Social Science Research 86 (2020) 102393

Contents lists available at ScienceDirect

Social Science Research

journal homepage: http://www.elsevier.com/locate/ssresearch

How dominance hierarchies emerge from conflict: A game theoretic model and experimental evidence

Wojtek Przepiorka [a,*], Charlotte Rutten [a], Vincent Buskens [a], Aron Szekely [b,c]

[a] Department of Sociology/ICS, Utrecht University, Netherlands
[b] Collegio Carlo Alberto, Università di Torino, Italy
[c] Institute of Cognitive Sciences and Technologies, CNR, Italy

ARTICLE INFO

Keywords:
Conflict
Hierarchy formation
Reputation
Social rank

ABSTRACT

We develop a game theoretic model of conflict and empirically test its predictions to study the emergence of social hierarchies in small groups. Previous research shows uncertainty about actors' ability may lead to more conflict; conflict demonstrates actors' ability and establishes relationships of dominance and submissiveness. Since we assume uncertainty regarding ability to be a crucial cause of conflict, we focus on the effects of different information conditions. We posit that actors know the distribution of abilities in their group and vary whether or not they know (1) their own ability and (2) their interaction partners' interaction histories. Our results from a laboratory experiment closely match qualitative model predictions. Most importantly, conflict produces information about actors' ability, which reduces subsequent conflict. In an exploratory analysis we investigate to what extent gender, social value orientation, risk preferences and a competitive personality account for the quantitative discrepancies between model predictions and subject behavior.

1. Introduction

Hierarchies, which we define as social structures that reflect a perceived ranking of people's attributes relative to others, are universal features of human societies that influence expectations and behaviours of the people intertwined in those hierarchies (Chase and Lindquist, 2009; Magee and Galinsky, 2008). In contrast to formal hierarchies that are backed by rules and enforced by agents in power, we study the emergence of *informal hierarchies* that operate without these. Our focus is on how hierarchies emerge *bottom-up* from social interactions, rather than by top-down design. The other important delimiting feature of the hierarchies we study is that they emerge from *competition* between individuals over scarce resources, rather than as a means to enhance coordination and reduce transaction costs in endeavours requiring the cooperation of many individuals (see Simpson et al., 2012).

Hierarchies that emerge through competition can be identified in myriad social settings. Colleagues in academia contend for promotion and scarce jobs with papers, critiques, and talks, creating prestige rankings; school children vie for friends and social status; athletes compete directly for victory to progress in the ability hierarchy and gain sponsorships and greater prize-money. But perhaps the most extreme, and thus fascinating, setting in which informal conflict-based hierarchies form is in prisons.

Prisoners are, for much of the time, unobserved by prison guards, allowing unconstrained actions to dominate, and since prisoners also typically interact repeatedly, bottom-up social orders can emerge. Differently to most interactions in the legal world, the

* Corresponding author.
E-mail address: w.przepiorka@uu.nl (W. Przepiorka).

https://doi.org/10.1016/j.ssresearch.2019.102393
Received 21 January 2019; Received in revised form 11 October 2019; Accepted 20 November 2019
Available online 26 November 2019
0049-089X/

incentives that prisoners face are much more intense because of two factors: resources are extremely scarce and many of the other prisoners are willing to use violence to obtain their ends (Gambetta, 2009: ch. 4; Sykes, 1958). Taken together, these features make prisons an illuminating microcosm in which to study the emergence of informal dominance hierarchies.

In a seminal chapter analysing hierarchy formation in prisons, Gambetta (2009) argues that the availability of credible information about one's ability and willingness to use violence and to receive it ('violence potential'), is an important determinant of the fights that occur during hierarchy formation. Specifically, he argues and finds evidence for the idea that conflict and credible information about violence potential are negatively related. The clearer it is that an inmate is tough or not, the less need there is for that inmate to fight. And, that conflict between inmates may create credible information about violence potential that can later allow conflict to be resolved without violence. While persuasive, Gambetta (2009) uses only observational data to test his conjecture and thus cannot rule out important confounds. We avoid the issue of confounds here as we study experimentally the formation of informal hierarchies in a situation of conflict.

In our experiment, subjects compete over scarce resources in the Hierarchy Contest Game (HCG), a modified version of the Hawk-Dove game, to test the conjecture that it is the lack of information that leads to conflict through which hierarchies are formed. By design, our experiment eliminates possibly confounding factors allowing us to cleanly test whether the informational predictions hold. Our predictions are derived from a game theoretic model that we devised for the purpose.

