)/1.94]*100 in fighting. A prisoner with no education has fewer than half the fights that an inmate with 18 years of education has (0.30 vs. 0.79). Black inmates are

[81] The predicted number of fights are obtained when the coefficients not under manipulation are set to their means.

[82] Forty prisoners in the sample are younger than 18 years of age. A study commissioned by U.S. Department of Justice found that in 1998 around 14,500 youth are housed in adult facilities on any given day. This is a widespread occurrence in the U.S., with 44 states housing juveniles in adult jails and prisons (Austin, Johnson, & Gregoriou, 2000, p. x).

predicted to have fewer than half the fights of white inmates (0.40 vs. 0.86) and Hispanic inmates have nearly half the fights of white inmates (0.48 vs. 0.86).

The author of this paper, male, 27 years of age, white, in education for 20 years, with no prior incarcerations, would, in a state facility be predicted to fight 2.03 times over the course of 31 months. An unappealing proposition, but far from the ever-violent milieux that prisons are sometimes portrayed as.

Table 7. Predicted number of fights in state facilities for significant violence capital variables.

State facilities	Predicted number
Age = 16 (min)	1.94
Age = 33 (mean)	0.55
Age = 89 (max)	0.01
Years in education = 0 (min)	0.30
Years in education = 18 (max)	0.79
Race/ethnicity	
White non-Hispanic	0.86
Black non-Hispanic	0.40
Hispanic	0.48

N.B. All other variables set to their survey weight adjusted means.

Also consistent with the violence capital hypothesis are the coefficients for age and ethnicity in federal prisons. Older inmates participate in less violence (β=-0.04) and Hispanic prisoners fight less than white prisoners (β=-0.61).

The oldest inmate in a federal facility, 83 in the sample, is predicted to fight 94.4% less [(0.18-0.01)/0.18]*100 than the youngest inmate, who is 19. A prisoner of average age, 38 years, is expected to fight 55.6% less [(0.18-0.08)/0.18]*100 than the youngest inmate. Prisoners of Hispanic ethnicity are predicted to fight 44.4% less [(0.09-0.05)/0.09]*100 than non-Hispanic white inmates (Table 8).

For the author of this paper, the situation in a federal facility would be much more favourable: only 0.11 fights over 35 months.

Table 8. Predicted number of fights in federal facilities for significant violence capital variables.

Federal facilities	Predicted number
Age = 19 (min)	0.18
Age = 38 (mean)	0.08
Age = 83 (max)	0.01
Race/ethnicity	
White	0.09
Hispanic	0.05

N.B. All other variables set to their survey weight adjusted means.

A single coefficient, for the number of prior convictions in state facilities, goes against the expected direction. Prisoners with more prior convictions fight more (β=0.09).

The other violence capital variables, incarceration for murder and body mass index in state facilities and number of prior incarcerations, years of education, race, incarceration for murder and body mass index in federal facilities, are non-significant. In addition, the interaction term between prior convictions and time served is not used in the regressions for it did not achieve significance at the 5%-level in either sample.

Where significance is achieved, giving us reasonable certainty about the estimated coefficient, the results generally support the violence capital hypothesis.

3.3.2 Information spread

All of the coefficients that attain significance are consistent with the information spread hypothesis. In state facilities, the coefficients for the number of inmates in a sleeping area (p=0.09), working within a facility, TV access (p=0.08), and security level, are

significant and go in the expected direction (Table 6). The more inmates share a sleeping area, the less violence an inmate experiences (β=-0.04). Prisoners who work inside their prison are involved in less fighting (β=-0.20), as are prisoners who have television access (β=-0.20). Inmates housed in medium security facilities fight less than those in maximum security ones (β=-0.17), and inmates in minimum security facilities also fight less than prisoners in maximum security facilities (β=-0.50).

A prisoner who shares sleeping quarters with the most number of other inmates has a 24.6% reduction [(0.57-0.43)/0.57]*100 in fighting relative to an inmate who has a cell to himself. Working inside the facility, relative to not working inside the facility, decreases violence by 18.3% [(0.60-0.49)/0.60]*100, and TV access decreases fighting by 17.5% [(0.63-0.52)/0.63]*100. Prisoners housed in medium security facilities are predicted to be in 16.1% fewer fights [(0.62-0.52)/0.62]*100 than prisoners in maximum security facilities, and inmates in minimum security prisons are predicted to fight 40.3% less [(0.62-0.37)/0.62]*100 than inmates in maximum security prisons (Table 9).

Table 9. Predicted number of fights in state facilities for significant information spread variables.

State facilities	Predicted number
Inmates in cell (log) = 0 (min)	0.57
Inmates in cell (log) = 6.55 (max)	0.43
Not working inside facility	0.60
Working inside facility	0.49
TV not available	0.63
TV available	0.52
Security	
Maximum	0.62
Medium	0.52
Minimum	0.37

N.B. All other variables set to their survey weight adjusted means.

Taken together, these variables have substantial effects on fighting. A lone prisoner, not working inside the facility, without a TV, and incarcerated in a maximum-security facility is predicted to fight 0.88 times. In contrast, an inmate who shares sleeping quarters with the maximum number of prisoners, works inside the facility, has a TV, and is incarcerated in a minimum-security facility is predicted to fight 0.27 times over 35 months.

In federal facilities, the number of inmates in a sleeping area interacted with the amount of time served, working inside the facility (significant at the 10% level), and security match the information spread hypothesis (Table 6). Inmates who have spent a long time in prison and have many cellmates report fewer fights than inmates who have spent a long time in prison but have fewer cellmates (β=-0.18). Although coefficient for the number of others who share a sleeping area with an inmate is positive, this is not surprising, and the effect of the two coefficients have to be considered together. Prisoners who work inside the facility are involved in fewer fights than those who do not work (β=-0.45). Inmates housed in low and minimum security facilities are involved in less violence than prisoners in high security facilities (β=-1.25 and β=-1.73).

Consider an inmate who has served 67 months, which corresponds to the 75th percentile time that prisoners in the sample have served. In a single occupancy cell, he is predicted to fight 0.22 times, while in a sleeping area with the maximum number of others, he is predicted to fight 0.13 times, a 40.9% reduction [(0.22-0.13)/0.22]*100. An inmate who works inside a prison facility fights 33.3% fewer times than an inmate who does not [(0.12-0.08)/0.12]*100. Turning to security level, the model predicts that a prisoner in a high security facility fights 0.19 times, 0.16 times in a medium security facility, 0.05 times in a low security facility, and 0.03 times in a minimum security facility. Moving from the

most secure to the least secure facility is associated with an 84.2% [(0.19-0.03)/0.19]*100 reduction in violence (Table 10).

Table 10. Predicted number of fights for the significant information spread variables in federal facilities.

Federal facilities	Predicted number
Inmates in cell (log) = 0 (min) Time served (log) = 4.21 (75th percentile) Inmates in cell (log) * time served (log) = 0	0.22
Inmate in cell (log) = 5.86 (max) Time served (log) = 4.21 (75th percentile) Inmates in cell (log) * time served (log) = 24.67	0.13
Not working inside the facility	0.12
Working inside the facility	0.08
Security	
High	0.19
Medium	0.16
Low	0.05
Minimum	0.03

N.B. All other variables set to their survey weight adjusted means.

Although the absolute effect is relatively small, for the number of fights that occur in federal facilities are few, together the significant information spread variables have a considerable relative impact. A prisoner housed alone in a maximum security facility who has served 67 months and who does not work is expected to fight 0.76 times, while an inmate in a minimum security facility sharing a sleeping area with the maximum number of others, who has served 67 months, and who works is predicted to fight 0.05 times. This corresponds to a 93.4% decrease [(0.76-0.05)/0.76]*100. The other information spread variables do not achieve significance.

Overall, these results are largely consistent with the information spread hypothesis since every statistically significant coefficient goes in the direction expected.

3.3.3 Summary

Table 11. The expected and observed relationships between each independent variable and fighting. Reported relationships are significant at the 5% level.

Variable	Expected	Observed	
		State facilities	Federal facilities
Age	-	-	-
Prior incarceration	-	+	0
Education	+	+	0
Black-non Hispanic[a]	-	-	0
Hispanic[a]	-	-	-
Murder	-	0	0
BMI	-	0	0
Cellmates (alone)[b]	-	0[e]	N/A
Cellmates (interaction)[c]	+	N/A	+
Cellmates by time served	-	N/A	-
Working within facility	-	-	0[e]
TV availability	-	0[e]	0
Reading material available	-	0	0
Security level[d]	-	-	-

[a] White non-Hispanic is the comparison group.
[b] When only cellmates without an interaction term is included in the model.
[c] When both cellmates and cellmates by time served are included in the model.
[e] Negative at the 10% level.

Concerning the violence capital hypothesis:

- In state facilities, inmate age, amount of time spent in education and race and ethnicity are consistent.
- In federal facilities, inmate age, and ethnicity are consistent.
- However, prisoners with more prior convictions in state prisons are associated with more fighting.

Regarding the information spread hypothesis:

- In state facilities, the number of inmates sharing a cell, whether an inmate is working within a facility, TV availability, and facility security level match expectations.

- In federal facilities, the number of inmates sharing a cell interacted with time served, working inside a prison, and prison security level support the hypothesis.
- None of the variables contradict the information spread hypothesis.

3.4 Discussion

My results are largely consistent with the informational theory of prisoner conflict: information availability is negatively related to the amount of violence that inmates experience. Prisoners with more signs and signals of their fighting prowess, and those who inhabit environments in which the spread of information is promoted, fight less.

Crucially, these results hold, or in some cases only emerge, after controlling for multiple variables that could feasibly confound the associations between the explanatory variables and the outcome. Some of the control variables, the ones that were found to be significantly linked, ended up in the final regression models. Others were not included in the final models, but were tested and found to have no effect. In total, 24 potential control variables were checked, and either included, or discarded because they had no effect.

Some control variables were surprisingly unrelated to fighting in both state and federal prisons and thus left out of the final models. Despite the oft-recounted fact that paedophiles are beaten, and sometimes killed, in prison, there is no association between such a conviction and fighting. Such a result could be interpreted as undermining this version of reality, or at least the extent to which it is widespread, and perhaps prisoners who report such values are typically relaying the macho culture of prisons without converting them into action. More likely, the threat is real and the lack of correlation

occurs because the state and federal prison services do a good job of separating and protecting such prisoners from others.

Growing up with both parents, a variable sometimes considered as an indicator of socialisation, stability, and positive outcomes (e.g. McLanahan & Sandefur, 1994) is not linked to fighting. Neither is having children linked to fighting. Intuitively, prisoners with children outside may be more motivated for release, less invested in building up a life in prison, than those without children, and so try to avoid violence whenever possible. This is not bourn out in my data.

Two further variables that were not used in the final analysis are worth mentioning. I find that participating in religious activity in the past week is not correlated with fighting. This is inconsistent with some previous findings regarding prisoner misconduct and adjustment (e.g. Clear & Sumter, 2002; Kerley, Matthews, & Blanchard, 2005; O'Connor & Perreyclear, 2002) but matching one (Sturgis, 2010). Finally, participating in prison activities is also unlinked to violence.

Prison gangs are one factor that may affect violence in prisons and that I did not address in this study. As Gambetta mentions, resources in prison may also be distributed according to the violence potential of groups, not only according to the violence potential of individuals (2009a, p. 79). Nevertheless, the theoretical links between individual violence potential, information, and gangs, are likely to be complex, and developing such a theory is not the aim of this work. Considering only one aspect briefly, prisoners with the lowest violence potential, those who are the most vulnerable, may have the most incentive to join a gang in order to derive the benefits from collective strength. On the other hand, some gangs have incentives to only selected the toughest

would-be candidates—the Aryan Brotherhood are reported to do this (Grann, 2004, p. 159). On an empirical level, the SISFCF does not contain information about gang affiliation. I leave it to future research to pursue the fascinating complication of considering the interplay between violence potential, information, and gangs.

Another factor that I did not, and could not, address—since there is no information about it in the SISFCF—is the relationship between prisoner turnover and prison violence. There are some reasons to think that higher turnover leads to more violence including one that is congruent with the informational theory: higher turnover implies more interactions among prisoners, some of whom will have low violence capital, leading to higher levels of violence among them to establish dominance. Indeed, Gambetta presents some data that match the prediction (2009a, pp. 94, 95). More broadly, the small number of past quantitative studies that have looked at the relationship between turnover and prison violence have found mixed results: Gaes and McGuire (1985, p. 51) and Harer and Steffensmeier (1996, p. 338) find no link, while Porporino and Dudley (1984, pp. 14, 15) and Porporino (1986, p. 231) find a positive relationship. In any case, the relationship between turnover and my explanatory variables are far from clear; there are no clear reasons why turnover would be a confound for the explanatory variables employed.

However, turnover could be one of the variables that explains the large difference in the number of fights in state and federal facilities (1.27 fights on average in the former vs. 0.33 fights in the latter)—although I was unable to find data to verify this conjecture. I suspect instead that the difference is largely down to two main variables: the average age of inmates in state facilities is lower than in federal facilities (33 years vs. 38 years) and a

much larger proportion of state prisoners are housed in medium or higher security facilities than federal prisoners (88% vs. 47%).

In the rest of the discussion, I focus on the findings from state facilities. This is for a number of reasons. First, apart from relationship between the number of inmates sharing a sleeping area, which requires specific discussion, all of the coefficients that are significant in federal facilities are also significant and in the same direction in state facilities. Second, the inmates in state facilities provide a broader cross-section of criminals. Third, far more violence is reported in state facilities than in federal ones, suggesting that the former are places in which more challenging and physical conflict over resources occurs, intensifying the need to obtain and emit information. Fourth, there are many more observations in state prisons providing greater power to detect statistical relationships. Whenever I refer to results that hold for only state or federal facilities, I explicitly mention it.

For some variables, I find results that do not match the findings reported by other papers. One explanation for the disparity is my dependent variable: self-reported self-recognised cases of fighting. Another possibility is that the combination of variables that I include in my regressions are different to that used by other studies. Below I also discuss differences in findings.

3.4.1 Violence capital

Age has an unparalleled effect on fighting. It is strongly, and consistently, negatively linked with fighting in my data, as well as in the literature. From a biological point of view, this is unsurprising, but it is also consistent with an information explanation. For this reason, it is not useful to dwell on this finding too much.

Fighting and years spent in education and 'whiteness', in state facilities, are positively linked. In federal facilities too, white inmates fight more than Hispanic prisoners. Parsimonious explanations, apart from the information one, are difficult to find. A cultural perspective on education and whiteness would both predict that these groups report less violence. The former because of the absorption of mainstream culture values that, at least explicitly, discourages violence, and the latter because it correlates with socioeconomic status, in which socialisation occurs, also discouraging violence. This is broadly the argument that Steiner and Wooldredge make to predict that white prisoners are less likely than black and Hispanic prisoners to violate prison rules (Steiner & Wooldredge, 2009b, pp. 461–463). In contrast, the violence capital conjecture accounts for them both. More educated and white prisoners have fewer signs of their violence ability, for they are likelier to grow-up in richer, higher socioeconomic status, neighbourhoods in which the day-to-day level of violence is lower, decreasing their opportunity to build up their displayable, and observable, indicators of violence prowess.

Interestingly, this paper is one of the few that finds such relationships. Most of the literature reports that more educated people are involved in less violence, and that white prisoners experience less violence than black or Hispanic inmates—albeit the findings for the former seem more consistent than for the latter.

Only the number of prior incarcerations in state facilities opposes the violence capital hypothesis. Yet, this variable is perhaps the one that has the least potential to falsify the hypothesis. This is because there is likely to be an underlying unmeasured variable that confounds, and potentially overpowers, any decrease in fighting due to information. Aggressive, risk-prone people are likelier to have more previous incarcerations and be

involved in more fighting; the character traits associated with one are the similar as with the other.

The relationship between prior incarcerations and fighting does not change over time. This is demonstrated by the relevant coefficient's lack of significance; the coefficient for the interaction variable between the number of prior incarcerations and the amount of time already served. Fighting is consistently higher for inmates with more previous convictions, and consequently, the interaction variable is not included in the final model.

Incarceration for murder and body mass index are not linked to fighting. Both of these variables were somewhat 'long shots'—nevertheless worthwhile to test—for observing the expected relationships. Concerning murder, the information that a crime conveys is highly context specific: a single variation may reveal the perpetrator to be a 'mouse' instead of a 'man'. Killing a young man in his sleep, for instance, first drugging him, conveys very different information than killing that same person in a one-on-one fistfight. Such, crucial, specific information is not possible to extract from the data. In addition, prosecutors and judges have considerable leeway over the exact offense that a person is tried for, which may be influence by biases or personal motivations—a prosecutor may be up for (re)election and, wishing to run a campaign for being tough on crime, chooses to prosecute a defendant for murder instead of manslaughter. For these reasons murder as an indicator of violence capital is murky.

While body mass index should be closely linked to strength, this is only one component of a person's fighting ability. It indicates how much damage a person is capable of and how likely he is to come out on top in a physical confrontation. Yet, it does not say anything about the most important aspect: whether or not a prisoner is willing to use

violence in the first place. On the first day that Michael Thompson, standing at six foot four and weighing near-three hundred pounds, was relocated to a new prison, he was challenged by an established inmate (Grann, 2004, p. 159). His size alone was not enough to deter attackers. In addition, body mass index does not capture whether a prisoner is overweight or fit and muscle-bound, important aspects for the information that others derive from the manifesta. For these reasons, it is plausible that body mass index is not linked to fighting.

3.4.2 Information spread

Despite the counteracting effect that being around a greater number of other inmates has—more opportunities for arguments, conflicts over resources, and generally violence to erupt—there is a tendency, the coefficient is significant at the 10% level, for the number of prisoners who share sleeping quarters to be negatively related to the fighting in state facilities. More inmates in a sleeping area are associated with fewer fights. The variable for interaction between time served and number of inmates sleeping in the same area, meanwhile, is not significant and not included in the final model for state facilities. In federal facilities, the relationship is different. The coefficient for the number of prisoners sleeping in the same area is significant and positive while the coefficient for the interaction variable with time served is negative. Why is there this difference between state and federal facilities?

One possible explanation stems from the greater mixing of prisoners in federal facilities than in state prisons. More criminals from different parts of the country are placed together in the former. This makes it less likely that prisoners in federal facilities know someone, directly or indirectly, in the institution in which they are placed, who can vouch for them, or spread word of their accomplishments. They need to use violence,

and given that they encounter more others in their sleeping area, there is an initially higher level. In state prisons, people may know of one another, or know of someone who does, who can spread information about them reducing the need for violence.

There are two alternative explanations for effects of cell size on fighting. One possibility is that prisoners less prone to violence are moved into cells with more others, while prisoners who are more prone to violence are separated and moved into smaller cells. In this perspective, larger cells are linked to less violence because there are less violent individuals within them. While this can directly account for the negative tendency found in the state facilities, it has a difficult time explaining the findings in federal facilities. If less aggressive prisoners are allocated to larger cells, then why do prisoners in larger cells fight more initially and only decrease fighting after time? Based only on a propensity explanation, such prisoners should fight less at each stage in their incarceration.

The other alternative is that prisoners in larger cells find it easier to build up and forge alliance since they have more of an opportunity to form links during their daily interactions.

Consistent with the information theory, working inside prison is negatively associated with violence, in both state and federal facilities. Prisoners who work fight less. There is a potential issue of selection: well-behaved or trustworthy prisoners are chosen, or agree, to work. While this explanation cannot be excluded, many control variables are included or tested, which diminishes the possibility that the effect is entirely down to selection.

Facility security level is consistently and strongly linked with fighting. In both state and federal prisons, more prisoners fight in more secure facilities. One interpretation of this

is that the prison service is allocating prisoners correctly: prisoners that are more violent are assigned to higher security facilities. This is likely to be true, yet, it is unlikely to be the entire story, for even after considering the proportion of prisoners with violent previous convictions and various other controls that could be expected to enable this result, the relationship remains clear.

Finally, there is a tendency for television access to be linked, at the 10% level, to a decrease in violence. There are two reasons why this is an interesting result. First, there is very little variation in the sample over access to television: almost every inmate, 88%, in a state facility reports access to a television. This reduces the power that statistical analysis has for detecting such a relationship. Second, providing television access also creates an additional resource over which there can be the possibility for violence to erupt. Prisons in which televisions are shared between prisoners, say as communal television rooms, conflict could arise over which programme to watch. But other dangerous also lay lurking for the unsuspecting. Consider what happened to an inmate who unwittingly took another prisoner's, known as 'Binbags', chair in the TV room of a high-security prison in England.

> Years earlier, in a high-security prison, Binbags had arrived for his regular mid-week fix of EastEnders only to discover his seat had been taken by a new prisoner. He knew he could not afford to lose the potential confrontation and so took what he believed to be the only appropriate action in the circumstances. Without a word he exited quickly, but returned minutes later carrying a jug of boiling water. In the dark, malodorous room, a couple of dozen pairs of eyes watched in silence as Binbags crept up behind the newcomer.
>
> The screams emitted as the hot liquid cascaded down the back of the innocent's neck temporarily drowned out the Albert Square drama's signature tune (James, 2003, p. 17).

Even when each prisoner has his own television set, the opportunity for conflict exists as inmates may steal one another's. To my knowledge, this is the first study that demonstrates a link (at the 10% level) between television access and decreased prison fighting.

There is also a potential alternative explanation for the tendency: television access takes prisoners' attentions away from everyday life. It is a distraction, and a way of relaxing, so prisoners are more likely to be calm in their interactions.

None of the variables used to investigate information spread contravened the hypothesis. Only the variable indicating access to newspaper, magazine, and book has no link with violence: neither in state nor in federal facilities. Perhaps this is because almost all inmates—like with TV availability—report access: 93% in state and 97% in federal prisons.

The goal of this chapter is to provide a multivariate quantitative test of the information theory of conflict among prisoners. Using negative binomial regressions to analyse data from the SISFCF on a nationally representative sample of prisoner, and controlling for multiple potential confound variables, I find that the majority of the explanatory variables are consistent with the hypotheses derived from the theory.

As is typical of observational data, it is not possible to identify causality. Neither can I identify the mechanisms causing the observed relationships; for some of the observed relationships, there are alternative explanations to the informational one. It is not possible to exclude, for instance, that security level is related positively with fighting because more risk-prone inmates are moved to higher security facilities. While it is true

that there are individual alternative explanations to many of the observed relationships, the informational theory is the most parsimonious one that can account for nearly all of them. Longitudinal or experimental data can be helpful to establishing causality; the latter is an approach I take in the following chapter.

Taken together, my results are largely consistent with the information theory of prison violence and suggests that prison violence, and the decisions of prisoners to use violence, may be affected by the amount of displayable and available information that they, and other prisoners, possess.

4. Communication and Information in Interpersonal Conflict

4.1 Introduction

When typical rules for allocating resources are lacking, or it is not possible to enforce them, as in prisons, force may be used by people who desire the resources, potentially leading to conflict and fighting among them. Information could play a crucial role in resolving such interpersonal conflict without violence. The presence or absence of relevant information can assuage or intensify the likelihood that violence erupts, and it may motive people to take communicative acts.

Here, I test both of the themes running through the thesis. First, how peoples' signalling behaviour changes when pre-existing signs are already present in an interaction. Will senders whose toughness is beyond reproach avoid costly signals, or will a failure to 'shout' for even these 'stand up guys' hurt their reputation? Second, whether information changes if physical violence emerges or not between the potential combatants. Does information about the ability (or lack of it) of each to hold one's own decrease fighting?

Prisoners inhabit environments in which others' beliefs about their ability to use violence is particularly important—as already discussed in the previous chapter. There are severe restrictions on the availability of resources that they value, many prisoners are incarcerated for taking others' resources, they inhabit an environment in which there is uncertainty about others type, it is difficult to avoid one another, and the consequences to violence are typically more severe than in the outside world (Gambetta, 2009a, pp. 78–110). Consequently, their incentives to change other prisoners' beliefs to their own

advantage are intense and the effect of information on their behaviour should be strong. For these reasons, I present, and discuss the following experiment in the context of prison conflict. Nevertheless, the same reasoning applies to a broad set of situations in which people desire others' resources, are not governed by enforceable allocation rules, and are capable of using violence.

Despite the potential that the informational perspective has for helping us to understand this aspect of the 'dark' side of human behaviour, no experimental research, to my knowledge, has been done to investigate the links between communication, information, and interpersonal conflict. To examine the research questions, I use a novel laboratory experiment that combines a strategic situation, of my own design, in which subjects use their actual toughness as a sign and signal and a physical competition that captures important aspects of fighting.

4.2 Theory

When people find themselves competing over access to scarce resources that many of them value, they may resort to using physical violence in order to allocate the resources, with the spoils going to the victor. Yet, doing this can be very costly for all of the people involved, and a cheaper solution, instead of actually resorting to violence, is to try to predict who would win and who would lose if a fight were to occur and to form a hierarchy based on the belief about these outcomes.

Forming accurate beliefs is a far from straightforward affair. Crucially, many forms of intentional communication cannot be relied upon to truthfully convey information. They suffer from the problem of mimicry—all prisoners could verbally claim to be the most

proficient with violence irrespective of their actual capabilities. At the same time, there are signals that mostly, or completely, retain their informative content.

4.2.1 (Counter)Signalling theory

Which fate awaits a signal—whether it is reliable or not—is a question that signalling theory answers. Proposed by Michael Spence (1973, 1974), the theory separates agents into senders, agents who take a signalling action, and receivers, agents who observe the signalling action of senders. There is a further separation: senders are either low type or high type, where type refers to the level of a quality that they possess.

Receivers gain the most by accurately estimating the type of senders—using each sender's signal and their knowledge of the distribution of types in the population—so, as the discrepancy between true and perceived quality increases the payoff to receivers decreases. Senders gain the most by being treated as high types by receivers.

The key prediction signalling theory makes is that when it is possible for them to do so, high type senders should take a signalling act that separates them from those below (see sections 1.1 and 1.2 for a deeper review of signalling theory).

In addition to signals, people often emit, and others perceive, prior information about their qualities. This prior information is (a) exogenous to the control of senders and (b) 'fuzzy' in that senders know the probability distribution that they have of emitting specific information but do not know the specific outcome that receivers actually observe. Reputation, accent, vocabulary, and body language, for instance, all fit the characteristics of extra information. They are frequently noisy—for instance one cannot be exactly sure what another knows of one's reputation—and, at least in the short term,

largely exogenous to the sender. While extra information can also be signals, typically they are signs, and, due to this experiment's design, the extra information is equivalent to a sign. Thus, in this paper I use the terms extra information and sign interchangeably.

When both extra information and signals are present in an interaction standard signalling theory becomes inapplicable—for it only includes the signal—and countersignalling theory may be more appropriate (Feltovich et al., 2002). It is an extension of signalling theory that deals with such richer situations and its main prediction is that medium types signal more intensely than both low and high types (see Section 2.2 for an in-depth explanation of countersignalling theory).

4.2.2 Information availability

Signalling and countersignalling theory predict the action that a sender should take assuming the opportunity to signal exists. They presuppose that communication can occur. However, it may not always be possible to communicate in conflict situations, for lack of available raw materials to transform into reliable signals, or because obstacles prevent the transmission or perception of relevant signals.

When it is not possible to send or receive information about violence prowess, Gambetta's information theory predicts that the proportion of interactions that result in fighting are higher than when it is possible to do so. As people cannot decide who would win and who would lose, in the absence of information, they have to use physical force and risk the emergence of a fight (see Chapter 3 for more discussion).

4.3 Design and hypotheses

I employ a three-step procedure that was fully developed by Gambetta and Székely (2014), Chapter 2 here, to investigate sign-signal interactions. In the first step of this procedure, which is shared with that used by others (Fehrler & Przepiorka, 2013; Gambetta & Przepiorka, 2014), uninformed subjects generate a sign of their qualities. Second, informed subjects generate a signal of these same qualities. Third, subjects participate in a strategic game in which subject-receivers assess the type of senders, using the sign-signal combinations of senders, and subject-senders reveal their type.

Uninformed subjects—uninformed for they do not know what will follow—carry out a wall squat exercise (more on this later). Next, it is revealed to them that they will participate in a strategic game, later in the experiment, as a sender and as a receiver. In addition, they are told that as a sender they will send their receiver-opponent an indicator of their initial squat time, that before the strategic game they will do the squat again, and that they will also send an indicator of this second squat to their opponent. As a receiver, they know that they will play the complementary role. Finally, they participate in the 'resident-rookie game'. To my knowledge, this is the first experiment in which the squat exercise is combined with a strategic game of conflict.

I use this procedure to explore sign-signal combinations related to a particular clustering of qualities—of which toughness is one—that correspond to peoples' willingness to engage in physical confrontation. However, by changing the tool used to measure and create subjects' sign and signal—from a squat to a dictator game for instance—and the strategic game employed—from my resident-rookie game to a trust game—it is readily adaptable to exploring the sign-signal interactions of other qualities.

Another noteworthy aspect of the design is its use of the 'true type' method. As articulated by Gambetta and Székely (pp. 284, 295) (also Section 2.3 of this thesis), the core concept is that subjects who are one type in the real world are allocated that type in the experiment. This endogenous allocation of type may encourage subjects to behave naturally—as they do in real life—in the experiment, increasing the external validity of the experimentally obtained results. It is a simple, but compelling, design feature that has wide applicability in experiments using Bayesian games.

4.3.1 Subject pool

Only males, between the ages of 18-40 were recruited as subjects. This was done for two reasons. First, as the experimenter is male, this allowed interaction effects between the experimenter and the subjects to be avoided. Second, the chosen age-range reduces physiological differences in strength and fitness—those within this group are, on average, comparable in isometric muscular strength and aerobic capacity (Asmussen & Heebøll-Nielsen, 1962, pp. 167, 168; Robinson, 1938).

4.3.2 Veiled squat

I employ a non-harmful, but pain generating, physical exercise called the 'wall squat', or simply 'squat', that has only been used by researchers a handful of times (Bastian, Jetten, & Ferris, 2014; Cabanac, 1986; Harrison & El Mouden, 2011; Harrison, Sciberras, & James, 2011; Madsen et al., 2007).[83, 84]

[83] To my knowledge, these few references comprise an exhaustive list of the papers that have previously employed the squatting exercise.
[84] I have been told that it is (or was) also used in the UK armed forces, and as punishment for fare-dodgers on trains in India.

To do the squat, subjects stand about two-feet away from a wall, lean back against it, and lower themselves down until there is a 90-degree angle between their shins and their thighs so that their shins are parallel to the wall and their thighs are parallel to the floor. They must place their hands flat on the wall behind them to avoid cheating (Figure 7).

Figure 7. The wall-squat.

In the veiled squat, subjects are paid €0.02 for every second that they hold the sitting position for. For instance, if a subject squats for 2 minutes (120s), he earns €2.40. The better they are able to do the exercise and the more motivated they are by money, the longer they hold the position for and the more money they earn (see the experimental instructions in Appendix E).[85]

[85] In both the veiled and unveiled squats, subjects wore sleeping masks and ear protectors in order to prevent them from perceiving one another's squatting times. In addition, the opportunity for subjects to do the veiled and unveiled squat was fixed at six minutes; irrespective of whether people were actually squatting or not, the squatting opportunity was not declared over until six minutes had passed. This was done to prevent subjects in the same squatting group from inferring what the longest squatting time was based on how long they all remained in the squatting room. The six-minute rule was public knowledge to the subjects.

4.3.3 Unveiling and the unveiled squat

Subjects are 'unveiled' and fully informed about the experiment. They learn that their time in the veiled squat is used to assign them a 'grade', A, B, or C, depending on which percentile in the session their veiled squat time corresponds to. If it is in the top 1/3rd they are allocated an 'A' grade, if it is in the middle 1/3rd they receive a 'B' grade, and if it is the bottom 1/3rd they receive a 'C' grade. The grade that they obtain in the veiled squat is the sign that they send to their receiver-opponent in the resident-rookie game, and into which noise is added. Sending squatting prowess as grades representing tertiles ensures that subjects have common priors about the distribution of types (1/3rd of each). Simply transmitting squatting time in seconds would not allow subjects to have common prior beliefs over squatting ability—an essential part of counter- and signalling theory.

By removing the possibility that the rookie sends a 'B' grade, and instead randomly deciding whether he sends an 'A' grade or a 'C' grade, substantial noise is added into the transfer of the sign. From the perspective of a receiver, observing an 'A' sign implies that the sender actually received an 'A' grade with 2/3rd probability and a 'B' grade with a 1/3rd probability. Observing a 'C' sign implies that the sender actually received a 'C' grade with a 2/3rd probability and a 'B' grade with a 1/3rd probability.

Before moving to the resident-rookie game, all subjects repeat the squat, this time *without* receiving payment for doing so.[86] By removing the direct financial incentive for squatting, Rookies can have only—assuming they care only about money—a signalling incentive for squatting, isolating the signalling motive. They are allocated an additional grade, also either A, B, or C, based on the distribution of times that was achieved in the

[86] Initially being paid and subsequently not being paid corresponds to reality too. Tattoos, scars, and a tough demeanour are often acquired for immediate benefit (here money), while subsequent signalling has no immediate gain but a gain instead through its information content.

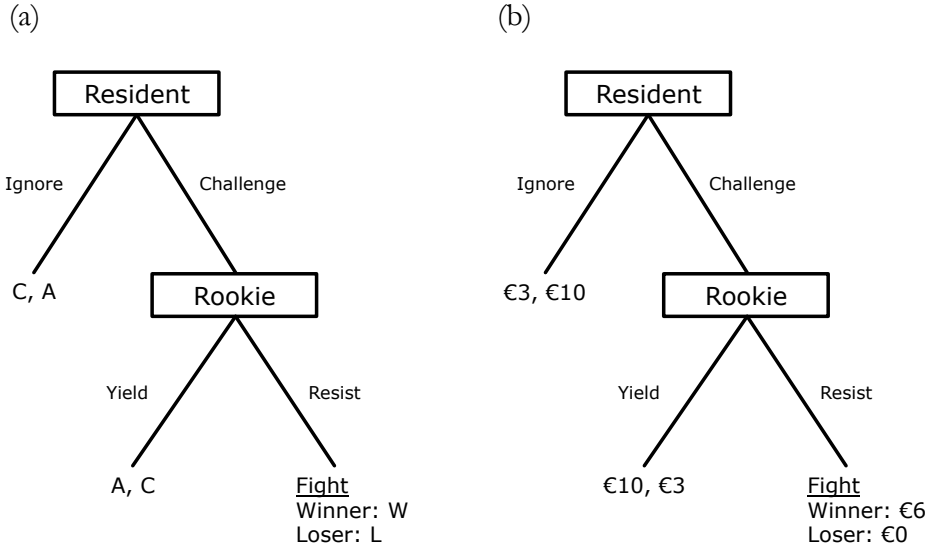
veiled squat. This second grade is their signal, which they send precisely, to their receiver-opponent.

4.3.4 Resident-rookie game

In the final part of the experiment, subjects participate in the resident-rookie game. This is a modified version of the hawk-dove game that I designed to capture some essential features of interpersonal conflict situations. It is a unique strategic game, as far as I know, in that it combines monetary outcomes and a physical competition in determining an outcome.

It is a two-player sequential game in which the resident, who is an established inmate, moves first, and can challenge or ignore the rookie and if challenged, the rookie, who is a new prisoner, moves, and decides whether to resist or yield to the resident's challenge (Figure 8).

Figure 8. (a) The resident-rookie game and (b) the specification of the resident-rookie game that I use in my experiment. A resident has to believe that the probability of the rookie resisting and winning is less than 0.7 before challenging (Appendix E).



The best outcome for the resident is to challenge the rookie and the rookie to yield to his demands: a situation of exploitation. His second best outcome is to fight and beat the rookie; next, ignore the rookie and leave him alone. Worst of all is to fight and lose to the rookie. Put succinctly:

$U_{RE}(\text{Exploitation}) > U_{RE}(\text{Fight and win}) > U_{RE}(\text{Status quo}) > U_{RE}(\text{Fight and lose})$.

The rookie's best outcome is to be ignored by the resident. Next is fighting and beating the resident and being exploited by the resident is his third best outcome. Finally, fighting and losing to the resident is his worst outcome:

$U_{RO}(\text{Status quo}) > U_{RO}(\text{Fight and win}) > U_{RO}(\text{Exploitation}) > U_{RO}(\text{Fight and lose})$.

Both players' payoffs may be represented by A>W>C>L. The two payoffs, 'A' and 'C', are directly monetary ones, while to determine W and L a 'squat-off' needs to take place. This involves both subjects who ended up at that outcome simultaneously doing the squat, and the subject who squats for longer wins and the one who does it for shorter loses. While squatting, subjects cannot see, or perceive, the other against whom they are competing.

Subjects make all of their decisions in one experimental session. In a second, follow-up session, subjects obtained their outcomes: receiving their earnings directly or, if a fight was the outcome, first participating in a squat-off and then receiving payment.[87]

[87] Separating the experiment into two sessions allows subjects to signal honestly in their veiled squat. If they made their decisions in the same sessions as they were signalling, some subjects may choose to signal less in order to save strength for a squat-off, making the signalling decision more complex than that of information.

As this game is designed for a context of (counter)signalling and information, I had to incorporate two important features. One is that rookies do not have incentives to signal dishonestly. A higher type rookie always wants to be perceived as a higher type rookie for he strictly gains more from deterring a resident than he does from fighting him.[88] The other, complementary feature is that both players have incentives to accurately assess the other's ability in fighting. Misidentification of type, both too high and too low, implies, respectively, either a missed opportunity or a waste of energy and pain in a fight.

I use the strategy method to elicit subjects' choices (Mitzkewitz & Nagel, 1993). Rookies make a single choice to resist or yield and residents make a decision to challenge or ignore for each of the possible, up to six, sign-signal combinations. The order that the sign-signal combinations are shown to residents is randomised across session.

Subjects are allocated to a control, a signalling treatment, or a countersignalling treatment, in a between-subject design. In the signalling treatment, rookies send their unveiled grades to the residents, and the residents observe only the unveiled grade of the rookies. In the countersignalling treatment, rookies emit both their veiled and unveiled grade to residents, and residents observe both. No information is transmitted in the control group.

In this setup, only rookies send information and only residents receive information. The key reason for implementing this is that it allows me to precisely identify how information affects the behaviour of receivers. Allowing both rookies and residents to emit and observe information would make it harder to disentangle how, and to what

[88] In prison, countersignalling may get the sender benefits from here too by helping them to discriminate between naïve and astute receivers. Astute receivers should treat them as high types and naïve receivers treat them as low types, in which case it could be used as a 'lure' to entrap naïve receivers.

exactly, receivers respond. Another compelling reason for implementing this is because new prisoners who enter a facility are unable to observe the signs and signals that multitude of residents have. They are overwhelmed by the many residents and have no possibility of coherently comprehending information about them. Residents, in contrast, can observe the information that rookies emit for, at any one time, they are only few in number.

4.3.5 Belief measurement

In order to gain a deeper understanding of the decision making process of both residents and rookies, I elicit rookies' beliefs about the probability that they will beat others in a squat off, residents' beliefs about the opponent-rookie's probability of resisting, and residents' beliefs about the probability that they will beat their opponent-rookie in a squat-off. I also ask subjects to report their reasoning behind their unveiled squat time.

4.3.6 Squatting validation

As the squatting exercise has rarely been used before, and never before as a tool for signalling, I investigate what it measures and the information it could be expected to transmit by asking subjects to report their age, height, weight, frequency of exercise, pain resistance, and fitness level, and fill out the Buss-Perry Aggression Questionnaire (Buss & Perry, 1992) (BPAQ). I also record subjects' second-to-fourth digit ratio (2D:4D).

A substantive amount of work in psychology, since the early 2000s, has found correlations between the ratio of the second digit (index finger) to the fourth digit (ring finger) and a variety of aggressive behaviours in men, but not women (Hönekopp & Watson, 2011). The basic theory is that males' exposure to higher levels of prenatal testosterone causes greater elongation of the fourth-digit relative to the second-digit.

Although the exact mechanism is not yet clear, experimental work on mice suggests that it is a result of a balance between androgen (of which testosterone is one) and oestrogen (Zheng & Cohn, 2011), the relationship between the ratio and behaviours appears to hold empirically (Hönekopp & Watson, 2011).[89]

Two fascinating examples, relevant here, concern athletes' and high frequency traders' outcomes and behaviours. The main result concerning the former is that successful male athletes have lower 2D:4D ratios than less successful athletes, or student-controls, which—as the authors argue—has implications for male-male fighting (Manning & Taylor, 2001). In the latter study, '2D:4D predicted the traders' long-term profitability as well as the number of years they remained in the business' (Coates, Gurnell, & Rustichini, 2009, p. 623).

Using a ballpoint pen, the basal crease of subjects' second and fourth digit on their right palm is marked (e.g. Bailey & Hurd, 2005). Next, they lay their right hand, palm-up, on a table, and I photograph that palm from directly above. Finger lengths are measured using GNU Image Manipulation Program (v. 2.8.10), from the middle of the basal crease to the middle of the top of each finger. Ratios are calculated by dividing the length of the second digit by the length of the fourth digit.

The Buss-Perry Aggression Questionnaire (Buss & Perry, 1992) is a widely used tool used to measure aggression that is composed of four sub-scales, physical aggression, verbal aggression, anger, and hostility. It has been found to correlate with multiple

[89] Two points should be kept in mind regarding this literature. First, the strength of the association between 2D:4D and an outcome is typically low, in their meta-review Hönekopp and Watson report a correlation coefficient of about -0.06 (p. 381). Second, the studies on humans are entirely non-experimental ones.

aggressive behaviours (e.g. Archer & Webb, 2006; Kim, Namkoong, Ku, & Kim, 2008; Palmer & Thakordas, 2005) and at least one paper finds a relationship between 2D:4D and the Buss-Perry Aggression Questionnaire (Bailey & Hurd, 2005).

4.3.7 Ethics

This design was reviewed by the Centre for Experimental Social Sciences (Nuffield College, Oxford) review panel and received ethical approval (see Appendix E for statement of ethical approval). Furthermore, the experiment was conducted at the Bologna Laboratory for Experiments in Social Sciences (University of Bologna) that accepted the ethical approval provided by the Centre for Experimental Social Sciences.

To ensure that there was no psychological or physical harm to subjects, I took a number of specific measures including:

- Highlighting the physical nature of the experiment in the recruitment material and requesting that people with back or knee problems not to register.
- Re-iterating to subjects multiple times during the experiment that people with back or knee problems should not participate, and that they were all free to leave at any point and they would receive a €4 show-up fee in that case. One of these instances included a time directly after subjects had practiced the squat exercise.
- Explaining the purpose of sleeping masks and ear protectors.
- Keeping a careful watch on the subjects while squatting and being ready to assist.

The exercise itself is painful but not harmful and does not pose a risk that is greater than that experienced by most of us in everyday life (e.g. when walking up or down stairs). As a final point, a number of recent studies have used the same procedure (Bastian et al., 2014; Harrison & El Mouden, 2011; Harrison et al., 2011; Madsen et al., 2007).

4.3.8 Hypotheses

Veiled squatting is a reliable *sign* of toughness so rookies who squat for longer when veiled are likelier to resist (H1a). Residents recognise and respond to this with fewer challenging 'tougher looking' rookies (H1b).

Unveiled squatting is a reliable *signal* of toughness so rookies who invest more in squatting are more likely to resist (H2a). Residents recognise H2a so that fewer residents challenge 'tougher looking' rookies (H2b).

Rookies realise the strategic value of signalling their toughness and will invest in signalling (H3). Specifically, rookies invest more in signalling their toughness in the signalling only treatment than in the control. In the former, they have an incentive to signal while in the latter, as no information is sent to residents, they do not have an incentive to signal.

When both a sign and signal are present in an interaction, countersignalling predicts that medium type rookies signal the most (H4). Both low and high types signal less than medium types.

In addition, high type rookies signal less in the countersignalling treatment than in the signalling treatment (H5). In the former, high types countersignal, while in the latter they signal.

Residents in the countersignalling treatment will evaluate rookies' types, according to countersignalling theory, as shown in Table 12 (H6).

Table 12. Predicted beliefs of residents by the veiled-unveiled grade combinations of rookies.

Veiled	Unveiled		
	A	B	C
A	High	Medium	High
C	No prediction	Medium	Low

Information about rookies decreases the number of fights that occur (H7). Opacity leads to more conflicts because it is harder to decide who would win and who would lose in a fight.

4.4 Results

The experiment was held at the Bologna Laboratory for Experiments in Social Sciences (BLESS) in May and June 2014. It was programmed and conducted with the software z-Tree (Fischbacher, 2007) and analysed using Stata v13.1.

In total, 144 subjects participated in the experiment. Five subjects were removed from the dataset because of issues in the collection of their data leaving information on 139 subjects: 26 in the control session (C), 49 in the signalling session (S), and 64 in the countersignalling session (CS) (Table 13). This gives me 139 rookie decisions, 557 resident decisions (Table 14), and 38 squat-offs to analyse.[90, 91, 92] Full model outputs for the results reported here are shown in the tables in Appendix D.

[90] Some squat-offs that were meant to take place could not be carried out, because the involved subjects did not turn up for their second session (12 subjects).
[91] The answers to the control questions strongly suggest that the subjects understood the experiment. Most (70%) subjects answered all of the control questions correctly, 18% answered 5/6 of the control questions correctly, 10% answered 4/6 of the control questions correctly, and only 2% answered 3/6 correctly. After they responded to the control questions, I read out and explained the correct answers.
[92] The median and mean session size is 16 and 15.44 subjects respectively.

Table 13. The number of observations on rookies' behaviour according to treatment and veiled grade.

Veiled grade obtained	Treatment			Total
	C	S	CS	
A	10	16	23	49
B	8	17	20	45
C	8	16	21	45
Total	26	49	64	139

Table 14. The number of observations collected on residents' behaviour according to grade combination and treatment.

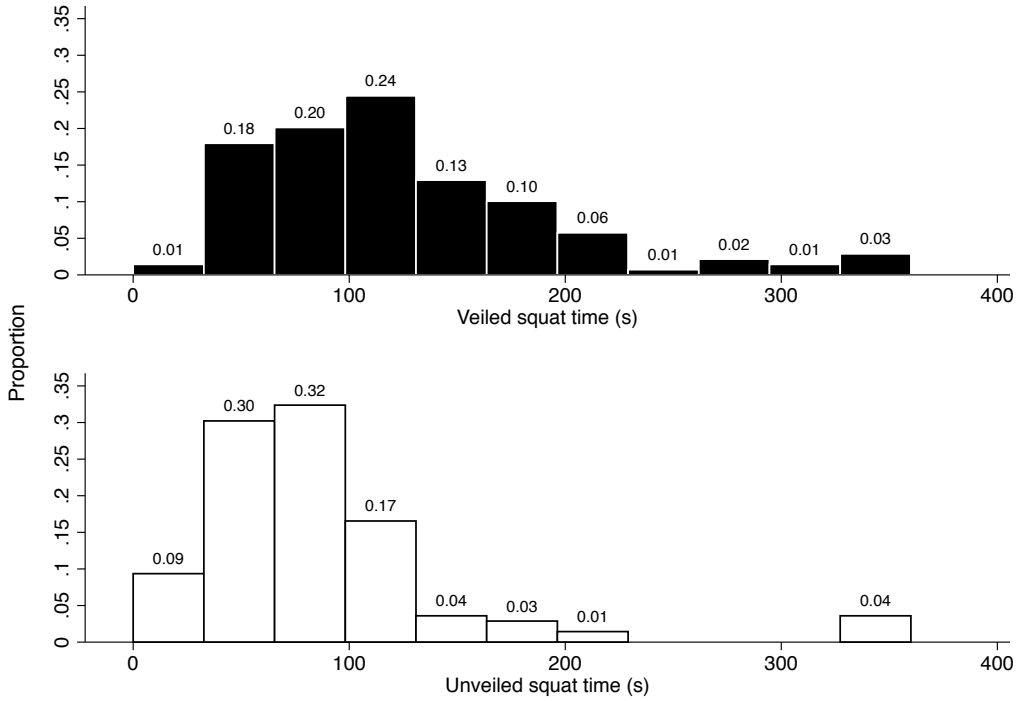
Veiled grade observed	Unveiled grade observed				Total
	A	B	C	None	
A	64^{CS}	64^{CS}	64^{CS}	-	192
B	64^{CS}	64^{CS}	64^{CS}	-	192
None	49^{S}	49^{S}	49^{S}	26^{C}	173
Total	177	177	177	26	557

Superscripts indicate the treatment group in which the subjects participated.

The mean veiled squat time of all subjects is 125.6s, the mean unveiled squatting time is 89.1s, and both distributions are positively skewed (Figure 9). The mean of the latter is significantly less than the mean of the former (t=7.20, $d.f.$=138, p<0.001).[93]

[93] The mean squatting time of subjects in the veiled section (125.6s) is broadly comparable to some of the average squatting times found by previous work—experiments 2 and 3 by Madsen et al. (2007, pp. 347, 352). Yet substantially different to those found by others (experiment 1 in the same paper (p. 345), and Cabanac's results (1986, p. 42)). The reason for these differences is currently unclear.

Figure 9. Veiled and unveiled squatting times (based on C, S, and CS groups; n=139).



Using the veiled squatting times of all subjects—for when veiled they all have the same incentives—I examine what squatting time captures. Bivariate correlations between veiled squat and the relevant collected measures (age, body mass index, frequency of exercise, ability to withstand pain, fitness, the four sub-scales of the BPAQ, the overall BPAQ score, and 2D:4D) reveal that five variables significantly correlate with veiled squatting time. Age (r=-0.20, p=0.018) and 2D:4D (r=-0.21, p=0.013) [94, 95] are negatively related to veiled squatting time. Frequency of exercise (r=0.17, p=0.021), ability to withstand pain (r=0.33, p<0.001), and fitness (r=0.35, p<0.001) are positively related to veiled squatting time. Older participants are somewhat worse at squatting, as are those

[94] The photographs of two subjects' palms were blurry and were removed before analysis.