With our experiment, we test the effect of two different types of information. Information that people obtain about their *own competitive abilities* and the information that they obtain about *other's abilities*. While knowledge of others' competitive abilities can come from many sources, we focus on the information that is contained within others' history of actions: whether the outcome of conflict was a win or a loss. This represents a simple reputation system. The knowledge that people have of their own abilities, meanwhile, is based on experience. Manipulating these two types of information allows us to explore our core research question: *How do different information conditions affect the amount of conflict involved in the emergence of informal hierarchical social structures?*

Intuitive reasoning, as well as some literature, suggests an alternative prediction (Benard, 2013, 2015). It could be that when people have the opportunity to create and build reputations about their abilities, they have increased incentives to compete. By competing, they reduce possible challenges in the future. Conversely, not competing could be perceived as a sign of weakness that makes them a target for contenders later on. In this case the addition of reliable information may increase conflict as no-one wants to back down and incur the latter cost due to a poor reputation.

We argue based on our model that this is not necessarily the case. Including information, both about own and about other's abilities reduces conflict. We also present experimental evidence showing that the information that is generated by conflict reduces uncertainty about others' abilities. This leads to a reduction in conflict and the emergence of hierarchical social structures, because actors can better assess the outcomes of future encounters and therefore behave more and more according to their relative abilities.

We use a 2×2 factorial design in which we vary whether or not subjects know their own ability and whether or not they are shown information about past action outcomes of their opponents. Specifically, in the four experimental conditions, subjects:

1. Do not know their own abilities nor their opponents' histories (*No Info*);
2. Know their own abilities but not their opponents' histories (*Own Ability*);
3. Do not know their own abilities but learn their opponents' histories (*Opponent History*);
4. Know their own abilities and learn their opponents' histories (*Own Ability* + *Opponent History*).

Each experimental condition captures a core type of interaction that occurs during competition and hierarchy formation. The *No Info* condition, for instance, represents the setting and informational constraints that are experienced by a new class of children entering school together, army recruits starting basic training, or newly convicted adolescents entering juvenile detention. At first, they have no knowledge about their own abilities nor do they know anything about the abilities of other prisoners whom they encounter in the institution. Conversely, the *Own Ability* + *Opponent History* treatment models cases in which a veteran criminal, with a history of violent confrontations and incarceration, enters an institution with prisoners whom they get to know.

Competitive or fighting ability in our experiment is exogenously imposed: some people are stronger than others allowing them to win conflicts with a higher probability. Hence, ability is an *individual trait*, and does not as such constitute a social structure. However, through dyadic conflict interactions, these traits are 'unveiled' and enable individuals to establish a hierarchical social structure at the group level. In other words, ability is an individual-level independent variable, conflict is the dyad-level mediating variable, and hierarchy is the group-level outcome variable that represents the full set of dominance relations in a group (Chase and Lindquist, 2009; Chase and Seitz, 2011).

Our work also relates to an important line of literature in sociology that explores the social construction of hierarchies. One position contends that hierarchy formation is primarily driven by arbitrary social processes that disassociate people's underlying ability from their position in a hierarchy (Chase and Lindquist, 2009; Podolny and Lynn, 2009). Thus, the association between ability and position is at best loose. We call this the 'arbitrary hierarchy' argument. Conversely, the 'ability hierarchy' perspective holds that ability and position in a hierarchy are tightly linked: people higher up in status, dominance, or wealth, are there largely because of their underlying attributes (e.g. Blau, 1964; Homans, 1961). Agent-based simulations demonstrate that both can, under different conditions, be true (Lynn et al., 2009; Manzo and Baldassarri, 2015). While we do not directly test these two competing positions, we can and do use our experiment to explore whether a tight or loose link emerges between ability and position depending on our information conditions. If a tight link arises, then this suggests that ability-based hierarchies can form. This does not rule out the possibility that in other settings and setups arbitrary hierarchies dominate.

2. Previous research

This paper primarily relates to four topics: conflict, hierarchies, reputation, and costly signalling.[1] Despite the rich literatures on each of these areas, surprisingly little work explores the particular intersection at which our question sits (although see Silverman, 2004). Most research on hierarchy formation (e.g. Boehm, 1999; Magee and Galinsky, 2008) and conflict (e.g. Chambers and De Dreu, 2014; De Dreu et al., 2016; DeScioli and Wilson, 2011; Edgar and Martin, 2001) does not consider the role of information, and the literatures on reputation (e.g. Milinski, 2016; Przepiorka et al., 2017) and costly signalling (e.g. Bolle and Kaehler, 2007; Kübler et al., 2008; Przepiorka and Diekmann, 2013) do not study conflict, instead focusing their investigations on cooperation and trust.