[95] 2D:4D does not correlate with other important outcomes. It is not associated with unveiled squat time (r=-0.08, p=0.36), choosing to resist (r=-0.11, p=0.18), challenging (r=-0.02, p=0.65), ending up in a squat-off (r=-0.10, p=0.25), the Buss-Perry Aggression Questionnaire physical aggression subscale (r=0.01, p=0.92), the verbal aggression subscale (r=-0.01, p=0.95), the anger subscale (r=0.04, p=0.66), the hostility subscale (r=0.05, p=0.60), and the overall score (r=0.03, p=0.72).

who have a longer index finger relative to their ring finger, while subjects who exercise more frequently, who are more pain resisting, and who are fitter are better at squatting. The sub-scales of the BPAQ and overall BPAQ, are unexpectedly, *not at all* correlated with veiled squatting ($-0.05 \leq r \leq 0.07$, $0.390 \leq p \leq 0.991$).[96] The correlation coefficient for body mass index is negative but not significant ($r=-0.15$, $p=0.086$).

Simple pairwise correlations allow possible confounders to interfere. Simultaneously including all of the variables in a single regression estimates the relationship that each of them has with veiled squatting time, controlling for all of the others. When this is the case, only the ability to withstand pain ($b=13.45$, $t=2.11$, $p=0.037$) and 2D:4D ($b=-283.74$, $t=-2.25$, $p=0.026$) remain significantly linked at the 5%-level, albeit fitness comes close ($b=13.78$, $t=1.87$, $p=0.064$) (Table D1). Every one-point increase in pain resistance,[97] is predicted to correspond to an 13s increase in squatting time increases, meaning that someone with the minimum value of 1 is predicted to squat for 52s less than someone who has the maximum value of 5. 2D:4D varies from a minimum of 0.824 to 1.086 at maximum, and based on this, someone with the lowest ratio is predicted to squat for 74s more than someone with the highest ratio. Fitness has a similar effect to pain resistance, at 14s extra squatting per score increase.[98]

[96] The overall Buss-Perry Aggression Questionnaire score is not linked to other important variables: unveiled squat time ($r=0.08$, $p=0.37$), choosing to resist ($r=0.02$, $p=0.81$), challenging ($r=0.01$, $p=0.82$), and ending up in a squat-off ($r=0.04$, $p=0.61$).

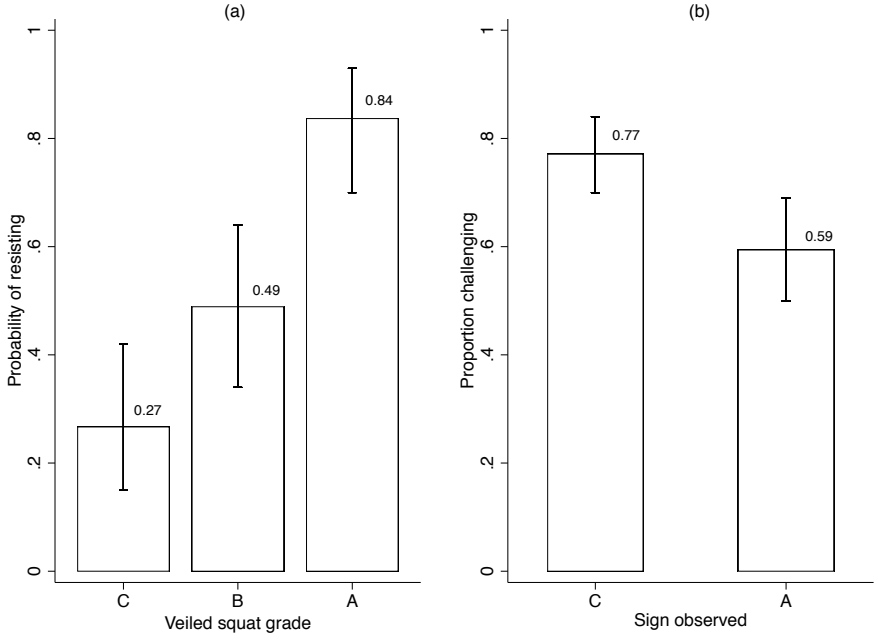
[97] Recorded on a 1-5 scale where 1 means that the statement 'I am good at resisting physical pain' is 'extremely unlike me' and 5 means that the statement is 'extremely like me'.

[98] Recorded using a 1-5 scale where 1 indicates that the statement 'I am physically fit' is 'extremely unlike me' and 5 indicates that the statement is 'extremely like me'

Taken together, these results imply that squatting measures a combination of one's resistance to pain and fitness, or toughness. Moreover, squatting time is not linked to aggressiveness.[99, 100]

H1a. Veiled squatting is a reliable sign of toughness so rookies who squat for longer when veiled are likelier to resist.

Figure 10. (a) The proportion of rookies choosing to resist by grade obtained when veiled (based on C, S, and CS groups; n=139). (b) The proportion of residents choosing to challenge based on the observed sign of rookies (based on CS group; n=384). Error bars represent 95% C.I.s.



Among all rookies, those who squat for longer when veiled are more likely to resist ($\exp(b)=1.02$, $z=4.68$, $p<0.001$) (Table D4). To help interpret this number, I group veiled

[99] Greed may be a factor in determining squatting time: greedier people squat for longer, as they have a greater subjective incentive for taking pain. Whilst possible, there should be no correlation between toughness and greed, so although this adds noise to the measure, it does not systematically bias it.

[100] Subjects filled out the questions concerning the relevant variables at the end of the experiment. This leaves open the possibility that their answers are influenced by the grade-feedback that they receive after each squat.

squatting times according to the corresponding allocated grade and plot the proportion of rookies who resist in each category (Figure 10a). Only 27% of rookies who receive a C grade resist, while 49% and 84% or rookies who obtain a B grade and an A grade, respectively, resist. The size of this effect is quite astonishing: a 57% difference.

Not only are rookies who resist for longer more likely to resist, they are also more likely to win in case a squat-off actually occurs. Subjects who obtained a longer veiled squat time as a rookie are much more likely to win in squat-offs that ensue than subjects who obtained shorter veiled squat times. In the vast majority of the squat-offs (79%), the subject with the longer veiled squatting time won (Fisher's exact test, p=0.001).

H1b. Residents recognise and respond to this such that fewer challenge rookies with higher signs.

Residents respond as expected to rookies' signs. Based on residents' responses to rookies in the countersignalling treatment—this is the only treatment in which residents observed rookies' veiled grades—I find support for this hypothesis. Most residents challenge rookies who emit a C grade (77%) while much fewer residents challenge rookies who emit an A grade (59%) (exp(b)=0.44, z=-6.07, p<0.001) (Table D5). (Figure 10b).

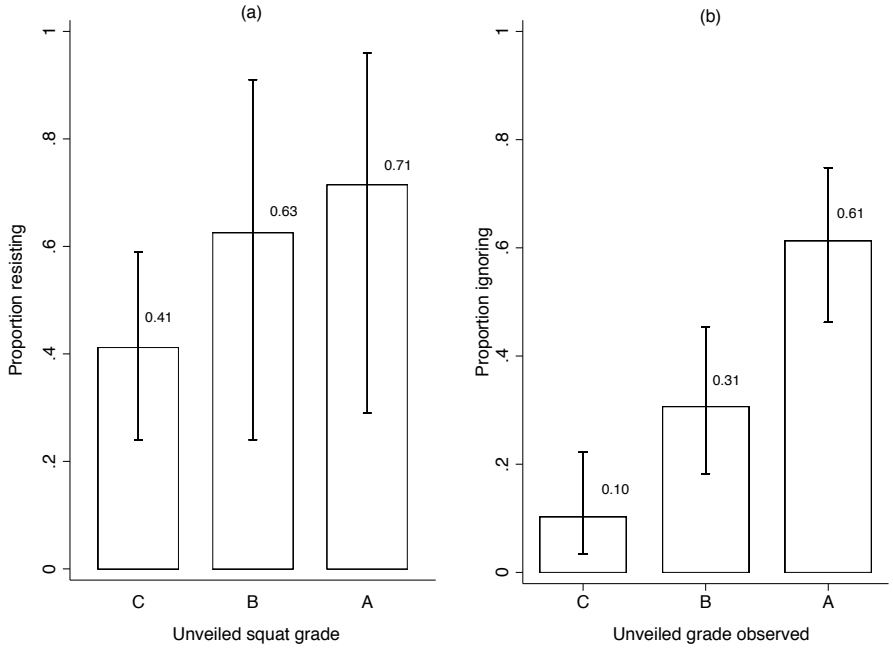
Although I never elicited the behaviour of receivers to signs alone, the different signal grades are equally represented in each of the two signs. For both C and A signs, 1/3rd of the responses of residents is towards a C signal, 1/3rd is towards a B signal, and 1/3rd is towards an A signal. Combined with the finding that there is no interaction effect in receivers' responses to veiled-unveiled grades (see Figure 17), this implies that the

difference between the responses to the two signs cannot be down to the signal grades that they are combined with.

Interestingly, the percentage of residents who challenge rookies emitting C grades (0.77) is close to the percentage of rookies who emit a C grade and who yield (66%) [1-(2/3*0.27+1/3*0.49)]*100, which suggests that the expectations of residents and rookies are near identical regarding the meaning of a C grade. Regarding rookies emitting an A grade, there is some disparity. While 59% of residents challenge, only 28% of rookies emitting an A grade choose to yield [1-(2/3*0.84+1/3*0.49)]*100.

H2a. Unveiled squatting is a reliable signal of toughness so rookies who invest more in squatting are more likely to resist.

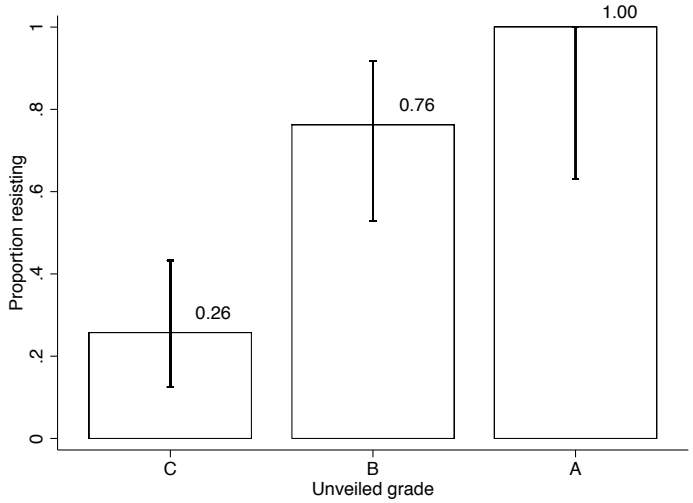
Figure 11. In the signal only treatment, the proportion of (a) rookies who resist according to their unveiled grade (based on S group; n=49). (b) Residents who ignore according to the observed unveiled grade (based on S group; n=147). Error bars represent 95% C.I.s.



Rookies who send a longer signal without a sign, those in the signal treatment, are likelier to resist at the 10% level (exp(b)=1.01, z=1.81, p=0.071) (Table D4): 41% of rookies who obtain a C grade choose to resist, while 71% of those who achieve an A grade do so (Figure 11a).

Rookies who send a longer signal with some sign (those in the countersignalling treatment) are also likelier to resist (Figure 12) (exp(b)=1.05, z=3.52, p<0.001) (Table D4). Only 26% of rookies with a C unveiled grade resist, in contrast every single rookie with an A unveiled grade resists (Figure 12).

Figure 12. The proportion of rookies resisting in the countersignalling treatment according to their unveiled grade (based on CS group; n=64). Error bars represent 95% C.I.s.



Subjects with longer unveiled squatting times obtained as rookies, are much more likely to win in case of a squat-off. In every single squat-off, the subject who obtained a higher unveiled squat time won in an ensuing squat off (Fisher's exact p<0.001).[101]

[101] Based on the congruence of signalling behaviour by rookies in the signalling and countersignalling group, I combined their behaviour for this analysis.

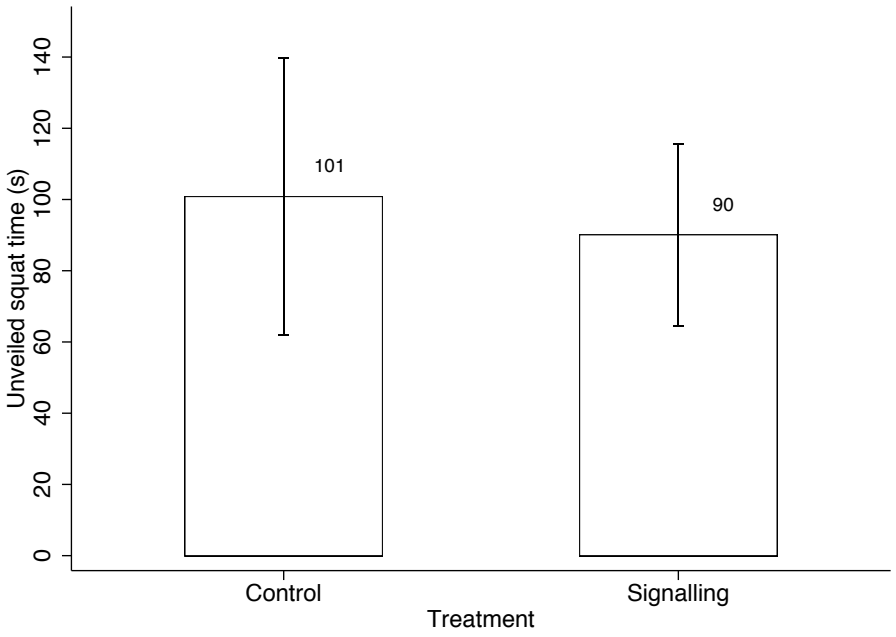
H2b. Residents recognise H2a so that fewer residents challenge rookies with higher signals.

Fewer residents challenge rookies with higher unveiled grades in both the signal only treatment ($\exp(b)=0.27$, $z=-5.04$, $p<0.001$) and in the countersignalling treatment ($\exp(b)=0.43$, $z=-6.97$, $p<0.001$) (Table D5). Most residents avoid rookies who emit an A grade (61%), while almost all residents challenge a rookie with a C grade (90%) (Figure 11b).

Depending on which sign the signal is combined with determines the exact proportion of residents who choose to challenge in the countersignalling treatment. However, for both signs it is unequivocal that fewer residents challenge rookies with higher signals (Figure 17).

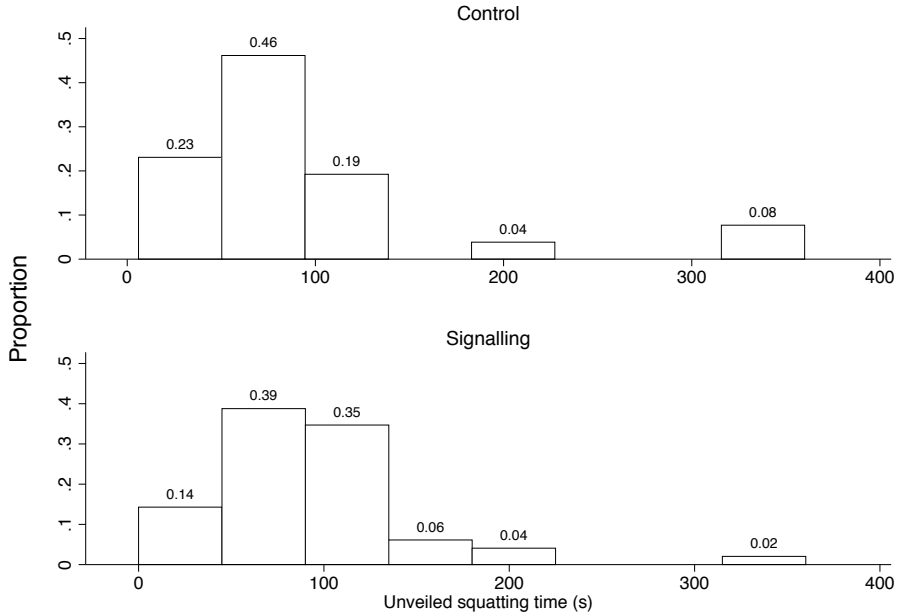
H3. Rookies realise the strategic value of signalling their toughness and will invest in signalling.

Figure 13. Unveiled squat times for the control and signalling treatments (based on C and S groups; n=75). Error bars represent 95% C.I.s.



Surprisingly, and contrary to the hypothesis, the mean unveiled squatting time for rookies in the signalling group (90s) is similar to the unveiled squatting time for rookies in the control group (101s) (b=-10.74, t=-0.65, p=0.52) (Figure 13 and Figure 14) (Table D6).

Figure 14. The distribution of unveiled squat times by treatment (based on C and S groups; n=75).



A potential explanation for this result is that subjects in the control condition had strong internal motivations for doing the unveiled squat—competiting with themselves and others to obtain a higher grade—that was replaced by a strategic motivation in the signalling condition. In this case, subjects realise the strategic value of signalling, supporting the hypothesis, but the average squatting of both groups ends up at the same level because one motivation replaces the other. If true, self-reported effort in the unveiled squat will be the same while the reported motivations for doing the squat will be different.

Indeed, this is what I find. In the control condition 23% (6/26) report putting in more effort, 35% (9/26) report putting in the same effort and 42% (11/26) report putting in less effort. In the signalling treatment, 29% (14/49) report putting in more effort, 33% (16/49) put in the same effort, and 39% (19/49) put in less effort. There are no statistically significant differences.

In contrast, regarding decision making, only 12% (3/26) report a strategic motivation in the control group while 41% (20/49) do so in the signalling group ($\boldsymbol{X}^2$ =6.85, p=0.009). Among those who increase their effort relative to their initial squat in the signalling treatment, 57% (8/14) explicitly report strategic motivations.[102]

H4. When rookies emit both a sign and a signal, countersignalling predicts that medium type rookies signal the most.

For rookies in the countersignalling treatment, there is a strong positive linear relationship between veiled squat and unveiled squat time (b=0.57, t=6.40, p<0.001) (Table D7) (Figure 15), which contradicts the hypothesis. Medium types squat for as long as or less than high types (b=30.03, t=1.72, p=0.090) and more than low types (b=-37.35, t=-2.09, p=0.040) (Figure 16) (Table D8).

[102] I investigated two other possibilities for the lack of expected difference. First, potentially, because of the small number of sessions that I ran, some confounders prevent the treatment's effects from being observed. However, controlling for subjects' age and time of day that the session took place does not lead to a change in results (t=0.79, $d.f.$=74, p=0.43)—although the relationship now goes in the expected direct of those in signalling treatment squat for longer. Adding further controls—body mass index, self-rated pain resistance, and self-rated fitness—does not change the results either. Second, a small number of outliers in the control group could increase the mean of that group and obscure a difference. Thus, I look at the distribution of unveiled squat times according to treatment. For the control, the standard deviation is 86.7 and for the signalling treatment it is 55.6. It is clear that a small group of subjects in both treatments squat for longer than 300s. Assuming that these subjects are extremely competitive and do not respond to strategic incentives, and on this basis removing them, still does not lead to a difference between the treatments. Both investigations suggest that the result is robust.

Figure 15. Unveiled squatting time by veiled squatting time in the countersignalling treatment (based on CS group; n=64).

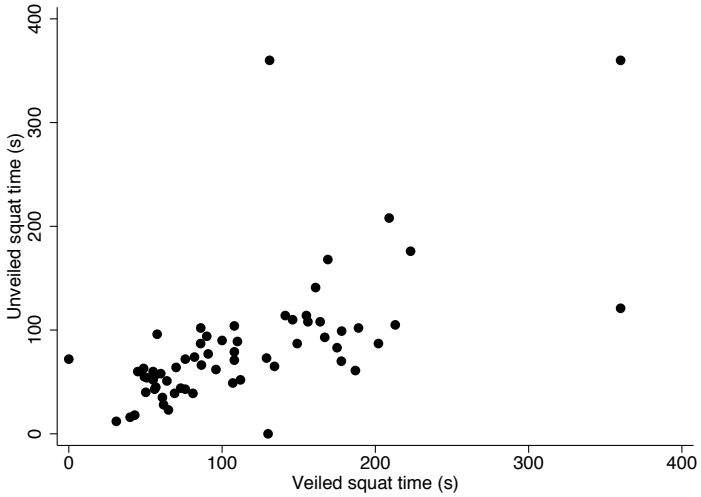
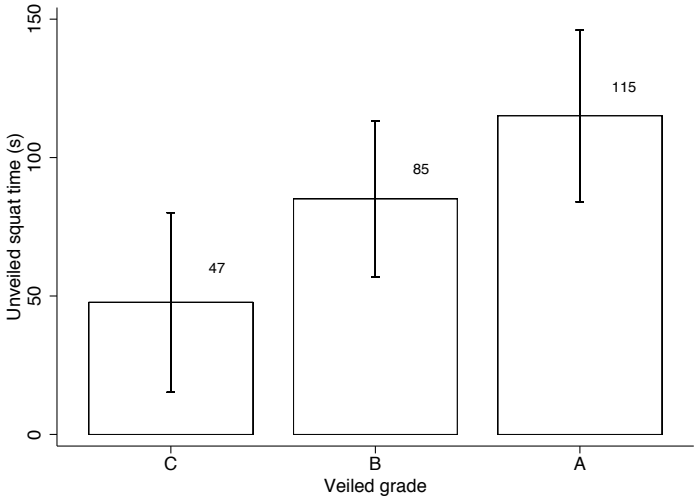


Figure 16. Unveiled squat time according to type in the countersignalling treatment (based on CS group; n=64). Error bars represent 95% C.I.s.



H5. High type rookies signal less in the countersignalling treatment than in the signalling treatment.

High type rookies (who obtained an A unveiled grade) in the countersignalling treatment have an average unveiled squatting time of 115s, while those in the signalling treatment

have an average unveiled squatting time of 113s. There is no difference between the two groups (b=2.40, t=0.10, p=0.921) (Table D9).

H6. Residents in the countersignalling treatment will evaluate rookies' types, according to countersignalling theory.

The predicted and actual behaviour of residents towards rookies is shown below in Table 15.

Table 15. Predicted and actual behaviour of residents towards rookies. Percentages represent residents who chose to ignore.

Veiled	Unveiled		
	A	B	C
A	59% (38/64)[a]	41% (26/64)[b]	22% (14/64)[c]
	High	*Medium*	*High*
C	38% (24/64)[b]	25% (16/64)[c]	6% (4//64)[d]
	No prediction	*Medium*	*Low*

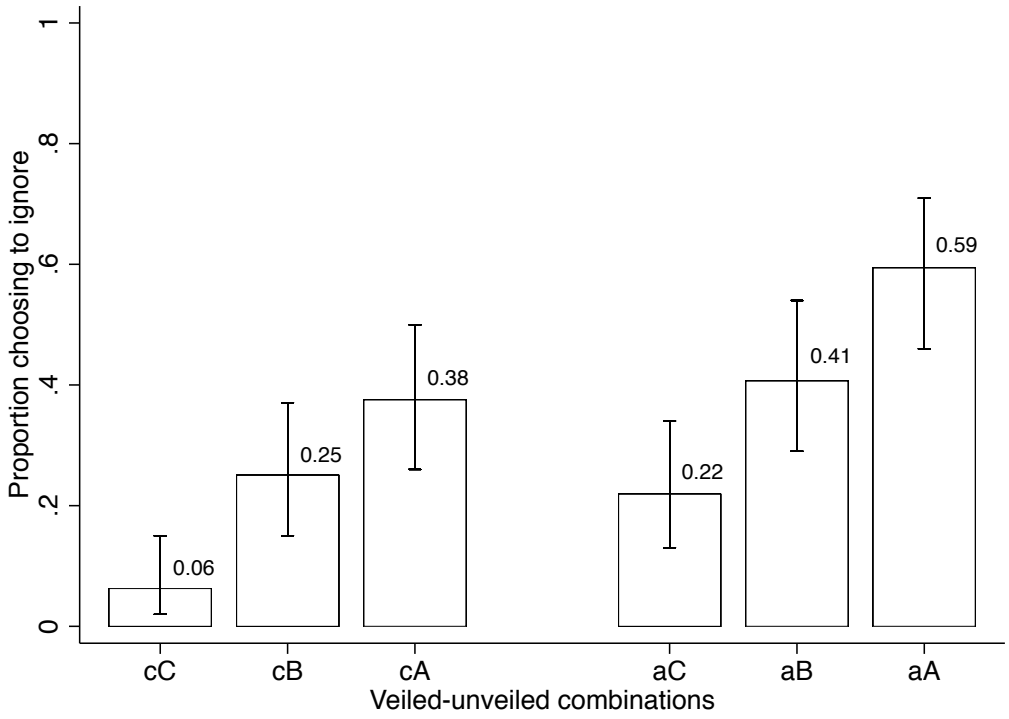
N.B. Predicted treatment is italicised. Groups that differ from one another are indicated with the superscripts a-d (Table D10).

There are two points to consider as I report residents' behaviour in the countersignalling treatment:

- Whenever I specify a veiled-unveiled grade combination, I use the format 'xY', and indicate the veiled grade with the first, lower-case, letter and the unveiled grade with the second, capitalised, letter. For instance, bC would indicate an veiled grade of B and a unveiled grade of C.
- When I say 'xY' rookie, I mean that this is the veiled-unveiled grade combination of a rookie that the residents observed, irrespective of whether the rookie actually obtained that veiled grade or only the noise included in the transmission made it seem so.

An interesting result is that residents broadly evaluate rookies in four groups, as opposed to three. cC rookies are unequivocally treated as low types—a staggering 94% of residents choose to challenge them—cB and aC are treated as medium-low types, cA and cB are treated as medium-high types, and aA rookies are treated as the kings of squatting (results are graphically displayed in Figure 17).

Figure 17. The proportion of residents choosing to ignore according to the veiled-unveiled combinations of rookies (based on CS group; n=384). Error bars represent 95% C.I.s.



The cA combination is an odd one, positing—with the exception of rookies who obtain a B veiled grade and happen to emit a C veiled grade—that rookies squat terribly when veiled and excellently when unveiled. As countersignalling makes no prediction for the cA combination, I do not discuss these results further. Indeed, only one rookie emitted a veiled grade of C and an unveiled grade of A in my experiment.

Comparing the actual treatment of rookies to their predicted treatment, I find a strong match. For all of the predicted combinations, except one, there is rough agreement: cCs are treated as low, cBs are treated as medium-low, aBs are treated as medium-high, aAs are treated as high. There is disagreement only between the predicted and actual treatment of aC rookies: these represent the countersignalling rookies who should be treated as ideal high types, yet, in reality, they are assigned to a lowly medium-low status.

H7. Information about the squatting ability of rookies decreases the proportion of interactions that end in fighting.

There is a negative correlation between the presence, or absence, of information and the proportion of rookies who end up taking part in a squat-off (r=-0.26, $d.f.$=132, p=0.003), and as shown by a Chi squared test, the squat-offs are not distributed randomly ($\boldsymbol{X}^2$=9.43, p=0.009).[103]

Table 16. The proportion and number of subjects who ended up in a squat-off according to treatment.

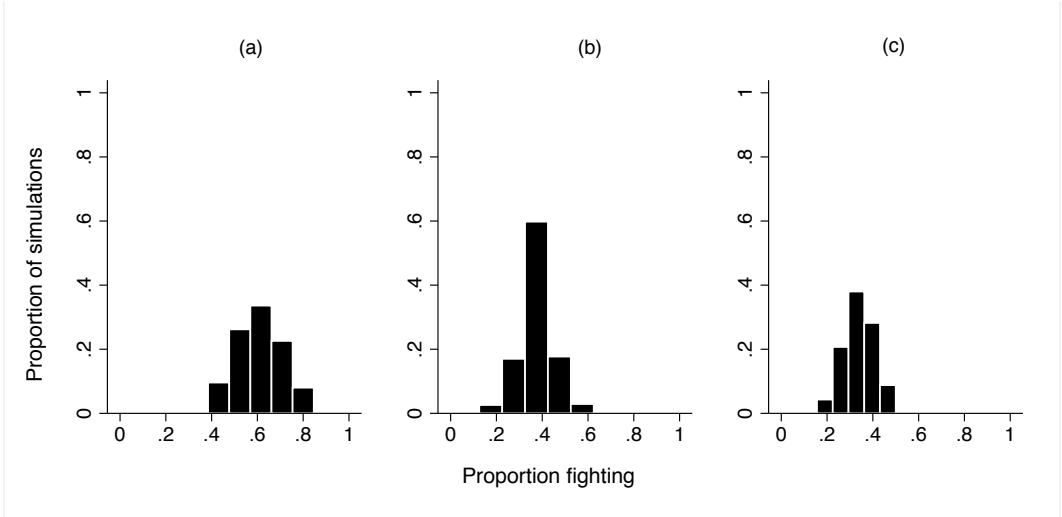
Outcome	Treatment			Total
	Control	Signalling	Countersignalling	
Squat-off	0.62	0.38	0.27	0.37
	(16)	(18)	(16)	(50)
Other	0.38	0.63	0.73	0.63
	(10)	(30)	(44)	(84)
Total	1.00	1.00	1.00	1.00
	(26)	(48)	(60)	(134)

[103] For the analyses presented in H7, I remove five subjects who were matched with the five that had to be initially excluded from the dataset because they would be responding to potentially anomalous decisions taken by the excluded five. I do this to ensure that there are no potential problems with their decisions. However, the results are not changed from their inclusion (r=-0.26, $d.f.$=137, p=0.002), ($\boldsymbol{X}^2$=9.72, p=0.008).

Although there is no difference between the proportion of fights in the countersignalling and signalling treatments ($\boldsymbol{X}^2$=1.45, p=0.228), and the hypothesis does not make predictions distinguishing the two, for transparency I keep these two treatments separate in the analysis.

The previous result is based on the outcomes that actually happened in the experiment. As there is an element of chance involved, in that subjects are randomly matched, it could be the case that the result is a statistical aberration arising not from sampling error but matching error. To ensure that the result is not down to matching error, I simulate the matching of subjects 1000 times. Each time, subjects in a treatment group are randomly paired with one another. This exercise demonstrates that the initial result is broadly stable (although the average proportion fighting in the countersignalling treatment increases from 26.7% to 33.7%). In the control group, 61% of the interactions end in a fight (mean=0.610, s.d.=0.087), 37.7% end in a fight in the signalling treatment (mean=0.377, s.d.=0.076), and 33.7% do so in the countersignalling treatment (mean=0.337, s.d.=0.061) (Figure 18).

Figure 18. The proportion of fights that occur after 1000 repetitions of the matching procedure in the (a) control, (b) signalling treatment, and (c) countersignalling treatment.



One interesting question to ask, and address here, is which player is driving the between-treatment differences in physical conflict? It could be that when there is more information residents react, or it could be that rookies are the ones to do so.

Residents in the control group choose to challenge in the vast majority of interactions (88%) and do so more than residents in the signalling treatment (65%) (exp(b)=0.25, z=-2.20, p=0.028) or the countersignalling treatment (68%) (exp(b)=0.27, z=-2.03, p=0.043) (Table D2). This behavioural difference is consistent with the beliefs that they report, for, while there is no difference across the groups regarding the probability of rookies resisting (63% in the control, 55% in the signalling group, and 61% in the countersignalling group, see Table D3), residents in the control group believe that they are likelier to win in a squat-off. In the control group they report that they would beat 62% in a squat-off, which is significantly higher than that reported in the signalling group (51%, p=0.048), and than in the countersignalling group (48%, p=0.017) (Table D3).

Whether rookies respond or not is ambiguous as p-values hover around the 10% level: 69% resist in the control group compared to 48% in the signalling treatment (exp(b)=0.41, z=-1.74, p=0.082) or 50% in the countersignalling treatment (exp(b)=0.44, z=-1.63, p=0.103) (Table D2). Even if there is some difference in behaviour, this is not down to their beliefs of winning in a squat-off for these do not differ across treatment groups, as expected by theory: 59% in the control, 51% in the signalling treatment, and 48% in the countersignalling treatment (Table D3).

The other interesting question to address here is what happens to the proportion of exploitations that occur in each treatment. While the proportion of interactions that

result in physical conflict might be lower in the presence of more information, this advantage may be offset by an increase in situations of exploitation.

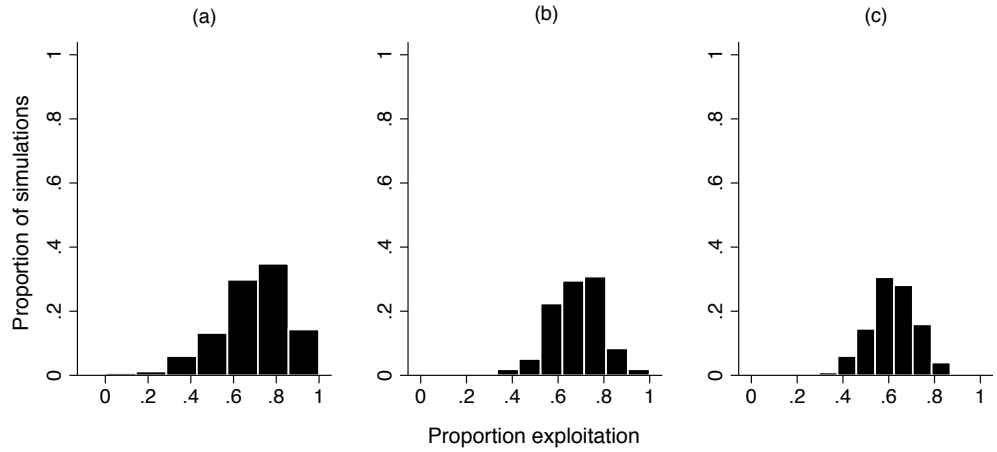
Removing the 50 subjects who ended up in conflict, I am left with 84 who ended up in a status quo outcome or an exploitation situation. The proportion who ended up in each, does not differ by treatment: in the control group 60% ended up in exploitation, 67% did so in the signalling treatment, and 64% of subjects in the countersignalling treatment (Fisher's exact test, p=0.901) (Table 17).

Table 17. Number and proportion of subjects ending up in exploitation or status quo.

Outcome	Treatment			Total
	Control	Signalling	Countersignalling	
Exploitation	0.60	0.67	0.64	0.64
	(6)	(20)	(28)	(54)
Status quo	0.40	0.33	0.36	0.36
	(4)	(10)	(16)	(30)
Total	1.00	1.00	1.00	1.00
	(10)	(30)	(44)	(84)

As before, this result may be driven by the stochastic matching procedure. To check for this possibility, I re-match subjects 1000 times and record the proportion of interactions, among those not in a fight, who end up in the exploitation situation. In the control group this is 69.1% (mean=0.691, s.d.=0.182), in the signalling treatment it is 67.9% (mean=0.679, s.d.=0.109), and in the countersignalling treatment it is 61.9% (mean=0.617, s.d.=0.100) (Figure 19). These show that including information in the conflict situation does not increase exploitation.

Figure 19. The proportion of exploitations that occur after 1000 repetitions of the matching procedure in the (a) control, (b) signalling treatment, and (c) countersignalling treatment.



4.5 Discussion

Squatting is a very reliable sign of resisting: almost all subjects who obtain an A grade for their veiled squat resist, and in case of a fight, longer veiled squatters are much likelier to win. Residents recognise the informational value of squatting and respond appropriately in that fewer challenge rookies who emit higher signs. Like in prisons, signs of toughness, scars and tattoos, are emitted and interpreted as indicators of violence prowess. It is a strong sign and people pay very close heed to it.

Squatting is a reliable signal of toughness too, as more rookies who obtain higher grades decide to resist, and like the veiled squat, rookies with a higher unveiled grade win in most of the fights that occur. Residents recognise this and respond as expected.

These findings demonstrate that squatting can function as a reliable source of information, and that it is a separating, or semi-separating signal such that low types cannot afford to squat for long, while high types can. More importantly, it implies that

people have the capacity to use and respond to signs and signals of toughness, making it likelier that people in real life observe and take notice of signs and signals of toughness.

As large inclusive fitness gains should be had from avoiding potentially deadly fighting, signs and signals of toughness may be some of the most important manifesta that humans have evolved to consider. If this is the case, women should consider men with signs and signals of toughness as more attractive partners than men lacking them for having a partner who avoids violence could increase fitness. Indeed, one experiment finds that men with scars are considered as more attractive partners for short-term relationships than men without scarring (Burriss, Rowland, & Little, 2009).

Despite the benefits of signalling, I do not find that people in the signalling group squat for longer upon unveiling than people in the control group. This is surprising for in the latter there is no (external) incentive for squatting. My working hypothesis, which is supported by exploratory analyses that I carried out, is that the internal incentives that people have in the control group are driven out by the external one in the signalling group such that the overall motivation level remains comparable. The motivation is different but the overall behavioural effect is the same.

The motivation could be partly driven by the inherent challenge and competition in the squatting procedure, and partly down to the lack of experience that subjects had with squatting. Were I to re-run this experiment with subjects who have experience with squatting as a punishment, I expect that subjects would not have the same internal motivation and squat for much less in the control group. Additionally, were the squat procedure to be repeated multiple times such that signalling takes place many times—like signalling does in real life—I expected that people would soon stop squatting in the

control, while those in the signalling treatment would not stop doing so, or only do so at a slower rate. Experience of the exercise and interpreting it as net cost should lead to the predicted difference. I do find that a substantial number of participants in the signalling group recognise and report strategic motivations for their behaviour.

Opposing the prediction derived from countersignalling, there is a strong linear relationship between signalling and type in the countersignalling treatment. Like in the signalling treatment, high types squatted for the longest in the unveiled stage of the countersignalling treatment. Moreover, I find no difference in the unveiled squat time of subjects in the countersignalling and signalling treatment. The sign, or extra information, had no impact on the signalling strategy of senders.

Why the lack of countersignalling? One possibility is that the cost of the signal is too high: medium types simply cannot pass themselves off as high types. It is less likely that countersignalling can emerge in such a case, for if the cost of the extra signal is small enough for high types, it may be better for them to use it. Countersignalling may be likelier to emerge when the signal is affordable to both medium and high types.

Whatever the explanation, senders clearly behave correctly given the behaviour of receivers. Irrespective of the sign initially emitted, a higher signal always garners more respect from residents. Receiving an A sign, as opposed a C sign, leads to between 16-24% more residents ignoring at all signal levels, and each additional grade increase in the signal obtains about 19% more residents ignoring. For instance, a 'countersignalling' high type, who emits aC is ignored only 22% of the time compared to a signalling one who emits aA and is ignored 59% of the time: attempting to countersignal backfires here. Each additional investment into signalling has about the same effect for both A and C

signs, and the sign increases or decreases the baseline such that there is no interaction effect.

Consistent with my findings in Chapter 3 and Gambetta's theorising and empirical evidence, I find that a greater availability of information leads to fewer occurrences of physical conflict. This is the first experiment that has demonstrated this effect. The magnitude of the effect is striking: fights are reduced from 62%, when there is no information, to 27%—a reduction of 56.5% [(0.62-0.27)/0.62]*100—when there are two sources of information. If this result were to hold outside of the lab,[104] in natural settings, it could have important policy implications for prison management. Simply allowing information to flow more freely in prisons could substantially reduce violence.[105] Indeed, this is a highly plausible explanation for why there is less violence in lower security category prisons—a stable result found in the criminological literature and in my previous chapter.

These findings may well be an underestimate too of the actual effect that information has in reducing conflict. In this experiment, only one party had information about the other—only the residents knew about rookies. Often however, at least some information is available about one another to all of the potential combatants allowing them all to condition their beliefs and reach 'agreement' without violence in more cases. Here, only the resident could respond to the rookie's information.

[104] My use of a non-cooperative game in which one of the outcomes is a physical competition involving both pain and money, makes it more likely that this result holds in the real world. The typical external validity shortcoming of laboratory experiments, with this methodology, is to some degree, assuaged.

[105] Whatever management's reasons for implementing such a procedure, the Pearsall detention centre in Texas, enforces colour-coded uniforms on prisoners that correspond to their criminal offence: blue for minor offences, orange for intermediate, and red for the most serious offenders (The Economist, 2014b).

One may correctly wonder about the negative consequences of increasing information and hypothesise that information leads to increased exploitation. Contrary to this intuition, conditional on the people not fighting, there is no difference in the proportion exploited in the three treatment groups. Although, in absolute terms, more subjects are exploited when there is more information, this only occurs because more people do not fight.

To illustrate the effect of information, consider three prisons, each corresponding to one of the levels of information availability and housing 100 prisoners. I use the average of the non-fighting prisoners who end up in a situation of exploitation, which is about 64% (Table 17). In the prison with no information circulation, say a high security facility, 62 prisoners fight, 24 prisoners exploit others or are exploited, and 14 do not interact. In the signalling prison, 38 prisoners fight, 40 are in exploitation, and 22 do not interact. Finally, in the countersignalling prison, a low security prison, 27 fight one another, 47 are in exploitation, and 26 are in status quo. In terms of physical safety, the third prison is clearly the best.

Congruent with the literature that more competitive and aggressive men have lower 2D:4D ratios, I find that a low ratio is associated with a longer veiled squat; the strength of the effect (r=-0.21) is a little stronger than that typically found (Hönekopp & Watson, 2011, p. 383). This matches with the theory, and available evidence, and it strengthens the validity of the squat as a measure for toughness. Moreover, it suggests that squatting correlates with behaviours outside of the laboratory. Yet, this finding should be taken with a caveat, for I do not find links between 2D:4D and *a priori* equally plausible correlates: it does not correlate with unveiled squat time, resisting, challenging, fight, or any variables of the BPAQ. In only 1/10 comparisons is 2D:4D significant. This poor

'hit rate' is a concern for the usefulness of 2D:4D for understanding human behaviour—and if it is a simple descriptive correlate (as opposed to a causally related one) it is not particularly interesting.[106]

Surprisingly, the Buss-Perry Aggression Questionnaire is not associated with any of my variables. This is surprising given the often-reported correlates of it. A reason for this result could be the questionnaire technique: questionnaires responses often do not translate, in a straightforward way at least, into behaviours. Social norms, for one thing, intervene, both in the self-report in the questionnaire, and in the actual behaviour.

Squatting ability is a correlate of pain resistance and fitness, based on subjects' questionnaire responses, and can be used as a sign and a signal of toughness. Consistent with signalling theory, higher type senders invest more in signalling, however, contrary to expectations, senders in the signalling treatment did not invest more than those in the control. Receivers base their decisions on both signs and signals of toughness and their evaluations follow the pattern predicted by signalling, but not countersignalling theory: senders with more costly signals are treated with more respect irrespective of their sign. There is a remarkably close link between the decisions of rookies to resist and the actions of residents to challenge. Finally, more information about toughness decreases the proportion of fights that occur without increasing in the proportion of exploitations, which can have important policy implications.

[106] Some studies conclude that 2D:4D is a measure that is reasonably sensitive to the experimenter and measurement methods, leading to the variation in findings (Kemper & Schwerdtfeger, 2009; Voracek, Manning, & Dressler, 2007). It is plausible that the procedure used is responsible for the few correlations.

5. Conclusions

The unifying theme of this thesis is its concern with the effect that the presence or absence of information has on peoples' signalling and fighting behaviour. Parts of it examine how senders act when signs about relevant qualities of theirs are naturally present, and conversely, how receivers interpret those sign-signal combinations and consequently behave towards the senders. Assuming there are two sources of information, unintentional and intentional created, how do senders choose to change the content of their intentionally generated information? Other parts speak to the question of how senders and receivers behave when the availability of the information is restricted or expanded. Is there more violence when few signs and signals about fighting prowess are available? Given that both natural and strategic sources of information are near-ever present in social situations, examining them simultaneously is a valuable endeavour that can help us to understand behaviour in a range of interactions.

5.1 How prior information affects signalling

Countersignalling theory provides a prediction, and explanation, for the behaviour of senders and receivers when a particular kind of sign is also present. Senders who possess a medium level of a relevant quality, not directly observable by receivers, should invest the most resources in signalling, and senders with a lower or higher level should invest less. Receivers recognise medium types by their signal, while low and high types pool in terms of their actions but are separable by their signs.

There are a number of good reasons why I employ this theory in my thesis: no other formal models include both a signal and a realistically modelled noisy and exogenous

sign, it yields an intuitively appealing result with broad generality, and its prediction corresponds to much anecdotal evidence. It is also worthwhile to investigate because a growing theoretical literature is developing that cites or builds on countersignalling.

The majority of the results found in Chapter 2 are consistent with predictions derived from countersignalling theory. A greater proportion of medium generosity senders choose to participate in a signalling situation than high and low generosity senders, most medium types choose to revise in order to signal, and receivers behave as expected. Yet, an important prediction was not fully realised: a greater proportion of intermediate type trustees chose to signal than high type trustees—as expected—but a similar proportion of low types chose to signal as intermediate types—contrary to the prediction. In this sense, medium types did not shout the loudest; low types shout just as loud.

Focusing on the high types, I find that they behave consistently with a weak interpretation of countersignalling theory; most high type senders avoid a signalling opportunity entirely—like students who do not want to enrol in a basic course—allowing their signs to come to the foreground, and receivers treat them as trustworthy types despite their refusal to signal. However, they are not consistent with a stronger interpretation in which the signalling opportunity is used and actively downplayed: dressing up in cheap clothes, giving inexpensive gifts, or answering obvious questions flippantly. Indeed, trustees who actively downplayed their generosity and decreased upon revision were evaluated poorly (see Section 2.5.2).

Perhaps one cause for this is the particular quality, trustworthiness, and corresponding signal, generosity, used in the experiment. For giving money increases the utility of senders who are more other-regarding, and other-regarding senders are likelier to be

trustworthy. Withholding money, downplaying the signal, is then strange, as the truly generous must trade-off the utility that they gain from transferring more money to others and the utility that they could obtain from receivers by transferring less money and countersignalling. By giving less, they may make themselves more recognisable yet at the same time directly decrease their benefits. Consider a generous sender who maximises his utility function—in absence of signalling motivations—by giving £4 to another, as measured by a dictator game. However, in a countersignalling type situation, giving £0, and downplaying his signal, is the best action to take to maximise the gain he obtains from a receiver. The two sources of benefits are in opposition with one another—transfer a lot and gain by considering other's payoffs or transfer little and gain from signalling—and the sender has to balance between these cost-benefits. This makes it harder for countersignalling to emerge because many senders may simply prefer to give than to separate themselves as high-type senders.

In Chapter 4, I once again put countersignalling theory to the test.[107] Here, the quality is senders' willingness to engage in 'fighting' and the signal is squatting. As the signal, squatting time, does not directly benefit others, the complication that occurred when using trustworthiness and generosity was avoided. Senders do not have to worry about their signal negatively affecting others.

Under these conditions, neither interpretation of countersignalling receives support, and senders do not behave according to the theory. In the countersignalling treatment, there

[107] My hope was to also test signalling and countersignalling theory in Chapter 3 using data on real prisoner behaviour. Despite my aim and multiple draft chapters, I was ultimately prevented from providing a convincing test of countersignalling because of limitations inherent in the data. The fundamental limitation is that the data do not record a reasonably comprehensive set of signalling actions, only a relatively small number of potential signalling actions. So, even if the data show that high violence potential inmates take certain actions less than medium violence potential inmates, it is possible that the high type inmates are simply signalling with other actions.

is a strong positive linear relationship between type and signalling (b=0.574, t=6.40, p<0.001), so that the most willing to resist are the ones who signal the most. High types are not blasé with signalling, and 'shout' louder than the medium types. Moreover, a comparison between the squatting times of high types in the signalling treatment and the high types in the countersignalling treatment reveals that there is no difference in the average investment between the two. Were countersignalling theory being followed, the toughest senders in the countersignalling treatment should signal less than the toughest senders in the signalling treatment.

Receivers too did not evaluate senders according to countersignalling theory and clearly did not value senders' attempts to countersignal. Countersignalling senders, who are perceived as emitting the combination aC, are evaluated equivalently to senders who are perceived as emitting the combination cB—a combination that is a full grade lower than aC. Additionally, aC senders are evaluated much more poorly than senders who display aA, both of whom should be treated equivalently under countersignalling (Figure 17). Downplayed squatting is not appreciated.

Of course, it is possible to posit *post hoc* explanations for why countersignalling is not found in this experiment. Perhaps squatting is a very differentially costly signal that has carry-over effects from the veiled squat to the unveiled one. Having already done the veiled squat, medium types find it too difficult to send a signal to separate themselves from low types. If this is the case, countersignalling high types would actually be pooling with both low and medium types, and so, have to revert to signalling to separate themselves. There is some evidence to support this, for consistent with this conjecture, the average veiled squatting time is less than the average unveiled squatting time (p. 149).

In combination with the findings from the initial experiment on countersignalling using generosity and trustworthiness, in which countersignalling receives limited support, the results from the latter experiment using squatting and fighting, in which countersignalling receives no support, I conclude that caution should be used when employing countersignalling as an explanation for human social behaviour. As is often the case with game theoretic models, countersignalling behaviour may be more easily observed in market interactions and be more likely to be taken by firms—as opposed to individuals. Still, my results suggest that caution should be applied when using countersignalling theory as an explanation and further experiments ought to be pursued.