Nevertheless, there are a handful of closely related studies. Most relevant is a recent experiment that was conducted in parallel to ours (Szekely and Gambetta, 2019). Similarly to our work, their experiment tests the effect that information has on aggression and conflict and they find that credible signs and signals of toughness reduce both. Yet their paper differs from ours in two important, albeit complementary, ways. First, we study how hierarchies emerge from *multiple interactions* within the same groups and how conflict today can be used as to deter conflict tomorrow. Conversely, Szekely and Gambetta study *one-shot interactions* and whether reliable signals generated before an interaction entailing conflict can affect the outcomes during the interaction—think tattoos and scars when entering prisons compared to fighting inside prison. Second, in our experiment we consider situations in which all participants have the *same amount of information* about each other. In contrast, Szekely and Gambetta study conditions in which competitors have *different amounts of information* in that one competitor knows the signs and signal of the other but not *vice versa*.

Also related is a series of experiments by Benard (2013, 2015) in which he tests whether multiple variants of reputation systems increase or decrease aggressive acts. In contrast to our predictions, he consistently demonstrates that information increases aggressive acts. However, because one side of the interaction is played by a pre-programmed computer, he cannot include the dynamic elements of reputation formation—being aggressive now to reduce conflict in the long term, and for the same reason he cannot test whether the increased aggression ultimately leads to more conflict. An additional difference is that Benard limits reputation to the transmission of action: whether a person decided to behave aggressively or not, but does not include the corresponding outcome. We transmit both the action and outcome as part of reputation formation in the relevant treatments.

Concerning observational studies—in addition to the work of Gambetta (2009) that we already mention—is done by Gould (2003) who explores the puzzle of why trivial snubs often escalate into serious violence. He argues that status concerns are a key motivation and that status concerns are most unclear between people of similar status, which he calls 'symmetric relations'. Consequently, Gould predicts that conflict arises most frequently among people with similar social rank, and he presents homicide data from America and India that are consistent with this prediction. Unlike Gould, we exclude the possibility of equal status by design (see Experimental design) and test whether it is the lack of information about social rank that leads to escalate conflict. Additionally, in a rich and analytical ethnography of life in a Polish prison, Marek Kaminski (2004:101–29) touches on the relationship between information and conflict but provides no systematic empirical evidence.

Our work also, but more loosely, relates to the abundant literature on economic contests (Dixit, 1987; O'Keeffe et al., 1984). A contest can be defined as a game 'in which players are able to expend scarce resources (such as money, time or effort) in order to affect the probabilities of winning prizes, the values of which are ranked identically by the players (but may not be identical in absolute terms). The distinguishing characteristic of a contest is the fact that a higher expenditure of the scarce resource(s) has a nonnegative (and sometimes strictly positive) effect on the probability of winning the more valuable prizes' (Dechenaux et al., 2015:613). While our conflict interaction fits this definition, it is distinct from the canonical models employed in the literature—Tullock contests, all-pay auctions, and rank-order tournament (Dechenaux et al., 2015)—in important ways. In our setup, and unlike in the standard models, players who challenge for a resource *only incur a cost if their opponent also challenges* but *when both players challenge they both incur a cost* irrespective of who wins and who loses. These features capture key elements of competitive interactions: unchallenged competitors automatically win the prize, while if a competition materialises then both the eventual winner and loser expend the energy/resource in competing. Another feature distinguishes our game from the economic contests literature: the prize is shared if neither player competes for it, which allows a more cooperative compromise. Nonetheless, our setup can be considered as a specific kind of economic contest that has not been studied before in this literature.

Biologists have extensively studied how information, transmitted through reliable signals such as the dominant frequency of frog calls, badges of status in birds, and weapon displays in crustaceans (Searcy and Nowicki, 2005:134–80) affects conspecific conflict among non-human animals (see also Maynard Smith and Harper, 2003). Additionally, evolutionary game theoretic models demonstrate how reliable signals that solve situations of conflict can evolve and remain stable (Enquist, 1985; Grafen, 1990). Yet even in this field, which has a history of using game theory to understand conflict, observational data dominates and experiments manipulating information availability are rare.

Finally, there is a broad literature in political science that uses the same line of thinking to understand war. Fearon's (1995) key paper argues that war emerges because states have private information about their own capacities that they cannot reliably communicate. Reliable information could solve the issue and allow the states to come to a mutually beneficial solution (see also Powell, 2004; Carrillo and Palfrey, 2009). The key difference between this literature and our work is the level of the agent: for us they

[1] Inherent in creation of an intimidating reputation is the requirement that the information generated is credible. Creating a history of challenging and winning is hard-to-fake for individuals lacking competitive ability but affordable to people with high competitive ability. In this way reputation is closely linked to costly signalling theory (Gambetta, 2009; Spence, 1973, 1974). Nevertheless, there are some important differences between reputation and signalling (Przepiorka and Berger, 2017).

are individuals while in this literature it is supra-individual entities, typically states. Whether individuals behave as expected by the theory is still an open question.