A small and growing theoretical literature has built up around countersignalling, and it is described in at least one game theory textbook (Rasmusen, 2007, pp. 345–348). 27 papers that cite Feltovich et al's paper (2002), as indicated by Web of Science, are theoretical ones that mention countersignalling and develop ideas in a similar area or directly build upon it. A substantive contribution of my thesis is that I have examined countersignalling theory and found mixed, or limited, support for it. Moving beyond anecdotes and stylised facts, and considering countersignalling seriously, has been my commitment throughout this thesis. Demonstrating its empirical limitations is a positive contribution that has arisen.

Another way to approach the question of how the presence of signs affects signalling is to employ data-driven retrodiction. What conjectures, or alternative theories, fit the data collected in the toughness experiment (for which countersignalling receives no support)?

By holding the signal constant, residents' responses to signs are clear. Consider a rookie with an 'A' sign and an 'A' signal and a rookie with a 'C' sign but the same 'A' signal. The

former is considered as significantly, and substantially tougher—59% residents avoid this rookie—than the latter—whom only 38% of residents ignore. The same relationship holds for all other grade combinations: senders who have the same signal but different signs, are treated differently such that senders possessing higher signs receive more respect (Table 15).

Turning the matter around, and holding signs constant, it is evident that they also matter. Residents avoid rookies who have an 'A' sign and an 'A' signal much more than they do rookies with an 'A' sign and a 'C' signal (59% ignore the former while 22% avoid the latter). This relationship holds for *every* veiled-unveiled grade combination, holding sign constant. Senders who emit a higher signal are treated with more deference than senders who emit a lower signal but the same sign (Table 15).

Residents seem to weigh the two sources correctly too, giving less weight to the veiled squat grade—which is highly probabilistic—than they give to the unveiled grade—which is a highly separating signal that has much less noise. Consider the grade pairings of cA and aC: rookies with a 'C' veiled grade and an 'A' unveiled grade are given significantly and substantially more respect than rookies who emit the opposite combination of an 'A' veiled grade an a 'C' unveiled grade (38% for the former and only 22% for the latter).

Perhaps the best way of understanding their behaviour is from the perspective of semi-sorting equilibria in signalling theory. Since toughness is not a perfect correlate of fighting ability, neither as a sign nor as a signal, some proportion of low type rookies will naturally emit a sign indicating that they are tough—even the meek occasionally pick up scars from schoolyard fights, and some proportion can afford to signal that they are tough. Very few, however, will have both a sign that indicates their toughness, and a

signal also indicating the same. Realising this, receivers look for coherence, and correspondence, between the signs and signals of senders to confirm the underlying quality: matching natural and strategic toughness provides reassurance. Like Gambetta and Hamill (2005) find, taxi drivers use multiple sources of information about their potential clients to judge whether he or she is a robber or a genuine fee-paying passenger, and in particular, consider the coherence or incoherence of the information (pp. 80, 88, 159). Consistent with my experimental results, they take coherence as a reassuring indication, while incoherence, '*not being drunk on a Saturday night*' (italics in original, p. 83) or 'a well-dressed man in a bad area' (p. 159), as a red flag.

Realising that their signs and signals are not foolproof, senders who can afford to do so signal: there is a strong positive linear relationship between type and unveiled squatting time (Figure 15). Senders' behaviour matches the treatment that they obtain from receivers.

On the path to testing countersignalling theory and combining signs and signals, I have developed and refined some specific experimental tools that others can use to further test countersignalling, or to test how multiple and different sources of information interact. The two experiments demonstrate how generosity, which has been used a small number of times before as a signal of trustworthiness (Fehrler & Przepiorka, 2013; Gambetta & Przepiorka, 2014), can be combined with a sign of generosity, and Chapter 4 is the first experiment to use real toughness, specifically squatting, as a signal of one's willingness to use force. The three-step procedure, explicitly described in Section 4.3, is a tool that has potentially broad application.

5.2 How information affects fighting

Chapter 3 and Chapter 4 examine the effect that information has on the likelihood of violence emerging. Signs and signals of a resource holder's willingness to resist a physical challenge should, and does, decrease fighting. This is because the conflict is less likely to be resolved through non-violent informational means when there is no, or little, possibility for communication to occur.

A range of variables indicates that more information leads to less fighting in prison (Chapter 3). Prisoners who can be expected to have more raw materials of their violence ability—more scars to displays, greater knowledge of prison argot from prior incarcerations—and prisoners who inhabit environments in which information spreads more easily—such as a lower security prison, or working inside a facility—fight less.

Almost all of the variables that are significant, except one, and are used to indicate violence capital go in the expected direction. All of the variables that are used to test the information-spread hypothesis, when significant, display the predicted relationship. The results are found *in situ*, from interviews with a nationally representative sample of incarcerated prisoners, and hold when controlling for multiple confounders.

These observational results are congruent with the results obtained experimentally in Chapter 4 and reported in Section 4.4. By manipulating the amount of information that is sent from one subject to another, I test how information availability about toughness affects fighting. Rookie-subjects who have the possibility of sending signs and signals of their toughness, and resident-subjects who can perceive such information, are involved in 57% fewer squat-offs than those who do not have the chance to send or receive relevant information. Simulation results confirm the robustness of the finding.

Rookies in all of the conditions receive the same amount of feedback about their own squatting abilities: they are all shown the veiled and unveiled grade that they obtained, so, the increase in fighting that occurs with less information cannot be because rookies know less about their own abilities—a possibility with the observational data. Thus, observational and experimental results support Gambetta's information hypothesis.

An important related finding emerges from the toughness experiment: more information does not lead to a greater proportion of people being exploited. This means that more information reduces violence without causing problems in another area.

Together, these findings from the SISFCF and the experiment may have potentially important policy implications. Most obviously, increasing information about one another's violence potential in prison may lead to prisons that are more peaceful. The general principle may apply too to other situations in which there is a potential for conflict over resources and violence is costly for those involved. As with any policy change, extreme care should be taken, and further experiments and observational analyses carried out.

A core result of this thesis is that information decreases violence, and it is a very promising avenue for further research.

5.3 Summary of findings

Explicitly listed below are the answers to each of the questions posed in the introduction.

1. Can signs created in experiments reliably indicate the qualities of senders and do receivers respond to them?

Generosity is a reliable sign of senders' trustworthiness. Subjects who naturally give more money to others are much more trustworthy: those giving nothing return only 11% of the time while those who give half return 75% of the time (p. 65). Trusters respond appropriately and behave more trustingly towards senders who give more naturally. More of them send money to senders with a sign indicating greater generosity. A majority, 60%, send money to trustees who appear to have given half of their money to another. In contrast, a clear minority, 29%, send money to trustees who give no money away (p. 65).

Toughness, measured by squatting, is a reliable sign of resisting, and winning, in case a squat-off competition occurs. A meagre 27% of the subjects who obtain the lowest grade for veiled squatting decide to resist, while nearly all, 84%, who obtain the highest grade resist. Having a higher veiled squatting time is strongly associated with winning in a squat-off (p. 151). Wisely, receivers are more likely to leave those subjects alone who have higher signs, although they take into consideration the probabilistic nature of the sign: 23% avoid senders with the worst sign and 41% avoid senders with the best sign (p. 151).

2. Can signals created in experiments reliably indicate the qualities of senders and do receivers respond to them?

Toughness is a reliable signal of resisting and winning in squat-offs. 71% of subjects in the signalling treatment and 100% of subjects in the countersignalling treatment who send the highest signal choose to resist. This contrasts sharply with the proportion who

decide to resist after sending the lowest possible signal: 41% in the signalling and 26% in the countersignalling treatment (pp. 153, 154).

Receivers respond to the signal by treating senders with higher signals with more care, and a greater proportion avoids senders with higher signals. In the signal only group, 10% of the receivers avoided a rookie with a C unveiled grade and 61% avoided a rookie with an A grade (p. 153). There is also a strong positive linear relationship between the signal and the proportion of residents who choose to ignore in the countersignalling treatment—although the specific proportion depends on the sign that the unveiled grade is combined with (p. 161).

3. How do senders behave when they can emit both signs and signals? Do the middle shout loudest?

In the generosity experiment, most senders decided to avoid sending a signal completely: they bypass the opportunity to send a precise signal of their trustworthiness. The generous here seem unconcerned with revision. Among the subjects who send a signal, a greater proportion of those with an intermediate sign increased their giving upon revision than those with a high sign (p. 69). Medium types signal more than high types but at similar levels to low types.

When senders have no possibility of avoiding sending a signal, as in the toughness experiment, there is a strong positive linear relationship between type and signal investment. Tougher subjects send a higher signal than those who are less tough (p. 159). Regarding toughness, the high shout the loudest.

4. How do receivers behave when they perceive both signs and signals?

Receivers consider signs as an additional source of information that reinforces or undermines the signal. They do so in different ways for generosity and toughness.

Perceiving that a trustee has a miserly sign makes receivers very reluctant to bestow trust in him or her, irrespective of the amount that the trustee gives after unveiling. Observing that a trustee has an intermediate sign leaves open the possibility that the trustee is trustworthy and trusters are sensitive to changes in signalling amounts for intermediate trustees. Trusters do not bestow additional trust on trustees who display a generous sign and subsequently signal. Signalling has no effect for high type trustees. Moreover, receivers treat trustees who initially display their generosity poorly if they go on to decrease their investment relative to their natural level upon revision (p. 71).

Resident-receivers respond astutely to the information emitted by rookie-senders. Holding their signs constant, rookies with higher signals are treated with more respect, and conversely, holding their signals constant, rookies with higher signs are treated with more care. Moreover, residents consider the noise of the information and weight their decisions according (p. 161).

5. Does information availability affect fighting?

Information decreases the amount of fighting that occurs when people compete for scarce resources. I find support for this effect from observational data *in situ* among prisoners (p. 113), and experimentally in a laboratory setting (p. 162).

Bibliography

Aldrich, N. W. (1988). *Old money: The mythology of America's upper class*. New York: Knopf.

Ambady, N., & Rosenthal, R. (1992). Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological Bulletin*, *111*(2), 256–274. http://doi.org/10.1037/0033-2909.111.2.256

Ambady, N., & Rosenthal, R. (1993). Half a minute: Predicting teacher evaluations from thin slices of nonverbal behavior and physical attractiveness. *Journal of Personality and Social Psychology*, *64*(3), 431–441. http://doi.org/10.1037/0022-3514.64.3.431

Archer, J. (2006). Testosterone and human aggression: An evaluation of the challenge hypothesis. *Neuroscience and Biobehavioral Reviews*, *30*(3), 319–345. http://doi.org/10.1016/j.neubiorev.2004.12.007

Archer, J., & Southall, N. (2009). Does cost–benefit analysis or self-control predict involvement in bullying behavior by male prisoners? *Aggressive Behavior*, *35*(1), 31–40. http://doi.org/10.1002/ab.20283

Archer, J., & Webb, I. A. (2006). The relation between scores on the Buss–Perry Aggression Questionnaire and aggressive acts, impulsiveness, competitiveness, dominance, and sexual jealousy. *Aggressive Behavior*, *32*(5), 464–473. http://doi.org/10.1002/ab.20146

Ashraf, N., Bohnet, I., & Piankov, N. (2006). Decomposing trust and trustworthiness. *Experimental Economics*, *9*(3), 193–208. http://doi.org/10.1007/s10683-006-9122-4

Asmussen, E., & Heebøll-Nielsen, K. (1962). Isometric muscle strength in relation to age in men and women. *Ergonomics*, *5*(1), 167–169. http://doi.org/10.1080/00140136208930570

Aumann, R. J. (1976). Agreeing to disagree. *The Annals of Statistics*, *4*(6), 1236–1239. http://doi.org/10.1214/aos/1176343654

Austin, J., Johnson, K. D., & Gregoriou, M. (2000). *Juveniles in adult prisons and jails: A national assessment*. Bureau of Justice Assistance.

Bacharach, M., & Gambetta, D. (2001). Trust in signs. In K. S. Cook (Ed.), *Trust in society* (pp. 148–184). New York: Russell Sage Foundation.

Bailey, A. A., & Hurd, P. L. (2005). Finger length ratio (2d:4d) correlates with physical aggression in men but not in women. *Biological Psychology*, *68*(3), 215–222. http://doi.org/10.1016/j.biopsycho.2004.05.001

Bastian, B., Jetten, J., & Ferris, L. J. (2014). Pain as social glue: Shared pain increases cooperation. *Psychological Science*, *25*(11), 2079–2085. http://doi.org/10.1177/0956797614545886

Bell, E. T. (1931). Mathematics and speculation. *The Scientific Monthly*, *32*(3), 193–209.

Belot, M., Duch, R., & Miller, L. (2010). *Who should be called to the lab? A comprehensive comparison of students and non-students in classic experimental games* (Discussion Paper No. 2010001). University of Oxford, Nuffield College. Retrieved from http://ideas.repec.org/p/cex/dpaper/2010001.html#biblio

Bench, L. L., & Allen, T. D. (2003). Investigating the stigma of prison classification: An experimental design. *The Prison Journal*, *83*(4), 367–382. http://doi.org/10.1177/0032885503260143

Berger, J., & Ward, M. (2010). Subtle signals of inconspicuous consumption. *Journal of Consumer Research*, *37*(4), 555–569. http://doi.org/10.1086/655445

Berg, M. T., & DeLisi, M. (2006). The correctional melting pot: Race, ethnicity, citizenship, and prison violence. *Journal of Criminal Justice*, *34*(6), 631–642. http://doi.org/10.1016/j.jcrimjus.2006.09.016

Berk, R. A., Ladd, H., Graziano, H., & Baek, J.-H. (2003). A randomized experiment testing inmate classification systems*. *Criminology & Public Policy*, *2*(2), 215–242. http://doi.org/10.1111/j.1745-9133.2003.tb00121.x

Binmore, K., Shaked, A., & Sutton, J. (1985). Testing noncooperative bargaining theory: A preliminary study. *The American Economic Review*, *75*(5), 1178–1180.

Bottoms, A. (1999). Interpersonal violence and social order in prisons. *Crime and Justice*, *26*, 205–281.

Bourdieu, P. (1984). *Distinction: A social critique of the judgement of taste*. Cambridge, Massachusetts: Harvard University Press.

Bracht, J., & Feltovich, N. (2009). Whatever you say, your reputation precedes you: Observation and cheap talk in the trust game. *Journal of Public Economics*, *93*(9–10), 1036–1044. http://doi.org/10.1016/j.jpubeco.2009.06.004

Brown, C., Garwood, M. P., & Williamson, J. E. (2012). It pays to cheat: Tactical deception in a cephalopod social signalling system. *Biology Letters*, *8*(5), 729–732. http://doi.org/10.1098/rsbl.2012.0435

Burriss, R. P., Rowland, H. M., & Little, A. C. (2009). Facial scarring enhances men's attractiveness for short-term relationships. *Personality and Individual Differences*, *46*(2), 213–217. http://doi.org/10.1016/j.paid.2008.09.029

Buss, A. H., & Perry, M. (1992). The aggression questionnaire. *Journal of Personality and Social Psychology*, *63*(3), 452–459. http://doi.org/10.1037/0022-3514.63.3.452

Cabanac, M. (1986). Money versus pain: Experimental study of a conflict in humans. *Journal of the Experimental Analysis of Behavior*, *46*(1), 37–44. http://doi.org/10.1901/jeab.1986.46-37

Cameron, R. E., & Neal, L. (2003). *A concise economic history of the world: From Paleolithic times to the present*. Oxford: Oxford University Press.

Camp, S. D., & Gaes, G. G. (2005). Criminogenic effects of the prison environment on inmate behavior: Some experimental evidence. *Crime & Delinquency*, *51*(3), 425–442. http://doi.org/10.1177/0011128704271471

Camp, S. D., Gaes, G. G., Langan, N. P., & Saylor, W. G. (2003). The influence of prisons on inmate misconduct: A multilevel investigation. *Justice Quarterly*, *20*(3), 501–533. http://doi.org/10.1080/07418820300095601

Chernow, R. (2009, March 23). Madoff and his models. *The New Yorker*. Retrieved from http://www.newyorker.com/magazine/2009/03/23/madoff-and-his-models

Chiu, H.-H., & Chen, C.-M. (2014). Advertising, price and hotel service quality: A signalling perspective. *Tourism Economics*, *20*(5), 1013–1025. http://doi.org/10.5367/te.2013.0324

Cho, I.-K., & Kreps, D. M. (1987). Signaling games and stable equilibria. *The Quarterly Journal of Economics*, *102*(2), 179. http://doi.org/10.2307/1885060

Clear, T. R., & Sumter, M. T. (2002). Prisoners, prison, and religion. *Journal of Offender Rehabilitation*, *35*(3-4), 125–156. http://doi.org/10.1300/J076v35n03_07

Clemmer, D. (1940). *The prison community*. Boston: Christopher Publishing House.

Coates, J. M., Gurnell, M., & Rustichini, A. (2009). Second-to-fourth digit ratio predicts success among high-frequency financial traders. *Proceedings of the National Academy of Sciences*, *106*(2), 623–628. http://doi.org/10.1073/pnas.0810907106

Colson, E. (1953). *The Makah indians: A study of an indian tribe in modern society*. Manchester: Manchester University Press.

Connell, A., & Farrington, D. (1996). Bullying among incarcerated young offenders: Developing an interview schedule and some preliminary results. *Journal of Adolescence*, *19*(1), 75–93. http://doi.org/10.1006/jado.1996.0007

Cunningham, M. D., Reidy, T. J., & Sorensen, J. R. (2005). Is death row obsolete? A decade of mainstreaming death-sentenced inmates in Missouri. *Behavioral Sciences & the Law*, *23*(3), 307–320. http://doi.org/10.1002/bsl.608

Dhami, M. K., Ayton, P., & Loewenstein, G. (2007). Adaptation to imprisonment: Indigenous or imported? *Criminal Justice and Behavior*, *34*(8), 1085–1100. http://doi.org/10.1177/0093854807302002

Diamond, J. (2003). Archaeology: Propaganda of the pyramids. *Nature*, *424*(6951), 891–893. http://doi.org/10.1038/424891a

Diggle, S. P., Gardner, A., West, S. A., & Griffin, A. S. (2007). Evolutionary theory of bacterial quorum sensing: When is a signal not a signal? *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *362*(1483), 1241–1249. http://doi.org/10.1098/rstb.2007.2049

Dobbs, R. R., & Waid, C. A. (2004). Prison culture. In *Encyclopedia of prisons & correctional facilities* (SAGE Reference Online, pp. 720–724). Thousand Oaks: SAGE. Retrieved from http://www.sagepub.com/hanserintro/study/materials/reference/ref2.2.pdf

Duncan, J. S. (1973). Landscape taste as a symbol of group identity: A Westchester County village. *Geographical Review*, *63*(3), 334–355.

Earley, P. (1992). *The hot house: Life inside Leavenworth prison*. New York: Bantam.

Ekland-Olson, S. (1986). Crowding, social control, and prison violence: Evidence from the post-Ruiz years in Texas. *Law & Society Review*, *20*(3), 389–421. http://doi.org/10.2307/3053581

Ellingsen, T., Johannesson, M., Tjøtta, S., & Torsvik, G. (2010). Testing guilt aversion. *Games and Economic Behavior*, *68*(1), 95–107. http://doi.org/10.1016/j.geb.2009.04.021

Engel, C. (2011). Dictator games: A meta study. *Experimental Economics*, *14*(4), 583–610. http://doi.org/10.1007/s10683-011-9283-7

Ermisch, J., Gambetta, D., Laurie, H., Siedler, T., & Noah Uhrig, S. C. (2009). Measuring people's trust. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *172*(4), 749–769.

Faris, R., & Felmlee, D. (2011). Status struggles: Network centrality and gender segregation in same- and cross-gender aggression. *American Sociological Review*, *76*(1), 48–73. http://doi.org/10.1177/0003122410396196

Fehrler, S., & Przepiorka, W. (2013). Charitable giving as a signal of trustworthiness: Disentangling the signaling benefits of altruistic acts. *Evolution and Human Behavior*, *34*(2), 139–145. http://doi.org/10.1016/j.evolhumbehav.2012.11.005

Feltovich, N., Harbaugh, R., & To, T. (2002). Too cool for school? Signalling and countersignalling. *RAND Journal of Economics*, *33*(4), 630–649.

Fischbacher, U. (2007). z-Tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics*, *10*(2), 171–178. http://doi.org/10.1007/s10683-006-9159-4

Fombrun, C. J. (1996). *Reputation: Realizing value from the corporate image*. Boston: Harvard Business School Press.

Fuller, D. A., & Orsagh, T. (1977). Violence and victimization within a state prison system. *Criminal Justice Review*, *2*(2), 35–55. http://doi.org/10.1177/073401687700200205

Gaes, G. G., & McGuire, W. J. (1985). Prison violence: The contribution of crowding versus other determinants of prison assault rates. *Journal of Research in Crime and Delinquency*, *22*(1), 41–65. http://doi.org/10.1177/0022427885022001003

Gaes, G. G., Wallace, S., Gilman, E., Klein-Saffran, J., & Suppa, S. (2002). The influence of prison gang affiliation on violence and other prison misconduct. *The Prison Journal*, *82*(3), 359–385. http://doi.org/10.1177/003288550208200304

Gambetta, D. (2005). Deceptive mimicry in humans. In S. L. Hurley & N. Chater (Eds.), *Perspectives on imitation: From neuroscience to social science*. Cambridge: MIT Press.

Gambetta, D. (2009a). *Codes of the underworld: How criminals communicate*. Princeton: Princeton University Press.

Gambetta, D. (2009b). Signaling. In P. Hedström & P. Bearman (Eds.), *The Oxford Handbook of Analytical Sociology* (pp. 168–194). Oxford: Oxford University Press.

Gambetta, D., & Hamill, H. (2005). *Streetwise: How taxi drivers establish customers' trustworthiness*. New York: Russell Sage.

Gambetta, D., & Przepiorka, W. (2014). Natural and strategic generosity as signals of trustworthiness. *PLoS ONE*, *9*(5), e97533.

Gambetta, D., & Székely, Á. (2014). Signs and (counter)signals of trustworthiness. *Journal of Economic Behavior & Organization*, *106*, 281–297. http://doi.org/10.1016/j.jebo.2014.07.009

Gluckman, M. (1963). Papers in honor of Melville J. Herskovits: Gossip and scandal. *Current Anthropology*, *4*(3), 307–316.

Goffman, E. (1956). *The presentation of self in everyday life*. Edinburgh: University of Edinburgh.

Goffman, E. (1971). *Relations in public: Microstudies of the public order*. New York: Basic Books.

Gover, A. R., Pérez, D. M., & Jennings, W. G. (2008). Gender differences in factors contributing to institutional misconduct. *The Prison Journal*, *88*(3), 378–403. http://doi.org/10.1177/0032885508322453

Grafen, A. (1990). Biological signals as handicaps. *Journal of Theoretical Biology*, *144*(4), 517–546. http://doi.org/10.1016/S0022-5193(05)80088-8

Grann, D. (2004, February 16). The brand: How the Aryan Brotherhood became the most murderous prison gang in America. *The New Yorker*, pp. 156–171.

Grayson, B., & Stein, M. I. (1981). Attracting assault: Victims' nonverbal cues. *Journal of Communication*, *31*(1), 68–75. http://doi.org/10.1111/j.1460-2466.1981.tb01206.x

Griffin, M. L., & Hepburn, J. R. (2006). The effect of gang affiliation on violent misconduct among inmates during the early years of confinement. *Criminal Justice and Behavior*, *33*(4), 419–466. http://doi.org/10.1177/0093854806288038

Haddad Junior, V., Cardoso, J. L. C., & Moraes, R. H. P. (2005). Description of an injury in a human caused by a false tocandira (dinoponera gigantea, perty, 1833) with a revision on folkloric, pharmacological and clinical aspects of the giant ants of the genera paraponera and dinoponera (sub-family ponerinae). *Revista Do Instituto de Medicina Tropical de São Paulo*, *47*(4), 235–238. http://doi.org/10.1590/S0036-46652005000400012

Han, Y. J., Nunes, J. C., & Drèze, X. (2010). Signaling status with luxury goods: The role of brand prominence. *Journal of Marketing*, *74*(4), 15–30. http://doi.org/10.1509/jmkg.74.4.15

Harer, M. D., & Langan, N. P. (2001). Gender differences in predictors of prison violence: Assessing the predictive validity of a risk classification system. *Crime & Delinquency*, *47*(4), 513–536. http://doi.org/10.1177/0011128701047004002

Harer, M. D., & Steffensmeier, D. J. (1996). Race and prison violence*. *Criminology*, *34*(3), 323–355. http://doi.org/10.1111/j.1745-9125.1996.tb01210.x

Harman, S. M., Metter, E. J., Tobin, J. D., Pearson, J., & Blackman, M. R. (2001). Longitudinal effects of aging on serum total and free testosterone levels in

healthy men. *The Journal of Clinical Endocrinology & Metabolism*, *86*(2), 724–731. http://doi.org/10.1210/jcem.86.2.7219

Harrison, F., & El Mouden, C. (2011). Exploring the effects of working for endowments on behaviour in standard economic games. *PLoS ONE*, *6*(11), e27623. http://doi.org/10.1371/journal.pone.0027623

Harrison, F., Sciberras, J., & James, R. (2011). Strength of social tie predicts cooperative investment in a human social network. *PLoS ONE*, *6*. http://doi.org/10.1371/journal.pone.0018338

Hawkins, A. (2007, November 14). In pictures: The frugal billionaires. *Forbes*. Retrieved from http://www.forbes.com/2007/11/14/billionaires-walton-buffett-biz-cz_ah_1114frugalbillies_slide.html

Hedström, P., & Udehn, L. (2009). Analytical sociology and theories of the middle range. In P. Hedström & P. Bearman (Eds.), *The oxford handbook of analytical sociology* (pp. 25–47). Oxford: Oxford University Press.