Our main contribution to this literature is twofold. (1) We present a novel game theoretic model that clarifies the relationship between conflict, reputation formation, and the emergence of informal hierarchies. The game theoretic model, which we describe in the next section and in the Appendix A, allows us to derive precise predictions about human behaviour under varying information conditions. (2) We design an experiment that allows us to test our model predictions and test these predictions with human subjects in a computerised laboratory experiment without using deception. Both our game theoretic model and experimental evidence allows us to gain a deeper understanding about how humans interact in conflict situations and how information about others' and own abilities influence competition and the emergence of dominance hierarchies.

3. Theory

Assume a group of N players who differ on a relevant characteristic A (for instance ability) such that $A_i > A_j$ for any pair of players i and j. That is, in the Hierarchy Contest Game (HCG), there is always a stronger player i and a weaker player j. We start with the assumption that players are rational, self-regarding, risk neutral and know the distribution of A in their group.

The HCG, a generalised version of the hawk-dove game, describes the situation in which two opponents can compete for a scarce resource (Fig. 1). In the HCG, two players, one stronger and one weaker, simultaneously choose to *challenge* (Ch) or *back down* (Bd). If both players choose Ch, a conflict occurs in which both incur c costs, the stronger player receives w winnings, and the weaker player receives nothing. If both choose Bd, each gets $0.5w$. If one player chooses Ch and the other Bd, the one who chooses Ch gets w while the other gets nothing, irrespective of ability. In the last two cases no conflict occurs and hence neither player incurs a cost. The payoffs are ordered as follows: $w > c > 0.5w > 0$ (see Carrillo and Palfrey, 2009 for a related game).

The Nash equilibrium of this game is *the strong player challenges and the weak player backs down*. This is based on the assumption that players know their own ability and their interaction partner's ability. In other words, in case of complete information, there is no conflict and the strong players 'exploit' the weak. In what follows, we analyse the HCG under different information conditions. We first look at the case where two players from a group of N players are randomly matched to interact only once and either have no information about their own ability and the ability of their interaction partner or know their own ability but do not know their interaction partner's ability. Then we consider the case in which players of a group of N players are repeatedly randomly matched with each other and, if they lack the information, can learn their own and their interaction partners' abilities over time. Based on these theoretical considerations and the parameter values we chose for our experiment, we derive our model predictions and behavioural hypotheses.

3.1. Information in the one-shot HCG

Start with the HCG in which players do not know their own nor other's abilities. In this case, the HCG becomes a chicken game (Rapoport and Chammah, 1966), which has two pure-strategy Nash equilibria [(Ch, Bd) and (Bd, Ch)] and a mixed strategy equilibrium (MSE) in which both players choose Ch with probability p_i^* such that $p_i^* = 0.5w/c$ for all i. This is because the probability of being the stronger player in the interaction is 0.5 and players' ability only matters if both choose Ch. In this case, solely the payoffs to conflict (the upper-left cell in Fig. 1) are altered to $0.5w - c$ for both players (see Appendix A).

If players know their own abilities but not the abilities of others, there is only one pure strategy Nash equilibrium: the strongest $N - \lfloor \tau_u \rfloor$ players choose Ch and the others Bd, where $\tau_u = [c(N-1) + w]/w$. So, the challenging probability in equilibrium is $p_i^* = 1$ for all i with $A_i > \tau_u$, and $p_j^* = 0$ for all j with $A_j < \tau_u$. In other words, choose Ch if you are stronger than the threshold τ_u and Bd otherwise.[2]

3.2. Information in the repeated HCG

We now analyse the effect of four exogenously assigned information conditions on a group of N actors from which pairs are randomly drawn to determine the two opponents in consecutive interactions. These four conditions correspond to our experimental treatments. These are *No Info*, *Own Ability*, *Opponent History*, and *Own Ability + Opponent History*. We assume self-regarding and risk neutral agents who are backward-looking. These agents use their knowledge, stemming from previous interactions, to decide whether to choose Ch or Bd in an interaction. Their strategic considerations do not take potential outcomes in future interactions into account implying that each interaction is perceived as a one-off encounter under changing information (within a particular information condition). This assumption is in line with experimental evidence showing that experimental subjects' strategic rationality in repeated games may be limited (Buskens, Raub, and van der Veer, 2010, also see footnote 3). Note, however, that actors obtain information through learning if they challenge each other; without mutual challenging no information is generated and no learning takes place.