Higginbotham, A. (2012, August 2). The narco tunnels of Nogales. *Businessweek: Global Economics*. Retrieved from http://www.businessweek.com/articles/2012-08-02/the-narco-tunnels-of-nogales

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81*(396), 945–960. http://doi.org/10.1080/01621459.1986.10478354

Hönekopp, J., & Watson, S. (2011). Meta-analysis of the relationship between digit-ratio 2d:4d and aggression. *Personality and Individual Differences*, *51*(4), 381–386. http://doi.org/10.1016/j.paid.2010.05.003

Hourani, A. H. (2005). *A history of the Arab peoples*. London: Faber.

Huebner, B. M. (2003). Administrative determinants of inmate violence: A multilevel analysis. *Journal of Criminal Justice*, *31*(2), 107–117. http://doi.org/10.1016/S0047-2352(02)00218-0

Huggins, D. W., Capeheart, L., & Newman, E. (2006). Deviants or scapegoats: An examination of pseudofamily groups and dyads in two Texas prisons. *The Prison Journal*, *86*(1), 114–139. http://doi.org/10.1177/0032885505284702

Hurd, P. L., & Enquist, M. (2005). A strategic taxonomy of biological communication. *Animal Behaviour*, *70*(5), 1155–1170. http://doi.org/10.1016/j.anbehav.2005.02.014

Irwin, J., & Cressey, D. R. (1962). Thieves, convicts and the inmate culture. *Social Problems*, *10*(2), 142–155. http://doi.org/10.2307/799047

James, E. (2003). *A life inside: A prisoner's notebook*. London: Atlantic Books.

Jiang, S., & Fisher-Giorlando, M. (2002). Inmate misconduct: A test of the deprivation, importation, and situational models. *The Prison Journal*, *82*(3), 335–358. http://doi.org/10.1177/003288550208200303

Jiang, S., & Winfree, L. T. (2006). Social support, gender, and inmate adjustment to prison life: Insights from a national sample. *The Prison Journal*, *86*(1), 32–55. http://doi.org/10.1177/0032885505283876

Kaminski, M. (2003). Games prisoners play: Allocation of social roles in a total institution. *Rationality and Society*, *15*(2), 188–217. http://doi.org/10.1177/1043463103015002002

Kaminski, M. (2004). *Games prisoners play: The tragicomic worlds of Polish prison*. Princeton: Princeton University Press.

Kemper, C. J., & Schwerdtfeger, A. (2009). Comparing indirect methods of digit ratio (2d:4d) measurement. *American Journal of Human Biology: The Official Journal of the Human Biology Council*, *21*(2), 188–191. http://doi.org/10.1002/ajhb.20843

Kerley, K. R., Matthews, T. L., & Blanchard, T. C. (2005). Religiosity, religious participation, and negative prison behaviors. *Journal for the Scientific Study of Religion*, *44*(4), 443–457. http://doi.org/10.1111/j.1468-5906.2005.00296.x

Kim, E. J., Namkoong, K., Ku, T., & Kim, S. J. (2008). The relationship between online game addiction and aggression, self-control and narcissistic personality traits. *European Psychiatry*, *23*(3), 212–218. http://doi.org/10.1016/j.eurpsy.2007.10.010

Kuanliang, A., & Sorensen, J. R. (2008). Predictors of self-reported prison misconduct. *Criminal Justice Studies*, *21*(1), 27–35. http://doi.org/10.1080/14786010801972662

Kuanliang, A., Sorensen, J. R., & Cunningham, M. D. (2008). Juvenile inmates in an adult prison system: Rates of disciplinary misconduct and violence. *Criminal Justice and Behavior*, *35*(9), 1186–1201. http://doi.org/10.1177/0093854808322744

Lahm, K. F. (2008). Inmate-on-inmate assault: A multilevel examination of prison violence. *Criminal Justice and Behavior*, *35*(1), 120–137. http://doi.org/10.1177/0093854807308730

Madsen, E. A., Tunney, R. J., Fieldman, G., Plotkin, H. C., Dunbar, R. I. M., Richardson, J.-M., & McFarland, D. (2007). Kinship and altruism: A cross-cultural experimental study. *British Journal of Psychology*, *98*(2), 339–359. http://doi.org/10.1348/000712606X129213

Maitland, A. S., & Sluder, R. D. (1998). Victimization and youthful prison inmates: An empirical analysis. *The Prison Journal*, *78*(1), 55–73. http://doi.org/10.1177/0032855598078001005

Manning, J. T., & Taylor, R. P. (2001). Second to fourth digit ratio and male ability in sport: Implications for sexual selection in humans. *Evolution and Human Behavior*, *22*(1), 61–69. http://doi.org/10.1016/S1090-5138(00)00063-5

Marcus, J. (2003). Monumentality in archaic states: Lessons learned from large-scale excavations of the past. In J. K. Papadopoulos & R. M. Leventhal (Eds.), *Theory and practice in Mediterranean archaeology: Old world and new world perspectives*. Los Angeles: Cotsen Institute of Archaeology.

Marquart, J. W., Ekland-Olson, S., & Sorensen, J. R. (1989). Gazing into the crystal ball: Can jurors accurately predict dangerousness in capital cases? *Law & Society Review*, *23*(3), 449–468. http://doi.org/10.2307/3053829

Marston, R. (2011, May 11). No rules for the rich: How China spends its new wealth. *BBC News*. Retrieved from http://www.bbc.com/news/business-13284481

Mauss, M. (1954). *The gift: Forms and functions of exchange in archaic societies*. London: Cohen & West Ltd.

McLanahan, S., & Sandefur, G. (1994). *Growing up with a single parent: What hurts, what helps*. Cambridge: Harvard University Press.

Meade, B. (2012). *Examining the effects of religiosity and religious environments on inmate misconduct*. University of South Carolina. Retrieved from http://gradworks.umi.com/35/09/3509537.html

Mead, G. H. (1934). *Mind, self and society from the standpoint of a social behaviorist*. Chicago: Chicago University Press.

Merton, R. K. (1968). *Social theory and social structure*. New York: Free Press.

Merton, R. K. (1987). Three fragments from a sociologist's notebooks: Establishing the phenomenon, specified ignorance, and strategic research materials. *Annual Review of Sociology*, *13*(1), 1–29. http://doi.org/10.1146/annurev.so.13.080187.000245

Merton, R. K., Fiske, M., & Curtis, A. (1946). *Mass persuasion: The social psychology of a war bond drive*. New York: Harper and Brothers.

Mills, C. W. (1959). *The sociological imagination*. Oxford: Oxford University Press.

Mitzkewitz, M., & Nagel, R. (1993). Experimental results on ultimatum games with incomplete information. *International Journal of Game Theory*, *22*(2), 171–198. http://doi.org/10.1007/BF01243649

Morris, R. G., Longmire, D. R., Buffington-Vollum, J., & Vollum, S. (2010). Institutional misconduct and differential parole eligibility among capital inmates. *Criminal Justice and Behavior*, *37*(4), 417–438. http://doi.org/10.1177/0093854810361672

O'Connor, T. P., & Perreyclear, M. (2002). Prison religion in action and its influence on offender rehabilitation. *Journal of Offender Rehabilitation*, *35*(3-4), 11–33. http://doi.org/10.1300/J076v35n03_02

Oltmanns, T. F., Friedman, J. N. W., Fiedler, E. R., & Turkheimer, E. (2004). Perceptions of people with personality disorders based on thin slices of behavior. *Journal of Research in Personality*, *38*(3), 216–229. http://doi.org/10.1016/S0092-6566(03)00066-7

Palmer, E. J., & Thakordas, V. (2005). Relationship between bullying and scores on the Buss-Perry Aggression Questionnaire among imprisoned male offenders. *Aggressive Behavior*, *31*(1), 56–66. http://doi.org/10.1002/ab.20072

Phillips, D. J., & W. Zuckerman, E. (2001). Middle-status conformity: Theoretical restatement and empirical demonstration in two markets. *American Journal of Sociology*, *107*(2), 379–429. http://doi.org/10.1086/ajs.2001.107.issue-2

Poole, E. D., & Regoli, R. M. (1980). Race, institutional rule breaking, and disciplinary response: A study of discretionary decision making in prison. *Law & Society Review*, *14*(4), 931–946.

Porporino, F. J. (1986). Managing violent individuals in correctional settings. *Journal of Interpersonal Violence*, *1*(2), 213–237. http://doi.org/10.1177/088626086001002005

Porporino, F. J., & Dudley, K. (1984). *An analysis of the effects of overcrowding in Canadian penitentiaries*. Ottawa: Ministry of the Solicitor General of Canada.

Posner, E. (1998). Symbols, signals, and social norms in politics and the law. *Journal of Legal Studies*, *27*(S2), 765.

Posner, E. (2002). *Law and social norms*. Cambridge, MA: Harvard University Press.

Powell, B., & Wilson, B. J. (2008). An experimental investigation of Hobbesian jungles. *Journal of Economic Behavior & Organization*, *66*(3–4), 669–686. http://doi.org/10.1016/j.jebo.2006.06.011

Pradel, J., Euler, H. A., & Fetchenhauer, D. (2009). Spotting altruistic dictator game players and mingling with them: the elective assortation of classmates. *Evolution and Human Behavior*, *30*(2), 103–113. http://doi.org/10.1016/j.evolhumbehav.2008.09.003

Rapaport, W. J. (2011). A triage theory of grading: The good, the bad, and the middling. *Teaching Philosophy*, *34*(4), 347–372. http://doi.org/10.5840/teachphil201134447

Rasmusen, E. (2007). *Games and information: An introduction to game theory* (4th Edition). Malden: Blackwell Publishing.

Reidy, T. J., Cunningham, M. D., & Sorensen, J. R. (2001). From death to life prison behavior of former death row inmates in Indiana. *Criminal Justice and Behavior*, *28*(1), 62–82. http://doi.org/10.1177/0093854801028001003

Reisig, M. D. (2002). Administrative control and inmate homicide. *Homicide Studies*, *6*(1), 84–103. http://doi.org/10.1177/1088767902006001005

Riley, J. G. (2001). Silver signals: Twenty-five years of screening and signaling. *Journal of Economic Literature*, *39*(2), 432–478. http://doi.org/10.1257/jel.39.2.432

Robinson, S. (1938). Experimental studies of physical fitness in relation to age. *Arbeitsphysiologie*, *10*(3), 251–323. http://doi.org/10.1007/BF02011412

Rothschild, M., & Stiglitz, J. (1976). Equilibrium in competitive insurance markets: An essay on the economics of imperfect information. *Quarterly Journal of Economics*, *90*(4), 630–649.

Satow, R. (1993). New Yorkers in the countryside: Status conflict and social change. *Journal of Contemporary Ethnography*, *22*(2), 227–248. http://doi.org/10.1177/089124193022002004

Schniter, E., Sheremeta, R. M., & Sznycer, D. (2013). Building and rebuilding trust with promises and apologies. *Journal of Economic Behavior & Organization*, *94*, 242–256. http://doi.org/10.1016/j.jebo.2012.09.011

Schotter, A., Weiss, A., & Zapater, I. (1996). Fairness and survival in ultimatum and dictatorship games. *Journal of Economic Behavior & Organization*, *31*(1), 37–56. http://doi.org/10.1016/S0167-2681(96)00863-3

Scott-Phillips, T. C. (2008). Defining biological communication. *Journal of Evolutionary Biology*, *21*(2), 387–395. http://doi.org/10.1111/j.1420-9101.2007.01497.x

Shields, I. W., & Simourd, D. J. (1991). Predicting predatory behavior in a population of incarcerated young offenders. *Criminal Justice and Behavior*, *18*(2), 180–194. http://doi.org/10.1177/0093854891018002006

Smith, J. M. (1991). Honest signalling: The Philip Sidney game. *Animal Behaviour*, *42*(6), 1034–1035. http://doi.org/10.1016/S0003-3472(05)80161-7

Sorensen, J. R., & Cunningham, M. D. (2007). Operationalizing risk: The influence of measurement choice on the prevalence and correlates of prison violence among incarcerated murderers. *Journal of Criminal Justice*, *35*(5), 546–555. http://doi.org/10.1016/j.jcrimjus.2007.07.007

Sorensen, J. R., & Cunningham, M. D. (2010). Conviction offense and prison violence: A comparative study of murderers and other offenders. *Crime & Delinquency*, *56*(1), 103–125. http://doi.org/10.1177/0011128707307175

Sorensen, J. R., & Davis, J. (2011). Violent criminals locked up: Examining the effect of incarceration on behavioral continuity. *Journal of Criminal Justice*, *39*(2), 151–158. http://doi.org/10.1016/j.jcrimjus.2011.01.003

Sorensen, J. R., & Pilgrim, R. L. (2000). An actuarial risk assessment of violence posed by capital murder defendants. *The Journal of Criminal Law and Criminology (1973-)*, *90*(4), 1251–1270. http://doi.org/10.2307/1144202

Spence, M. A. (1973). Job market signaling. *Quarterly Journal of Economics*, *87*(3), 355–374. http://doi.org/10.2307/1882010

Spence, M. A. (1974). *Market signaling: Informational transfer in hiring and related screening processes*. Cambridge: Harvard University Press.

Sperber, D. (2010). The guru effect. *Review of Philosophy and Psychology*, *1*(4), 583–592. http://doi.org/10.1007/s13164-010-0025-0

Staples, B. (1986). Just walk on by: Black men and public space. *Ms. Magazine*.

Steiner, B., & Wooldredge, J. (2008). Inmate versus environmental effects on prison rule violations. *Criminal Justice and Behavior*, *35*(4), 438–456. http://doi.org/10.1177/0093854807312787

Steiner, B., & Wooldredge, J. (2009a). Individual and environmental effects on assaults and nonviolent rule breaking by women in prison. *Journal of Research in Crime and Delinquency*, *46*(4), 437–467. http://doi.org/10.1177/0022427809341936

Steiner, B., & Wooldredge, J. (2009b). The relevance of inmate race/ethnicity versus population composition for understanding prison rule violations. *Punishment & Society*, *11*(4), 459–489. http://doi.org/10.1177/1462474509341143

Steiner, B., & Wooldredge, J. (2013). Implications of different outcome measures for an understanding of inmate misconduct. *Crime & Delinquency*, *59*(8), 1234–1262. http://doi.org/10.1177/0011128709335151

Sturgis, P. W. (2010). Faith behind bars: An explicit test of the moral community hypothesis in the correctional environment. *Journal of Offender Rehabilitation*, *49*(5), 342–362. http://doi.org/10.1080/10509674.2010.489468

Sykes, G. M. (1958). *The society of captives: A study of a maximum security prison*. Princeton: Princeton University Press.

Thaler, R. H. (1988). Anomalies: The ultimatum game. *Journal of Economic Perspectives*, *2*(4), 195–206. http://doi.org/10.1257/jep.2.4.195

The Economist. (2008, December 18). Con of the century. *The Economist*. Retrieved from http://www.economist.com/node/12818310

The Economist. (2011, February 17). The middle blingdom. *The Economist*. Retrieved from http://www.economist.com/node/18184466

The Economist. (2014a, January 25). Doing it their way. *The Economist*. Retrieved from http://www.economist.com/news/briefing/21595019-market-growing-furiously-getting-tougher-foreign-firms-doing-it-their-way

The Economist. (2014b, February 8). The great expulsion. *The Economist*. Retrieved from http://www.economist.com/news/briefing/21595892-barack-obama-has-presided-over-one-largest-peacetime-outflows-people-americas

Thompson, H. S. (1967). *Hell's angels*. London: Penguin Books.

Toch, H. (1977). *Living in prison: The ecology of survival*. New York: Free Press.

Tolstoy, L. (1952). *War and peace*. Chicago: William Benton.

Trawalter, S., Hoffman, K. M., & Waytz, A. (2012). Racial bias in perceptions of others' pain. *PLoS ONE*, *7*(11), e48546. http://doi.org/10.1371/journal.pone.0048546

United States Department of Justice, B. of J. S. (1997). Survey of Inmates in State and Federal Correctional Facilities, 1997: Codebook. Inter-university Consortium for Political and Social Research.

Veblen, T. (1899). *The theory of the leisure class*. New York: Macmillan Company.

Voracek, M., Manning, J. T., & Dressler, S. G. (2007). Repeatability and interobserver error of digit ratio (2D:4D) measurements made by experts. *American Journal of Human Biology*, *19*(1), 142–146. http://doi.org/10.1002/ajhb.20581

Wolff, N., Shi, J., & Siegel, J. (2009). Understanding physical victimization inside prisons: Factors that predict risk. *Justice Quarterly*, *26*(3), 445–475. http://doi.org/10.1080/07418820802427858

Wooldredge, J., & Steiner, B. (2009). Comparing methods for examining relationships between prison crowding and inmate violence. *Justice Quarterly*, *26*(4), 795–826. http://doi.org/10.1080/07418820802427841

Worrall, J. L., & Morris, R. G. (2011). Inmate custody levels and prison rule violations. *The Prison Journal*, *91*(2), 131–157. http://doi.org/10.1177/0032885511404380

Wright, R. F. (2006). Federal or state? Sorting as a sentencing choice. *Criminal Justice*, *21*(2), 16–21.

Wwd Staff. (2010, July 29). China's luxury profile: Men lead the spending, at home and overseas. *WWD: Women's Wear Daily*. Retrieved from http://www.proquest.co.uk/en-UK/

Zahavi, A. (1975). Mate selection—A selection for a handicap. *Journal of Theoretical Biology*, *53*(1), 205–214. http://doi.org/10.1016/0022-5193(75)90111-3

Zakkour, M. (2014, January 17). The China luxury downturn is real - global luxury brands must adjust. *Forbes*. Retrieved from http://www.forbes.com/sites/michaelzakkour/2014/01/17/the-china-luxury-downturn-is-real-global-luxury-brands-must-adjust/

Zheng, Z., & Cohn, M. J. (2011). Developmental basis of sexually dimorphic digit ratios. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(39), 16289–16294. http://doi.org/10.1073/pnas.1108312108

Appendices

Appendix A. Chapter 2 regression tables

Table A1. Logistic regression outputs for the odds of trustees choosing to return according to their veiled dictator game giving (H1a) and the odds of trusters bestowing trust according to trustees' veiled dictator game giving (H1b). The coefficients are odds ratios.

Dependent variable	Return exp(*b*)	Trust exp(*b*)
Veiled giving	2.22**	1.40**
	(0.34)	(0.05)
Constant	0.15*	0.20**
	(0.15)	(0.10)
Controls		
Age	0.98	1.02
	(0.03)	(0.02)
Female	0.85	1.05
	(0.34)	(0.21)
Session 1	3.03	1.87
	(2.21)	(0.71)
Session 2	1.38	1.20
	(1.02)	(0.41)
Session 3	1.86	1.30
	(1.38)	(0.44)
Session 4	0.48	0.74
	(0.42)	(0.28)
Session 5	1.29	0.84
	(0.97)	(0.33)
Session 6	2.76	1.09
	(1.99)	(0.38)
Session 7	1	1
	(omitted)	(omitted)
Observations	155	2450

Standard errors in parentheses. Standard errors adjusted for clustering at subject level in the regression in which Trust is the dependent variable.
* $p<0.05$, ** $p<0.01$

Table A2. Logistic regression output for the odds of trustees choosing to revise according to their veiled dictator game giving (H2). The coefficients are odds ratios.

Dependent variable	Revise exp(*b*)	Revise exp(*b*)
Veiled giving	4.50**	
	(2.17)	
Veiled giving squared	0.68**	
	(0.09)	
High generosity		0.26*
		(0.17)
Medium generosity		1
		(omitted)
Low generosity		0.29**
		(0.13)
Constant	0.12*	0.42
	(0.12)	(0.39)
Controls		
Age	1.02	1.01
	(0.03)	(0.03)
Female	0.62	0.63
	(0.24)	(0.24)
Session 1	1.23	1.26
	(0.93)	(0.95)
Session 2	2.43	2.40
	(1.76)	(1.74)
Session 3	1.87	1.94
	(1.42)	(1.47)
Session 4	1.68	1.77
	(1.27)	(1.32)
Session 5	1.41	1.41
	(1.11)	(1.11)
Session 6	0.57	0.61
	(0.47)	(0.50)
Session 7	1	1
	(omitted)	(omitted)
Observations	155	155

Standard errors in parentheses. *N.B.* Only one observation per trustee was collected. Therefore, we do not use cluster robust standard errors here.
* $p<0.05$, ** $p<0.01$

Table A3. The logistic regressions show the effect that an increase in unveiled giving, relative to not revising, has on the trusting behaviour of trusters (H4a, H5, H6b). The unveiled giving variable shows the impact of increasing; the constant corresponds to a decision to not revise. The coefficients are odds ratios.

Dependent variable	Sign £0 Trusting exp(*b*)	Sign£1 Trusting exp(*b*)	Sign £2 Trusting exp(*b*)	Sign £3 Trusting exp(*b*)	Sign £4 Trusting exp(*b*)
Unveiled giving	1.27**	1.31**	1.29**	1.12	1.02
	(0.06)	(0.08)	(0.08)	(0.09)	(0.15)
Not revise (constant)	0.04**	0.07**	0.43	2.42	3.04
	(0.03)	(0.07)	(0.30)	(2.14)	(2.85)
Controls					
Age	1.05*	1.06	1.00	0.98	1.00
	(0.03)	(0.04)	(0.02)	(0.03)	(0.04)
Female	1.18	1.16	1.43	1.09	0.98
	(0.36)	(0.36)	(0.44)	(0.38)	(0.36)
Session 1	2.31	0.63	2.45	1.51	2.01
	(1.37)	(0.36)	(1.55)	(1.05)	(1.31)
Session 2	1.90	0.66	1.35	0.87	1.40
	(0.92)	(0.38)	(0.79)	(0.60)	(0.92)
Session 3	0.79	2.41†	1.27	2.69	0.48
	(0.50)	(1.26)	(0.70)	(1.92)	(0.32)
Session 4	0.85	0.56	0.62	0.36	0.53
	(0.47)	(0.37)	(0.37)	(0.23)	(0.32)
Session 5	0.81	1.78	0.92	0.40	0.35
	(0.63)	(0.97)	(0.54)	(0.25)	(0.23)
Session 6	1.57	1.04	1.03	0.62	1.26
	(0.81)	(0.56)	(0.62)	(0.39)	(0.85)
Session 7	1	1	1	1	1
	(omitted)	(omitted)	(omitted)	(omitted)	(omitted)
Observations	447	317	318	328	270

Standard errors in parentheses, adjusted for clustering at subject level. *N.B.* Observations where unveiled giving was equal to or below its corresponding veiled giving value were dropped (e.g. observations which had a veiled giving of £3 and an unveiled giving of £3 or less were dropped). The decision to not revise was then coded to replace the situation where unveiled giving was equal to veiled giving and—veiled giving values were centred on the unveiled giving values.

† p<0.10, * p<0.05, ** p<0.01

Table A4. Output for the logistic regression showing the difference between the odds of trusters placing trust in trustees who appear generous (those appearing to give £4 when veiled) and do not revise and trustees who appear generous and re-give the same amount (H6a). The coefficients are odds ratios.