3.3. Model predictions and hypotheses

In each of the four information conditions we consider two limit cases for which we derive predictions: subjects' first interactions at $t = 0$ and their interactions after an extended number of repetitions, at $t = T$.

[2] This holds for groups that are not too large (see Appendix A).

		Weaker player: Ch	Weaker player: Bd
Stronger player	Ch	$w-c$ / $-c$	w / 0
	Bd	0 / w	$0.5w$ / $0.5w$

Fig. 1. Hierarchy contest game (HCG).

In the *No Info* condition, at $t = 0$, we expect subjects to behave according to the MSE in the one-shot HCG (without information). This is because they do not know anything about themselves nor about others. However, over time subjects are able to discover their own abilities and so at $t = T$ they behave as players in the one-shot HCG who know their own abilities.

In the *Own Ability* condition, at $t = 0$, we expect subjects to behave according to the one-shot HCG in which they know their own ability, because they know their own ability from the start. Since they have no chance of observing others' histories and learn about others' abilities, they behave as in the one-shot HCG with information about own ability, even at $t = T$. In other words, in the *Own Ability* condition, subjects' behaviour will not change over time.

In the *Opponent History* condition, at $t = 0$, we expect subjects to behave like in the *No Info* condition at $t = 0$, because they do not know their own abilities. Over time, they have the chance to test themselves, and figure out their own abilities, and, during these interactions they observe information about the abilities of others. We therefore expect subjects to behave according to the predictions of the one-shot HCG with full information at $t = T$.

Finally, in the *Own Ability* + *Opponent History* condition, we expect subjects to behave like in the *Own Ability* condition at $t = 0$, because these subjects know their own abilities from the start, but have no knowledge of others' abilities. However, following sufficient interaction with others, the abilities of others' also become apparent. We therefore expect subjects to behave as in the full information one-shot HCG at $t = T$.[3]

If players know both their own and others' abilities, the game corresponds to the HCG in Fig. 1 and the sole pure strategy Nash equilibrium is that the stronger player chooses Ch and the weaker player Bd. Player i's challenging probability is $p_i^* = \frac{A_i - 1}{N-1}$. Like experienced prisoners who know each other, they face little incentive to fight: both know who would win and who would lose if a conflict were to occur.

Using the parameters from our experiment (see Methods) of $n = 4$, $w = 5$ and $c = 3$, we make the following predictions for the repeated HCGs. We derive individual-level predictions about actors' challenging probabilities (p_i) in the four information conditions at both $t = 0$ and $t = T$. Based on these challenging probabilities, we can make group-level predictions about the challenging rate $p_{\mathrm{Ch}} = \frac{1}{N}\sum_{i=1}^{N} p_i$, fighting or conflict rate $p_{\mathrm{Fi}} = \frac{2}{N(N-1)}\sum_{i}^{N}\sum_{j>i}^{N} p_i p_j$ (the proportion of interactions in which both actors choose Ch), sharing rate $p_{\mathrm{Sh}} = \frac{2}{N(N-1)}\sum_{i}^{N}\sum_{j>i}^{N}(1-p_i)\left(1-p_j\right)$ (the proportion of interactions in which both actors choose Bd) and exploitation rate $p_{EX} = 1 - p_{Fi} - p_{Sh}$ (the proportion of interactions in which one actor choose Ch and the other chooses Bd). Table 1 presents these predictions. Drawing on our model predictions, we derive six behavioural hypotheses.

When subjects do not know their own ability (*No Info* and *Opponent History*), they all challenge at the same probability for they cannot differentiate how likely they are to prevail in conflicts based on their abilities. They have to estimate this probability based on the distribution of abilities. Conversely, when subjects know their own ability (*Own Ability* and *Own Ability* + *Opponent History*) they can condition their challenging according to their ability from the start.

H1. At $t = 0$, subjects of all ability challenge equally in *No Info* and *Opponent History* while higher ability subjects challenge more in *Own Ability* and *Own Ability* + *Opponent History* than lower ability subjects.

In all treatments, except for *Own Ability*, subjects learn about their relative ability as the experiment progresses, so higher and lower ability subjects adjust their beliefs and challenging accordingly.

H2a. At $t = T$, higher ability subjects challenge more than lower ability subjects in all treatments.

If subjects know their opponent's history (*Opponent History* and *Own Ability* + *Opponent History*) ability 2 and 3 subjects challenge at intermediate rates while if they do not (*No Info* and *Own Ability*) then they either never challenge or always challenge. This is because when subjects know the opponents' history, intermediate subjects adjust their expectations of winning accurately and only challenge those whom they will beat while without this information they are forced to challenge or concede to everyone.