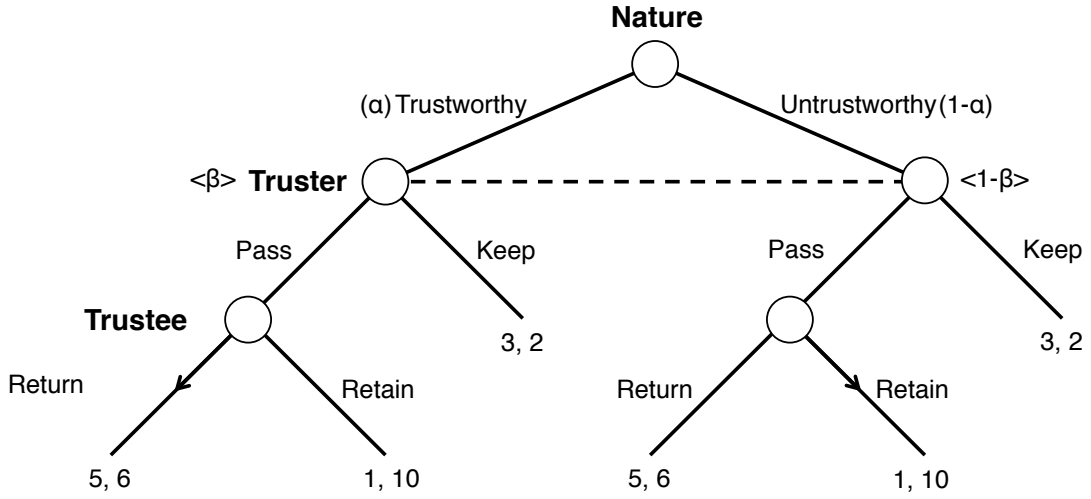
Dependent variable	Trusting exp(*b*)
Revise and re-give the same amount	0.68
	(0.26)
Not revise (constant)	9.13*
	(9.19)
Controls	
Age	0.94†
	(0.03)
Female	0.73
	(0.33)
Session 1	3.00
	(2.55)
Session 2	1.13
	(1.00)
Session 3	0.59
	(0.51)
Session 4	1.35
	(1.11)
Session 5	0.99
	(0.77)
Session 6	1.55
	(1.28)
Session 7	1
	(omitted)
Observations	131

Standard errors in parentheses, adjusted for clustering at subject level.
† $p<0.10$, * $p<0.05$

Appendix B. Trust game, signalling analysis, and instructions

Trust game

Our version of the trust game has the following payoffs. Suppose that there are two types of trustees, a trustworthy trustee who always chooses to return and an untrustworthy trustee who always chooses to retain. Parentheses '( )' contain the probabilities of Nature choosing a type and chevrons '< >' contain the truster's beliefs.



Given our assumption about the two trustee types, the expected utility of a truster from choosing to pass is

$$EU(P) = 5\beta + \left(1-\beta\right), \tag{1.1}$$

and his expected utility from choosing to keep is

$$EU(K) = 3. \tag{1.2}$$

When $5\beta + \left(1-\beta\right) = 3$, the truster is indifferent between his two actions. At this point β, his belief that he is interacting with a trustworthy trustee must be 0.5. When this is the case, the truster is indifferent between keeping and passing and chooses any mix of

the two. When β is greater than 0.5 the truster should pass and when β is less than 0.5 the truster should keep.

Signalling analysis

In this section we carry out a simple analysis of the amounts that the two types of trustees would be able to send in the unveiled dictator game as a signal.

Let γ represent the unconditional probability that a trustworthy trustee believes that he will be trusted, and let s_{TW} indicate the amount that a trustworthy trustee spends on signalling. Assume that trustees will never be trusted unless they spend some amount on signalling which is greater than 0. Here the trustee spends more than 0, thus he can expect to receive

$$EU_{TW}\left(s_{TW}\right)=6\gamma+2\left(1-\gamma\right)-s_{TW}, \tag{1.3}$$

which simplifies to $4\gamma+2-s_{TW}$.

As trustees are not trusted unless they spend some amount on signalling, when they spend nothing on signalling (s_{TW}=0) their expected utility is 2. This trustee becomes indifferent between his two actions, of signalling something and signalling nothing, when his two expected utilities are equal: $4\gamma+2-s_{TW}=2$, which simplifies to $4\gamma=s_{TW}$.

Turning to an untrustworthy trustee. Let δ represent the unconditional probability that he attaches to being trusted by a truster and let s_{UW} be the amount that he spends on signalling in the unveiled dictator game. If s_{UW} is more than 0 he will expect to receive

$$EU_{UW}\left(s_{UW}\right)=10\delta+2\left(1-\delta\right)-s_{UW}, \tag{1.4}$$

which simplifies to $8\delta + 2 - s_{UW}$.

This trustee becomes indifferent between signalling and being trusted some of the time and not signalling when $8\delta + 2 - s_{UW} = 2$, which simplifies to $8\delta = s_{UW}$.

When $2\delta = \gamma$ a trustworthy and an untrustworthy trustee can afford to spend an equal amount on signalling. Before the trustworthy trustee is able to spend more on signalling, $2\delta < \gamma$. That is, the probability that a trustworthy trustee places on being trusted by a truster has to be more than twice as much as the probability that an untrustworthy trustee places on being trusted.

If we fix the probability that a trustworthy trustee and an untrustworthy trustee are trusted at an equal amount ($\delta = \gamma = 1$) we can see that a trustworthy trustee can afford at maximum to spend 4 on signalling ($4 = s_{TW}$) before preferring to not signal and that an untrustworthy trustee can afford to spend 8 on signalling ($8 = s_{UW}$) at most before preferring to not signal.

If instead we allow the cost of signalling to vary according to whether the trustee is trustworthy or untrustworthy we see that the cost of signalling has to be half as much for a trustworthy trustee than it is for an untrustworthy trustee before the former can spend as much as the latter.

Experimental instructions

Dear participant,

Welcome and thank you for participating in this experiment. Please take your time to read through the following instructions thoroughly. You can take notes if you wish to do so.

Important rules

First, we wish to inform you about a number of rules and practical details.

- This session will last for about 60 minutes.
- Your participation is voluntary and you are free to leave at any point, if you wish to do so. In that case, we will only pay you the show-up fee of £4.
- We ask you not to communicate verbally or otherwise with the other participants and generally to remain quiet from now until the end of the session. Those who do not respect the silence requirement will be asked to leave.

What will happen at the end of the session?

Once the session is finished, please remain seated. We will need around 10 minutes to prepare your payment. You will be called up successively by the number on your table, you will then receive an envelope with your earnings and you will be asked to sign a receipt.

General information about the experiment

You are participating in an experimental session in which you will earn some money. The money you earn during this session will be added to your show-up fee of £4 and paid in cash at the end of the session. Your earnings depend on both the decisions you will make and the decisions other participants will make. Up to 30 people are participating in this session. All decisions taken in this experiment are anonymous and the payments are administered by a person not involved in the implementation of this experiment. The data collected in this experiment are for scientific purposes only.

This experiment consists of three linked parts: Part 1, Part 2 and Part 3. If you are in the experimental laboratory sitting in front of a computer, you will be paid for a decision in Part 1 or Part 2. You will also be paid for a randomly chosen decision in Part 3. If you are in the seminar room you will be paid for Part 1 and Part 2.

You will now be given instructions concerning the first part of the experiment. After, you will be asked to answer a few control questions about these instructions on your computer. Next, the correct answers will be shown and the experimenter will read out and explain them. Finally you will have the opportunity to ask questions. The first part of the experiment will then start. The following parts will proceed in a similar way.

Instructions for Part 1

For the first part of this experiment (Part 1) you have been randomly assigned to be a "Sender" or a "Recipient". If you are a Sender, you will be endowed with £8. If you are a Recipient, you will receive no endowment.

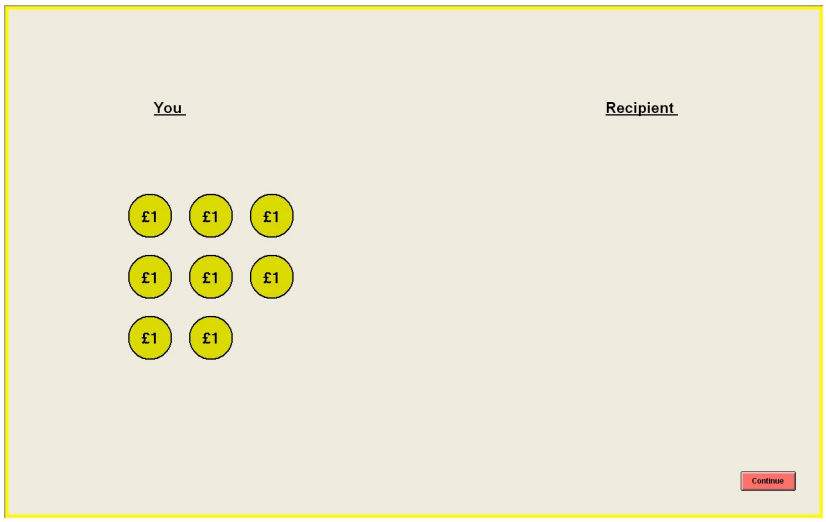
Five participants in this session have been randomly assigned to be Recipients. The other participants have been assigned to be Senders. The Senders are in the experimental laboratory in front of a computer and the Recipients are in the seminar room on the same floor. Thus, you are a Sender. All participants are reading the same instructions.

Senders will be asked to select how many pounds they would like to allocate to a Recipient. A Recipient will receive this amount and the Sender will keep the rest. For example, if the Sender decides to send £3, a Recipient will receive £3 and the Sender will keep £5. That is, Senders will keep the amount according to their decision.

The Recipients' earnings are calculated as follows. Once all Senders have made their decision, as many numbered cards (from 0-8) will be put in an opaque box as there were £0 to £8 choices made by the Senders. Next, each Recipient will draw one card from the box. Whatever the number the Recipient draws will be the amount in pounds he or she receives. Each card will be put back into the box before the next Recipient gets to draw a card. Remember, this interaction is anonymous.

As a Sender, you will see eight £1 coins on your computer which start off on the left-hand side of the screen (shown in Figure 1). This represents how much money you – as the Sender – will keep. To allocate money to a Recipient, click on the image of a one pound coin. This sends it to the Recipient's side of the screen. To move it back to your side of the screen, click on any coin on the Recipient's side. Simply put, clicking a coin moves it from your endowment to a Recipient's endowment and *vice versa*. You can repeat this process as many times as you wish before settling on a value and clicking "Continue". Once you have clicked the "Continue" button you will be moved to the next part of the experiment.

Figure 1. The screen you will see as a Sender.



If you are a Sender please proceed on the screen and answer the control questions. If you are a Recipient please wait for further instructions. You may use these instructions to answer the questions.

Instruction for Part 2

In the second part of the experiment (Part 2) Senders have the opportunity to revise how much they sent to a Recipient. As in the previous part Senders will be endowed with £8 and Recipients will be given no endowment. You will keep your role from the previous part as a Sender or a Recipient.

After the control questions have been answered and feedback has been given, you will be asked – as a Sender – whether you wish to revise your decision (shown in Figure 2). If you wish to revise how much you sent to a Recipient you will be taken to a screen identical to that in Part 1. If you do not wish to revise your decision you will be taken to a waiting screen while others make their choice.

Figure 2. The screen you will see when you decide if you wish to revise your decision.

If you choose to revise your decision, you will be paid randomly for how much you kept in Part 1 or how much you kept in Part 2. If you choose not to revise your choice, you will be paid the amount you kept in Part 1. The payments of Recipients are calculated identically to that in Part 1 (page 3); Senders who choose not to revise their choice will automatically send the same amount to a Recipient as they did in Part 1.

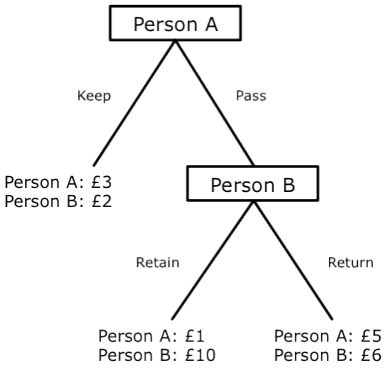
You will then proceed to Part 3 of the experiment where you will receive new roles. Recipients will be paid and leave at the end of Part 2.

Instructions for Part 3

In the final part of the experiment you will take part in the following situation. First as a "Person B" and then as a "Person A".

As a Person A, you will be endowed with £5 and asked to choose between the actions "Keep" and "Pass". As a Person B, you will receive no endowment and asked to choose between the actions "Retain" and "Return" (shown in Figure 3).

Figure 3. A diagram of the situation you will participate in for Part 3 of the experiment.

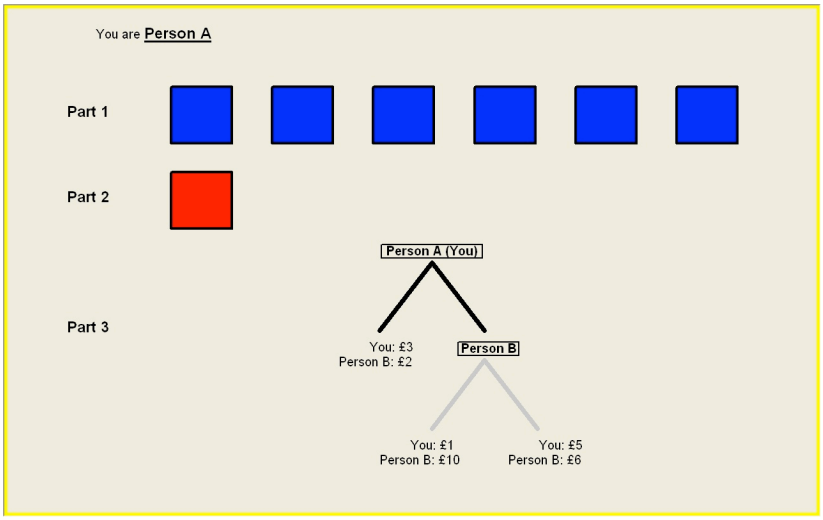


- If Person A chooses the action "Keep", the £5 will be divided between Person A and Person B such that Person B will receive £2 and Person A gets to keep £3.
- If Person A chooses the action "Pass", the £5 will be turned into £11 and sent to Person B. Person B will be asked to choose between the action "Retain" and the action "Return".

- If Person B chooses the action "Retain", the £11 will be divided between Person A and Person B such that Person A will receive £1 and Person B gets to keep £10.
- If Person B chooses the action "Return", the £11 will be divided between Person A and Person B such that Person A will receive £5 and Person B gets to keep £6.

When you make your choice as Person A, you will see a screen which contains six blue cards and a diagram identical to the one shown in Figure 3. You may or may not also see a single red card (shown in Figure 4).

Figure 4. The screen you will see as Person A.



The six blue cards cover numbers that show how much Person B sent to a Recipient in Part 1 of the experiment. Four of the six cards cover the exact value that Person B sent, while one card covers a value that is one pound more and one card covers a value that is one pound less than what he or she sent. As a Person A, you will click on one card to reveal the value underneath it. You can only reveal one card.

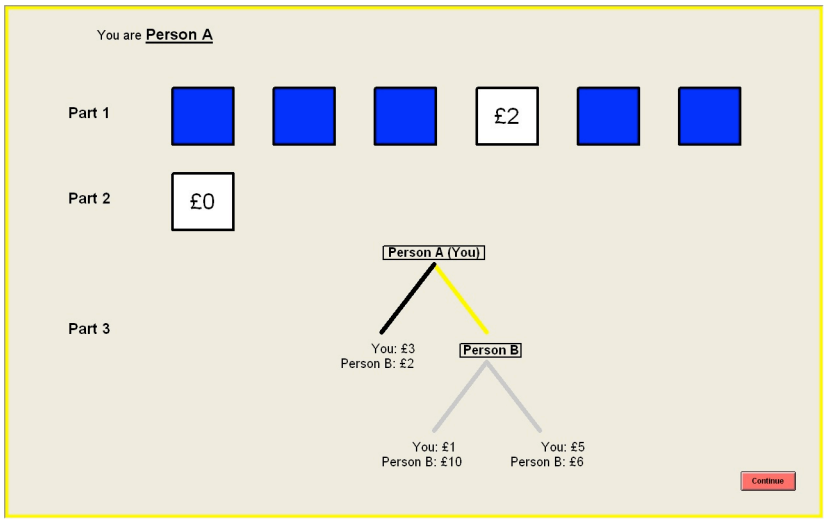
The single red card, which may or may not be present, refers to Person B's decision in Part 2 of the experiment. If the red card is present, it means that Person B chose to revise his or her decision in Part 2. Moreover, the red card covers the exact amount that Person B chose to send in Part 2. Clicking on the card reveals the value underneath it. If Person B did not choose to revise his or her decision, you will not see the red card. You must select a blue card before clicking on the red card (if it is present).

The diagram on-screen (below the cards) is where you will make your decision to Keep your endowment or Pass it to Person B. The branches correspond to a decision, and the branches that are available for you to select are shown in black. The branches you cannot select are shown in grey. As a Person A you will be asked to choose between one of the two upper

branches. Clicking on a branch selects it and this is shown by the branch being highlighted in yellow.

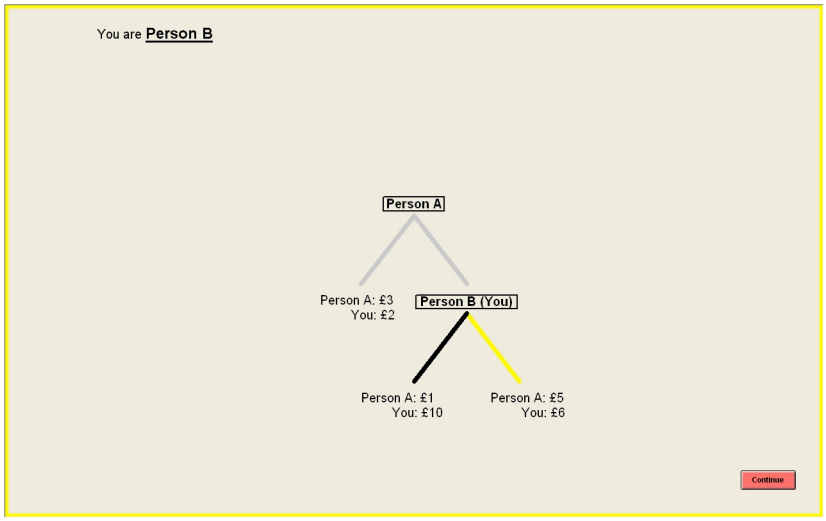
Once the cards have been revealed and a decision has been made, the "Continue" button will appear (shown in Figure 5).

Figure 5. The screen you will see as Person A once you have revealed the cards and made a decision.



When you make your choice as Person B, you will see a screen which contains a diagram identical to the one in Figure 3. Here you will be asked two choose between one of the two lower branches (shown in Figure 6). The "Continue" button will appear once you have selected a branch.

Figure 6. The screen you will see as Person B.



You will play the role of Person B once and the role of Person A a number of times. In one of the rounds as Person A you will be randomly matched with another person in this room. In the other rounds you will face hypothetical situations. You will not know when you are matched with another person or when you are facing a hypothetical scenario.

In the round where you are matched with another person, you will see the decisions this person made in Part 1 and Part 2 of the experiment and you will decide whether to Pass your endowment to this person or Keep it. The outcome of this interaction will be determined by whether the person you are paired with chose to Retain or Return the endowment when he or she made his or her choice as Person B. The same procedure applies for the person you are paired with: he or she will see your decisions in Part 1 and Part 2 of the experiment and his or her outcome depends on the choice he or she makes as Person A and the choice you made as Person B.

At the end of the experiment you will be randomly paid for the outcome of the interaction when you made your choice as Person A or the outcome of the interaction when you made your choice as Person B.

Simply put, you will be paid for the interaction in which you are paired with another person, and your payment will depend on what you chose and what the other person chose.

To summarise, in Part 3 you will play one round as a Person B and then a number of rounds as a Person A. One of your choices as Person A or Person B – in the round you are matched to another person – will be randomly selected and used for payment.

As a Person A you will:

- See six blue cards. Underneath four are the exact amount Person B sent in Part 1. Underneath one is a value that is one pound higher and underneath another is a value that is one pound lower.
- See a single red card if Person B chose to revise his or her decision. Beneath this is the exact value he or she sent. If you do not see the red card it means Person B did not revise his or her decision.
- Select one of the two actions "Keep" or "Pass" using the branches in the diagram.

As a Person B you will:

- Select one of the two actions "Retain" or "Return" using the branches in the diagram.

You will now proceed to the control questions and the answers to the control questions. After this you will proceed to Part 2 where you can revise how much you send. Then you will proceed to Part 3 where you will first make your choice as Person B and then your choice as Person A in a number of situations.

Appendix C. Chapter 3 regression table

Table C1. Full output from negative binomial regressions.

Variables	State facilities		Federal facilities	
	Coefficient (S.E.)	IRR	Coefficient (S.E.)	IRR
Age	-0.07*** (0.01)	0.93	-0.04*** (0.01)	0.96
Number of prior incarcerations	0.09*** (0.03)	1.10	0.04 (0.07)	1.04
Years in education	0.05** (0.02)	1.06	-0.03 (0.03)	0.97
Race/ethnicity				
White	Reference category		Reference category	
Black non-Hispanic	-0.78*** (0.09)	0.46	-0.02 (0.19)	0.98
Hispanic	-0.59*** (0.11)	0.56	-0.61* (0.27)	0.54*
Other non-Hispanic	-0.54** (0.17)	0.58	1.10*[a] (0.48)	2.99*
Murder	0.04 (0.08)	1.04	0.17 (0.30)	1.19
BMI (log)	-0.29 (0.22)	0.75	-0.65 (0.70)	0.52
Number of inmates in cell (log)	-0.04† (0.03)	0.96	0.69* (0.29)	1.99
Number of inmates in cell (log) * time served (log)		-	-0.18* (0.08)	0.83
Working within the facility	-0.20** (0.07)	0.82	-0.45† (0.27)	0.64
TV available to watch	-0.20† (0.11)	0.82	-0.26 (0.37)	0.77
Newspapers, magazines, books, available to read	0.09 (0.11)	1.09	-0.58 (0.66)	0.56
Security				
Maximum	Reference category		-	
High	-		Reference category	
Medium	-0.17* (0.08)	0.85	-0.16 (0.19)	0.85
Low	-		-1.25*** (0.28)	0.29
Minimum	-0.50* (0.21)	0.61	-1.73*** (0.45)	0.18

Administrative		-	-1.29*** (0.34)	0.28
Time served (log)	1.13*** (0.05)	3.08	1.61*** (0.17)	4.98
Mental health issues	0.41*** (0.11)	1.50	0.60* (0.25)	1.81
Involved with antisocial peer group	0.37*** (0.09)	1.45	0.76*** (0.17)	2.14
Work outside facility	-0.37** (0.14)	0.69		-
Drug use	0.20** (0.08)	1.22		-
Abused	0.57*** (0.10)	1.77		-
Violent crime		-	0.55* (0.19)	1.74
Time left (months)		-	-0.003** (0.001)	0.996
Married		-	-0.49* (0.21)	0.61
Employed		-	0.49* (0.21)	1.63
Region				
North East		Reference category		
Mid-West	0.17 (0.12)	1.18	0.60* (0.29)	1.82
South	0.48*** (0.10)	1.62	0.24 (0.30)	1.27
West	0.11 (0.15)	1.12	-0.02 (0.35)	0.98
Mean prisoner age	-0.04*** (0.01)	0.96	0.09** (0.03)	1.09
Proportion used drugs	-0.42 (0.44)	0.66		-
Proportion in activity		-	-2.25* (0.89)	0.11
Constant	-0.01 (0.86)	0.99	-4.72 (3.08)	0.01
Observations (facilities)	207		32	
Observations (inmates)	9911		2721	

† p<0.10, * p<0.05, ** p<0.01, *** p<0.001

[a] Positive relationship based entirely on a single observation: one prisoner who is other non-Hispanic and who reports that he was involved in fights 45 times. Removing this single observation removes the significance of this coefficient.

Appendix D. Chapter 4 regression tables

Table D1. Regression output for the determinants of unveiled squatting time.

Dependent variable	Veiled squat time
	b
Age	-2.18
	(1.80)
BMI	-1.92
	(2.04)
Frequency of exercise	2.86
	(5.10)
Pain resistance	13.45*
	(6.39)
Fitness	13.78†
	(7.38)
Physical aggression (BPAQ)	0.15
	(1.84)
Verbal aggression (BPAQ)	2.08
	(2.60)
Anger (BPAQ)	-0.11
	(3.06)
Hostility (BPAQ)	Omitted
BPAQ score	-0.59
	(1.49)
2D:4D	-283.74*
	(126.23)
Constant	402.15**
	(140.80)
Observations	137
R-squared	0.22

Standard errors in parentheses
† p<0.10, * p<0.05, ** p<0.01

Table D2. Logistic regression output for residents' decisions to challenge and rookies' decisions to resist according to treatment group.

Dependent variable	Challenge exp(*b*)	Resist exp(*b*)
Countersignalling treatment	0.27*	0.44
	(0.17)	(0.22)
Signalling treatment	0.25*	0.41†
	(0.16)	(0.21)
Control group	Omitted	Omitted
Constant	7.67**	2.25†
	(4.72)	0.96
Observations	530	134
Pseudo R-squared	0.01	0.02

Standard errors in parentheses, cluster robust standard errors for 'Challenge'
† p<0.10, * p<0.05, ** p<0.01

Table D3. OLS regression outputs for residents' and rookies' beliefs in the treatment groups.

Dependent variable	Residents' beliefs about rookies retaining *b*	Residents' beliefs about rookies winning *b*	Rookies' beliefs about rookies winning *b*
Countersignalling treatment	-1.16	-13.78*	-11.16†
	(4.82)	(5.691)	(6.42)
Signalling treatment	-7.25	-11.21*	-7.88
	(5.01)	(5.63)	(6.66)
Control group	Omitted	Omitted	Omitted
Constant	62.54***	61.77***	58.73***
	(4.34)	(4.84)	(5.36)
Observations	530	530	134
R-squared	0.01	0.01	0.02

Standard errors in parentheses. Cluster robust errors for regressions with 530 observations.
† p<0.10, * p<0.05, ***p<0.001

Table D4. Logistic regression outputs for veiled and unveiled squatting and resisting. Coefficients are odds ratios.