[3] By assuming backward looking agents we circumvent the full-fledged game theoretic analysis of the repeated game. This is a limitation in the conditions in which agents can see the opponent's history because in these conditions forward looking agents might have an incentive to establish a reputation. Although such an extension of our model would be worthwhile pursuing, it will hardly change our predictions. First, realise that in the *Own Ability* + *Opponent History* condition, low ability agents have no opportunity to build a reputation for being high ability given that eventually they will be exposed as low ability by high ability agents, who challenge from the start. In the *Opponent History* condition, agents do not know their ability from the start and thus will be uncertain what the value of building a reputation might be. In this case it seems plausible to expect agents to behave according to the MSE in the one-shot HCG (without information).

Table 1
Experimental conditions and predictions.

Treatment conditions	No information		Own ability		Opponent history		Own ability + opponent history	
Time	$t = 0$	$t = T$	$t = 0$	$t = T$	$t = 0$	$t = T$	$t = 0$	$t = T$
Individual level predictions of challenge probabilities								
$p_1 = \Pr(a_i = \text{Ch} \mid A_i = 1)$	.83	0	0	0	.83	0	0	0
$p_2 = \Pr(a_i = \text{Ch} \mid A_i = 2)$	.83	0	0	0	.83	.33	0	.33
$p_3 = \Pr(a_i = \text{Ch} \mid A_i = 3)$	.83	1	1	1	.83	.66	1	.66
$p_4 = \Pr(a_i = \text{Ch} \mid A_i = 4)$	.83	1	1	1	.83	1	1	1
Group level predictions of challenge, conflict, sharing and exploitation rates								
$p_{\text{Ch}} = \Pr(a_i = \text{Ch})$	.83	.5	.5	.5	.83	.5	.5	.5
$p_{\text{Fi}} = \Pr(a_i = \text{Ch} \,\&\, a_j = \text{Ch})$	.69	.17	.17	.17	.69	0	.17	0
$p_{\text{Sh}} = \Pr(a_i = \text{Bd} \,\&\, a_j = \text{Bd})$	.03	.17	.17	.17	.03	0	.17	0
$p_{\text{Ex}} = \Pr(a_i = \text{Ch} \,\&\, a_j = \text{Bd})$	.28	.66	.66	.66	.28	1	.66	1

H2b. At $t = T$ intermediate ability subjects challenge at intermediate rates in *Opponent History* and *Own Ability + Opponent History* while intermediate ability subjects challenge at extreme rates in *No Info* and *Own Ability*.

Conflict, we argue, is associated with information negatively such that more information leads to less conflict. There are two ways in which this can occur in our experiment: between the different treatments and within each treatment over time.

H3a. At t = 0, conflict is highest in No Info and Opponent History and lowest in Own Ability and Own Ability + Opponent History.

H3b. At $t = T$, conflict is lower than at $t = 0$ in all experimental conditions except for *Own Ability*.

Hypothesis H3b holds in all treatments except *Own Ability*, because in the *Own Ability* treatment no additional information is generated over time. However, the decrease in conflict predicted by H3a and H3b is driven by an increase in exploitation and not an increase in sharing. Thus, when conflict is reduced, exploitation is high.

H4. Exploitation is highest when conflict is lowest.

4. Methods

4.1. Experimental design

Subjects play four variants of the repeated HCG that vary in terms of the information they receive. These four conditions correspond the information conditions described in the theory section. We test the model predictions using a 2 × 2 factorial design that varies whether or not subjects know their own ability and whether or not they know their interaction partner's history (Table 1). The experimental conditions are varied within-subject, so every subject participates in every condition. To control for order and learning effects we systematically vary the order of the conditions between sessions.

Each of the four conditions consists of 15 periods (i.e. $T = 15$).[4] Before the first period in each condition, we randomly divide subjects into groups of $n = 4$ and randomly assign each of them an ability level between 1 and 4, such that there is one of each ability level in each group. Subjects keep their ability level for the condition. A higher number means that a subject is better in competing, so the subject with ability 1 is the worst and the subject with ability 4 is the best and subjects are made explicitly aware of this. In every period, subjects are randomly paired with one of the other subjects in their group and are told that they interact with that other subject in a situation in which they simultaneously choose between action A and action B, whereby the payoff to their choice depends on their own choice, the other person's choice and their number. These combined action choices and the resulting payoffs come from the Hierarchy Contest Game (Fig. 1). The decision situation is presented to the subjects in a table (the complete instructions are provided in the online appendix). Choosing A is equivalent to challenging (Ch), whereas choosing B is equivalent to backing down (Bd). If both interacting subjects choose A, a conflict emerges and the subject with the higher number earns €0.20, whereas the subject with the lower number loses €0.30. If both subjects choose B, which means that they divide the resource, they earn €0.25 each. And if only one of the two subjects chooses A, then this subject earns the full €0.50, whereas the other subject receives nothing. Thus, they compete over €0.50. We use neutral wording throughout the experiment to avoid framing effects.

In conditions in which subjects know their own ability from the start, there is a sentence on the subjects' screen which shows them their ability. Furthermore, in conditions in which subjects know their interaction partners' history, there is a table on subjects' screens, which states 'The other person chose … ', followed by the number of times the current interaction partner chose A and B in the previous periods and the number of times this interaction partner generated different payoffs. In other words, the table allows subjects to infer how many times their opponent had chosen Ch in the past and how many times this resulted in a conflict that the opponent won. We

[4] Results from a pilot session suggested that 15 periods are sufficient for subjects to reach a behavioral steady state.

consider knowledge of both—fighting and whether a win or loss occurred—as realistic and important components of reputation formation.

To ensure that differences in memory do not affect subjects' behaviour, there is always a table on the subjects' screen, which informs them about their own previous actions and payoffs. Moreover, if both interacting subjects choose A, we explain why they received the payoffs from the competition; 'This is because your number is [higher/lower] than the number of the other person.' By indicating this specifically, we ensure that subjects realise the information that is generated about their own number from the competition.

4.2. Experimental procedure

We programmed our experiment in z-Tree (Fischbacher, 2007) and conducted it in the Experimental Laboratory for Sociology and Economics of Utrecht University. Via the web based ORSEE recruitment system (Greiner, 2015) we invited Utrecht University students and graduates of a wide variety of disciplines and nationalities to participate. In total we ran eight experimental sessions with 128 subjects meaning that we collected data on 32 groups of $n = 4$ per treatment.

Before the start of a session, subjects were randomly assigned to a computer in the laboratory and received general instructions in English on paper (see online appendix). They received additional treatment-specific instructions before the start of each condition. Instructions were the same for all subjects. To ensure that subjects understood the conditions, we let them answer control questions and we explained the correct answers before the start of each condition. Sessions lasted approximately one and a half hours.

4.3. Measures

To study the effects of ability and the experimental conditions on the initial and final challenging rates, we consider the action choices of individual subjects that are made in the *first three* ('initial') and the *last three* ('final') periods of the different conditions. Likewise, we study the effects of experimental conditions on the initial and final competing and dividing rates by considering the combined action choices of interaction partners, where both subjects choosing action A is considered as conflict and both actors choosing action B is considered as sharing. We consider the first three periods to be the early stage of (potential) learning and the last three periods to be the 'equilibrium' stage of our game. By analysing the data of more periods we reduce the likelihood of biased results due to outliers and potential end-game effects.

We infer the emergence of hierarchies based on both observed behaviours and consequent outcomes. In particular, we consider that an ability-based hierarchy is established when the challenging rate is monotonically increasing with ability at time $t = T$ (see Table 1 and hypotheses H2a and H2b), and when the conflict rate is low in dyadic interactions (hypotheses H3a and H3b). The former criterion is necessary for a dominance relationship to form: there should be consistent aggressors and consistently deferential individuals, and, this should correspond to ability. The latter criterion is necessary to say that aggression and submission are well-targeted: higher ability individuals should be aggressive towards lower ability individuals but not to other high ability ones and low ability individuals should be submissive towards higher ability individuals but not to other low ability ones. Together, these imply that an ability-based hierarchy exists in a group.

Still, we anticipate that a non-negligible proportion of subjects will not behave according to equilibrium predictions. To understand why subjects might deviate, we elicit a range of other measures. We elicit their risk and social preferences, and, in a post-experimental questionnaire, we obtain information on their competitiveness, degree of dominance personality, and demographic information.

We elicit subjects' risk preferences using an adapted version of Binswanger's (1980) measure in which subjects use their €2.00 show-up fee to decide between various lotteries the outcome of which was based on a virtual coin flip. These lotteries, of which there were five in total, formed a risk preference continuum. Choosing the lottery with a certain outcome of €2.00 is considered as risk averse and choosing the lottery with a fifty-fifty outcome of €0.00 or €4.00 reflects the highest possible risk taking.

To determine subjects' social preferences, we used a z-Tree implementation (Crosetto et al., 2019) of the Social Value Orientation (SVO) Slider Measure (Murphy et al., 2011). Subjects are presented with a series of six dictator games in which they allocate money to themselves and a random and anonymous other subject in the experiment. Since the payoff combinations vary also in efficiency, the extent to which subjects care about the outcomes of others can be quantified.