Dependent variable	Resist (all rookies)	Resist (signalling treatment)	Resist (countersignalling treatment)
	exp(*b*)	exp(*b*)	exp(*b*)
Veiled squat time (s)	1.02***		
	(0.004)		
Unveiled squat time (s)		1.01†	1.05***
		(0.08)	(0.01)
Constant	0.11***	0.28†	0.04**
	(0.06)	(0.20)	(0.04)
Observations	139	49	64
Pseudo R-squared	0.19	0.07	0.27

Standard errors in parentheses
† p<0.1, ** p<0.01, *** p<0.001

Table D5. Logistic regression output for the decision of residents to challenge based on the veiled and unveiled grade that they observed in the signalling and countersignalling treatment. Coefficients are odds ratios.

Dependent variable	Challenge (countersignalling treatment)	Challenge (signalling treatment)	Challenge (countersignalling treatment)
	exp(*b*)	exp(*b*)	exp(*b*)
Veiled grade A	0.43***		
	(0.06)		
Unveiled squat time (s)		0.27***	0.43***
		(0.07)	(0.05)
Constant	3.36***	31.71***	12.67***
	(0.72)	(19.65)	(4.26)
Observations	384	147	384
Pseudo R-squared	0.03	0.16	0.08

Cluster robust standard errors in parentheses
*** p<0.001

Table D6. Regression output for unveiled squatting times in the control and signalling groups.

Dependent variable	Unveiled squat time *b*
Control group	Omitted
Signalling treatment	-10.74 (16.46)
Constant	100.81*** (13.31)
Observations	75
Pseudo R-squared	0.01

Standard errors in parentheses
*** p<0.001

Table D7. Regression output comparing unveiled squatting times against veiled squatting times.

Dependent variable	Unveiled squat time *b*
Veiled squat time (s)	0.57*** (0.09)
Constant	18.7 (11.85)
Observations	64
Pseudo R-squared	0.40

Standard errors in parentheses
*** p<0.001

Table D8. Regression output comparing unveiled squatting times by type.

Dependent variable	Unveiled squat time *b*
High type	30.03† (17.46)
Medium type	Omitted
Low type	-37.35* (17.84)
Constant	85.01*** (12.77)
Observations	64
Pseudo R-squared	0.20

Standard errors in parentheses.
† p<0.10, * p<0.05, *** p<0.001

Table D9. Regression output for comparing high type unveiled squatting times in the countersignalling treatment to their unveiled squatting in the signalling treatment.

Dependent variable	Unveiled squat time *b*
Countersignalling treatment	2.40 (24.05)
Signalling treatment	Omitted
Constant	100.81*** (13.31)
Observations	75
Pseudo R-squared	0.01

Standard errors in parentheses
*** p<0.001

Table D10. Logistic regression output for the treatment of rookies by residents. Coefficients are odds ratios.

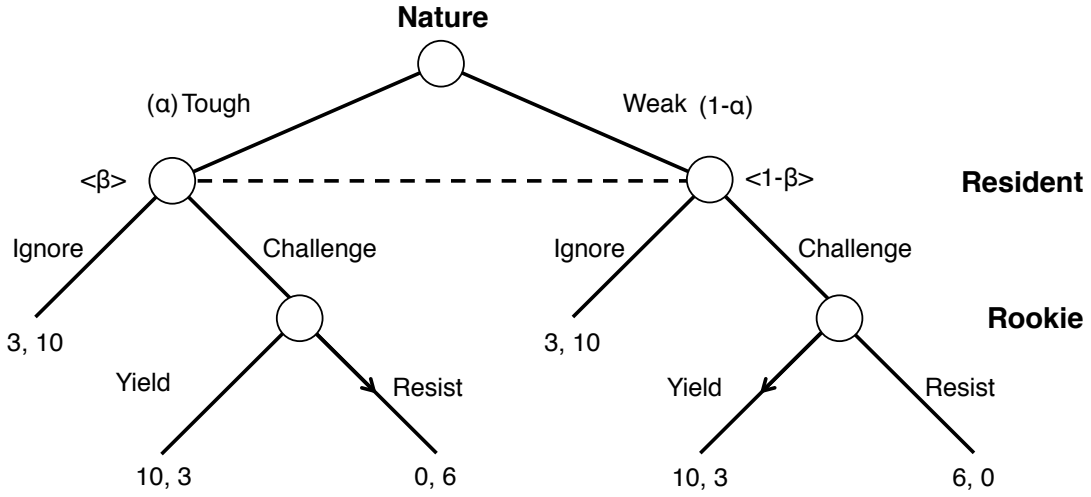
Dependent variable	Ignore exp(*b*)	Ignore exp(*b*)	Ignore exp(*b*)	Ignore exp(*b*)	Ignore exp(*b*)	Ignore exp(*b*)
CC	Omitted	0.20** (0.12)	0.11*** (0.06)	0.24** (0.11)	0.10*** (0.05)	0.05*** (0.03)
CB	5.00** (2.98)	Omitted	0.56* (0.14)	1.19 (0.42)	0.49** (0.13)	0.23*** (0.07)
CA	9.00*** (4.49)	1.80* (0.45)	Omitted	2.14** (0.54)	0.88 (0.14)	0.41*** (0.10)
AC	4.20** (1.92)	0.84 (0.30)	0.47** (0.12)	Omitted	0.41** (0.11)	0.19*** (0.06)
AB	10.26*** (5.17)	2.05** (0.54)	1.14 (0.18)	2.44** (0.65)	Omitted	0.47** (0.11)
AA	21.92*** (12.30)	4.38*** (1.39)	2.44*** (0.58)	5.22*** (1.76)	2.14** (0.52)	Omitted
Constant	0.07*** (0.03)	0.33*** (0.09)	0.60† (0.16)	0.28*** (0.09)	0.68 (0.18)	1.46 (0.37)
Observations			384			
Pseudo R-squared			0.11			

Cluster robust standard errors in parentheses
† p<0.10, * p<0.05, ** p<0.01, *** p<0.001

Appendix E. Resident-rookie game, instructions, and ethical approval

Resident-rookie game

My resident-rookie game has the following payoffs. Suppose that there are two types of rookies, a tough type who always resists and wins in a squat-off, and a weak type who never resists and would lose in a squat-off. Parentheses '( )' contain the probabilities of Nature choosing a type and chevrons '< >' contain the resident's beliefs.



Given the assumption about the two rookie types, the expected utility of a resident from choosing to challenge is

$$EU(C) = 10\left(1-\beta\right), \tag{1.5}$$

and his expected utility from choosing to ignore is

$$EU(I) = 3. \tag{1.6}$$

When $10\left(1-\beta\right) = 3$, the resident is indifferent between his two actions. At this point β, his belief that he is interacting with a tough type rookie must be 0.7. When this is the case, the resident is indifferent between challenging and ignoring and chooses any

probability mix of the two. When β, the resident's belief of encountering a tough type rookie, is greater than 0.7 the resident should ignore, and when β is less than 0.7 the resident should challenge.

Experimental instructions

The following instructions concern the countersignalling treatment, which is the most complex treatment. They are the English versions of the Italian language instructions. Payoffs shown in the figures of the instructions are in pound sterling; in the Italian instructions they are displayed in euros on a one-to-one basis e.g. £6 is converted to €6.

Dear participant,

Welcome and thank you for participating in this experiment. Please take your time to read through the following instructions carefully. You can take notes if you wish to do so. You will have the chance to ask questions later.

Important rules

- This experiment consists of two sessions. The first session (the one you are currently in) lasts for about 80 minutes. The second session, scheduled for Friday 6th of June, lasts for about 20 minutes.
- Your participation is voluntary and you are free to leave at any point if you wish to do so. In that case, we will only pay you the show-up fee of €4.
- We ask you not to communicate verbally or otherwise with the other participants and generally to remain quiet from now until the end of the session. If you breach this rule you may be asked to leave.
- All data collected in this experiment are anonymous and are used for scientific purposes only.

What will happen at the end of the session?

Once this session ends, please remain seated. We will need around 5 minutes to prepare your payment.

You will then be provided with your show-up fee and asked to sign a receipt. At this point, we will take a scan of your palm.

General information about the experiment

You are participating in an experimental session in which you will earn some money. At the end of this session you will be paid your first show-up fee of €4.

Your total earnings from the experiment will be calculated at the end of the second session. You will then be paid in cash both a second show up fee of €4 as well as the earnings from the entire experiment. You are required to attend both sessions to receive full payment.

Your earnings depend on both the decisions you make and the decisions other participants make.

This session consists of three parts: Part 1, Part 2, and Part 3. Your total earning will be based on your actions in Part 1 and on one of your decisions in Part 3. In this session you will complete Part 1, Part 2, and most of Part 3. In the second session you will finish off Part 3.

The following instructions concern Part 1 of the experiment.

Part 1 Instructions

In this part you are asked to carry out a simple 'wall sit' exercise in order to earn some money.

To perform the exercise:
1) Stand about two feet away from the wall with your back to it.
2) Gently lean back against the wall and slide down until your knees form right angles, so that there is a 90 degree angle between your shins and thighs.
3) Place both hands on the wall behind you.

The correct form is shown in Figure 1 below.

Figure 1. The correct form of the wall sit exercise.

The longer you stay in the seated position of this exercise, the more money you earn. For every second that you stay in the seated position you earn 2 cents. For instance, if you stay sitting for 3 minutes (180 seconds) you earn €3.60. Your payment will be rounded to the nearest 10 cents.

Remain seated for as long as you wish to do so. To stop the task, gently lower yourself to the ground onto the pillow provided. Please remain in this position until the experimenters stop the exercise, which will be 6 minutes after the start of the exercise.

During the exercise the experimenters may adjust your technique.

Each of you will receive, and wear, a sleeping mask and a pair of ear protectors. These will prevent you from seeing or hearing the other people also doing this exercise. We require you to wear these because we do not want others to influence you and *vice versa*.

When an experimenter tells you to, first put the mask on and then the ear protectors. Please keep them on until an experimenter tells you to remove them (this will be at the end of the 6 minutes).

Before the exercise is started an experimenter will demonstrate the correct form for the wall sit. After this you will have the opportunity to practice it.

If you have any joint, back or knee problems, or feel that you cannot carry out this task safely, inform an experimenter now.

The experimenters will indicate the start of the exercise by saying 'start the exercise'. You will be able to hear this through the ear protectors.

Please now raise your hands if you have any questions.

The instructions for Parts 2 and 3 will be described after you complete Part 1. Now proceed with Part 1.

Everyone who drew an even number will now move, along with the experimenters, to the other room to do their wall sits. People who drew an odd number will wait here until we return. Then, they will move, along with the experimenters, to the other room to do their wall sits while the people who drew an even number wait here.

Part 2 Instructions

You have been allocated a grade according to how many seconds you held the wall sit position for in Part 1.

The experimenters will shortly hand each of you a piece of paper with your grade on it.

Grades represent how long you held the sitting position for relative to other people in this session.

- An A grade indicates that your wall sit time is in the top third of the wall sit times.
- A B grade indicates that your wall sit time is in the middle third of the wall sit times.
- A C grade indicates that your wall sit time is in the bottom third of the wall sit times.

Simply put, an A grade indicates the longest wall sit time, a B grade indicates an intermediate time, and a C grade indicates the shortest wall sit time.

In Part 2 of the experiment, you will do the wall sit exercise again. Based on how long you do the wall sit for in Part 2, you receive another grade. This is assigned by comparing your Part 2 wall sit time to the distribution of wall sit times that were created in Part 1. Thus, your Part 2 grade indicates how long you hold the sitting position for relative to the times that everyone in this session obtained in Part 1.

In Part 2 you are not paid for doing the wall sit.

You will then proceed to Part 3 of the experiment.

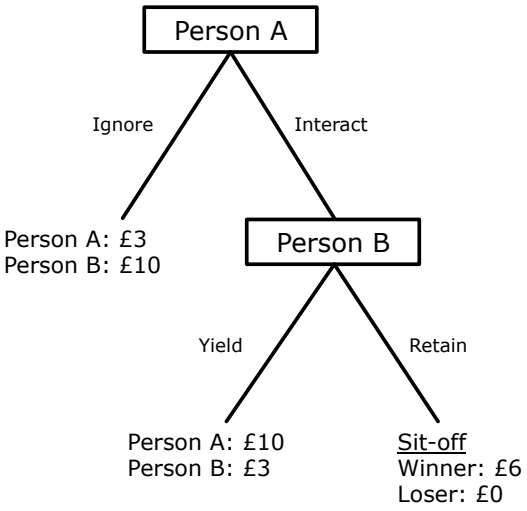
Before you perform Part 2 please read carefully the instructions of Part 3 below.

Part 3 Instructions

The final part of this experiment takes place in front of the computers in the lab. Here you will be in a situation against other people who are also taking part in this experiment.

This situation is shown in the diagram below (Figure 2). Refer back to this figure to help make sense of the instructions.

Figure 2. A diagram of the situation that you will engage in for Part 3 of the experiment.



Keep the following in mind to make sense of the diagram:

- The two rectangles containing 'Person A' and 'Person B' show that there are two people in this situation, one called Person A and the other called Person B.
- Each person has two possible actions to choose from—represented by the two pairs of branches with names next to them.
- The amount of money Person A and Person B receive from each combination of actions is indicated at the end of each branch.

You will participate in the situation as both Person A and Person B.

Person B
Put yourself first in the role of Person B.

As Person B we randomly match you with a Person A, another participant in the experiment.

We will give some information to this Person A about the grades that you achieved in Part 1 and Part 2.

Information on your Part 1 grade is sent in the form of two blue cards. Person A will be allowed to pick only one of these two cards to discover the information that is underneath it. What the two blue cards cover depends on the grade that you achieved in Part 1 of this experiment.

- If you received an A grade in Part 1, then both cards cover the letter 'A'.
- If you received a B grade in Part 1, then one card covers the letter 'C' and the other card covers the letter 'A'.
- If you received a C grade in Part 1, then both cards cover the letter 'C'.

Because Person A can select only one of the two blue cards to reveal the grade underneath it, he will have some uncertainty about the grade that you actually attained in Part 1.

In addition to your Part 1 grade, we automatically send your Part 2 grade to Person A. This is sent as a single red card.

This single red card covers the grade that you really obtained in Part 2 of the experiment. Person A will turn it over to reveal the letter underneath it. The red card always reveals your Part 2 grade correctly.

Finally we will put €10 at your disposal; how much of that sum you will actually earn will depend both on your decision of whether to 'Yield' or to 'Retain' and on the decision of the Person A you are against. Each person will make their decision without knowing what the other person will decide.

If you choose to 'Yield' and:
- Person A chooses to 'Ignore', you earn €10 and Person A earns €3.
- Person A chooses to 'Interact', you earn €3 and Person A earns €10.

If you choose to 'Retain' and:
- Person A chooses to 'Ignore', you earn €10 and Person A earns €3.
- Person A chooses to 'Interact', you both take part in a 'sit-off' situation where the winner earns €6 and the loser earns €0. Sit-offs will be conducted in the second session on Friday the 6th of June.

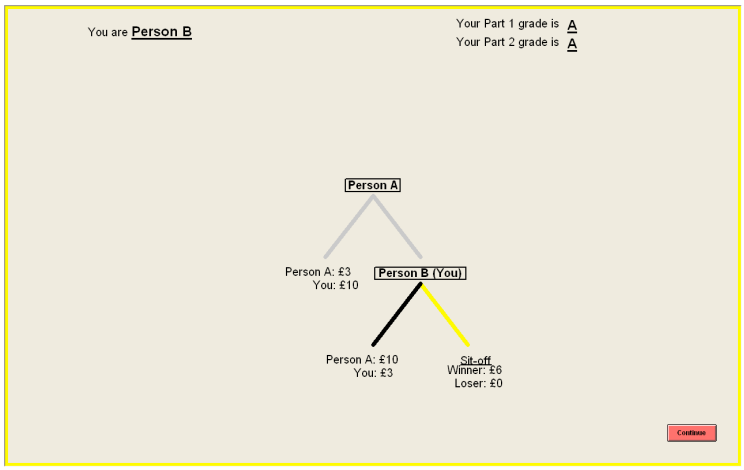
In a sit-off situation, both Person A and Person B carry out the wall sit exercise simultaneously without seeing or otherwise perceiving each other's performance. The

person who maintains the position for longer receives €6, while the person who maintains the position for a shorter time receives €0.

As Person B you receive no information about the grades of the Person A you face.

When you make your decision as Person B you will see a screen containing a diagram of your situation, and you are asked to choose between one of the two options (shown in Figure 3). Clicking on a branch selects it and this is shown by the branch being highlighted in yellow. You may change your decision as many times as you wish before clicking on the 'Continue' button.

Figure 3. The screen you will see as Person B.



A 'Continue' button will appear once you have selected a branch. Upon clicking 'Continue' you will be asked a question about your decision on-screen.

Person A
Now put yourself in the position of Person A. We will repeat what we said above, but this time from the perspective of a Person A.

As Person A we randomly match you with a Person B.

Before you make your decision to 'Ignore' or 'Interact' you discover two things about the Person B you are against.

You will see two blue cards on your screen that will give an imperfect indication of Person B's grade for Part 1 of the experiment. As described before, if Person B obtained an A grade in Part 1 then both cards cover the letter 'A', if Person B obtained a B grade in Part 1 then one card covers the letter 'A' and one card covers the letter 'C', and if Person B obtained a C grade in Part 1 then both cards cover the letter 'C'. As a Person A, you will click on one of the blue cards to reveal what is underneath it.

You can only reveal one card (introducing some uncertainty).

In addition, you will see a single red card which covers the grade that Person B obtained in Part 2 of the experiment. This always covers Person B's Part 2 grade correctly.

Finally we put €10 at your disposal, but how much of that sum you will earn depends both on your decision to 'Ignore' or to 'Interact' and on the action of the Person B you are against.

If you choose to 'Interact' and:

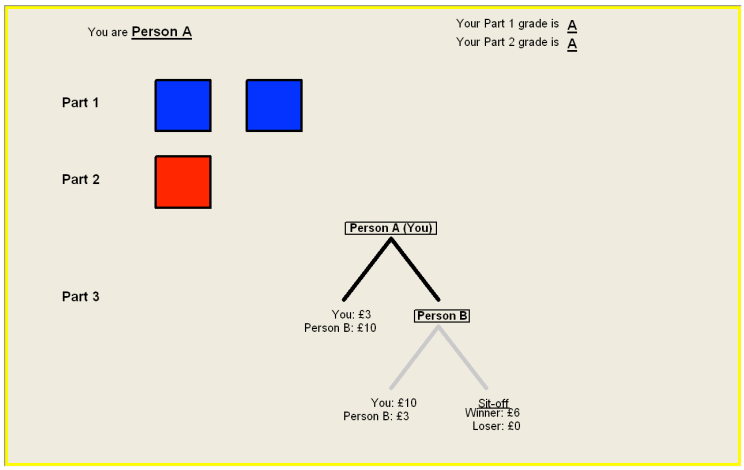
- Person B chooses to 'Yield', you earn €10 and Person B earns €3.
- Person B chooses to 'Retain', you both take part in a 'sit-off' where the winner earns €6 and the loser earns €0. These will be conducted in the second session on Friday the 6th.

If you choose to 'Ignore', regardless of what action Person B takes the situation ends, you earn €3, and Person B earns €10.

When you make your decision as Person A you will see a screen containing two blue cards, a single red card, and a diagram of the situation that you are taking part in (shown in Figure 4 below). Clicking on a card uncovers the grade beneath it.

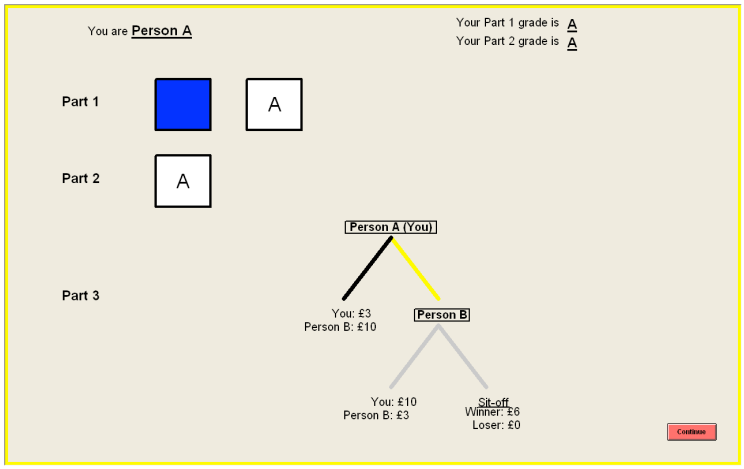
The diagram on-screen (below the cards) is where you make your decision to 'Ignore' or 'Interact' with Person B. As Person A you are asked to choose between one of the two upper branches. Clicking on a branch selects it and turns it yellow.

Figure 4. The screen you will see as Person A.



Once the cards have been revealed and a decision has been made, the 'Continue' button will appear (shown in Figure 5).

Figure 5. An example of the screen you will see as Person A once you have revealed the cards and made a decision.



After clicking the 'Continue' button two questions will appear on-screen that ask you about your decision. Once you have completed these, click 'Continue' again and you will move to the next screen.

Other Information
You will play the role of Person B once and make a single decision.

You will play the role of Person A six times. Each time you will make a single decision in response to a different Person B who will have achieved different combinations of grades in Part 1 and Part 2. Every decision corresponds to a different subject in this experiment playing as a Person B, so you will never meet the same subject twice.

Each screen will display your current role (whether you are Person A or Person B for this round) and remind you of the wall sit grades that you achieved in Part 1 and Part 2.

Part 3 Payment
During the experiment, as you know, you will be asked how to play against a series of different people. Each interaction between you and one of these different people may yield a different outcome (which is determined by what you decide and what the other person decides) that will change how much money you can potentially earn.

To decide how much money you actually earn, we will randomly choose one of these outcomes from your interactions and you will be paid the resulting outcome directly or you will engage in a sit-off and then receive your payment.

Thus, your action and the action of the other person in each interaction may determine your final earnings, but neither you nor we will know which interaction your payment is based on until the end.

Summary
In Part 2 you will do the wall sit exercise (without being paid).

In Part 3 you will play one round as a Person B and then six rounds as a Person A.

As a Person B:
- The grade you achieved in Part 1 will be sent with some imprecision to Person A in the form of two blue cards.
- The precise grade you achieved in Part 2 will be sent to Person A in the form of a single red card.
- You will select one of the two actions 'Yield' or 'Retain' using the branches in the diagram.

As a Person A you will:
- See two blue cards which cover, with some error, Person B's wall sit grade from Part 1. If he received an A grade both cards cover 'A', if he received a B grade one

card covers 'A' and the other covers 'C', and if he received a C grade both cards cover 'C'.
- See a single red card which covers Person B's precise grade from Part 2.
- Select one of the two actions 'Interact' or 'Ignore' using the branches in the diagram.

You will all come back on Friday the 6th to take part in a second 20 minute session.

You will now be given some questions to check your understanding and the answers to these questions. After this you will proceed to Part 2 where you will wall sit. Then you will return to the computer room and participate in Part 3 where you will first make your choice as Person B and then your choice as Person A in six situations.

Ethical approval

Nuffield
Centre for Experimental
Social Sciences

24 May 2014

To whom it may concern,

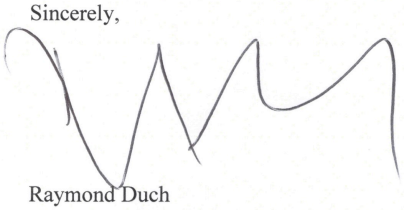
RE: "The Interaction between Signs and Signals of Toughness"

We note that the above entitled experiment has been reviewed by the members and associates of the Centre for Experimental Social Sciences (CESS), Nuffield College, University of Oxford, and received a favourable ethical opinions. The experimental protocols is essentially the same as the one successfully implemented in the CESS lab in December 2013, and we do not consider that the modification of the protocol, the use of sleeping mask, would provoke serious ethical problems.

However, please note that there are several points raised by the reviewers. The experimental protocols have to conform to the following suggestions to satisfy the ethical standard set by the CESS:

- Some of the protocols of this experiment can be obtrusive to subjects. In particular, involvement of physical activities has to be clearly stated in the recruitment, and the purpose of using a sleeping mask has to be explained to the subjects during the experimental instruction.
- When giving the experiment instructions, the experimenter has to explicitly state that the participation of this experiment is absolutely voluntary and subjects can leave the experiment at any moment having the show-up fee paid. Also, the experimenter may want to consider the use of blacked-out sunglasses instead of sleeping masks.

Sincerely,



Raymond Duch

Director
Nuffield Centre for Experimental Social Sciences