To determine how competitive subjects are we used the competition items of Helmreich and Spence (1978). Subjects are asked questions such as 'Do you enjoy working in situations involving competition with others?' and 'Do you feel that winning is important in both work and games?', followed by five-point Likert items. We consider the answers 'Don't know/can't say' as missing. This has led to a five-item competitiveness scale with a reliability of $\alpha = 0.76$.

The degree to which subjects have a dominant personality was based on the work of Ray (1981) and assessed by asking them questions such as 'Do you tend to boss people around?' and 'Are you easily swayed by other people's opinions?', followed by five-point Likert items. Recoding was done such that a higher value reflects a more dominant personality and the answer 'Don't know/can't say' is considered as missing. By analysing the respective data via a principle component analysis, we generated a seventeen-item dominance scale with a reliability of $\alpha = 0.88$.

Out of our 128 subjects, 60.9% (78/128) were female and their ages ranged from 17 to 60 (M = 24.13, SD = 6.17). Since subjects made 15 decisions in each treatment, we have data on 1920 actions and 960 outcomes (one per dyad) in each condition and thus a total of 7680 actions and 3840 outcomes (see Table B in Appendix B for descriptive statistics).

5. Results

We first give a detailed account of the individual-level results testing hypotheses H1, H2a and H2b and then do the same for group-level outcomes testing hypotheses H3a, H3b and H4. The last part of the results section consists of an exploratory analysis that aims at explaining the discrepancy between our model predictions and subject behaviour.

All test statistics are based on regression model estimations. We only use logistic regression models because all our target variables are binary (Long, 1997). Statistical significance is set at the 5% level (i.e. $\alpha = 0.05$) for two-sided tests and we estimate coefficients with robust standard errors accounting for clustering, where every group of four subjects in a particular experimental condition forms a cluster (Fitzmaurice et al., 2004; Snijders and Bosker, 2012). We use Stata's *margins* command to calculate proportions from saturated logit models and test the statistical significance of the differences between proportions using Wald tests of linear hypotheses or Bonferroni corrected pairwise comparison tests.

5.1. Individual-level actions

Fig. 2 shows the (individual-level) challenge rates over the course of 15 rounds and across ability types and our four experimental conditions. The results are mostly in line with H1, H2a and H2b: (1) Concerning H1, there are hardly any differences in challenge rates across ability types within conditions *No Info* and *Opponent History* at $t = 0$ (Wald tests of equality of challenge rates: $\chi^2_{(3)} = 5.71$, $p = 0.127$ and $\chi^2_{(3)} = 6.26$, $p = 0.100$, respectively), whereas in *Own Ability* and *Own Ability + Opponent History*, high ability subjects challenge substantially more than low ability subjects ($\chi^2_{(3)} = 362.45, p < 0.001$ and $\chi^2_{(3)} = 498.05, p < 0.001$, respectively). (2) In line with H2a, at $t = T$, the higher subjects' ability, the higher are their challenge rates. Bonferroni corrected pairwise comparison tests (comparing challenge rates across abilities within conditions) produce only four statistically insignificant differences: between abilities 1 and 2 in *No Info*, between abilities 1 and 2 in *Own Ability*, and between abilities 1 and 2 and 2 and 3 in *Opponent History*. The other 22 differences are statistically significant and only the difference between abilities 1 and 2 in *No Info* is negative albeit small. (3) We find only partial support for H2b. In line with H2b, the difference in challenge rates between abilities 2 and 3 is substantial and statistically significant in *No Info* and *Own Ability* ($z = 4.44$, $p < 0.001$ and $z = 6.35$, $p < 0.001$, respectively). Contrary to our expectations, in the *Own Ability + Opponent History* condition, the difference between abilities 2 and 3 is also substantial and statistically significant ($z = 4.25, p < 0.001$). We expected the two intermediate abilities to be closer to each other at an intermediate level of challenging, but this is only the case in *Opponent History*, where the difference is small and statistically insignificant ($z = 1.35$, $p = 0.176$).

These results corroborate H1, H2a, and H2b. If information about ability is lacking, subjects challenge equally, irrespective of ability. However, the more information about ability is available (between conditions) or becomes available over time (within conditions), the more challenging is positively related with ability; higher ability subjects challenge more than lower ability subjects. But does, as we claim, more information lead to less conflict? We address this question next.

5.2. Group-level outcomes

We next focus on conflict and exploitation as these are the two group-level outcomes that are relevant for our remaining three hypotheses. Results concerning sharing can be inferred from the conflict and exploitation rates (see Model predictions and hypotheses) and will therefore not be discussed explicitly.

Fig. 3 shows the predicted and observed average conflict rates across experimental conditions in the first and the last three rounds. Recall that conflict is a dyad-level outcome that occurs if both subjects choose Ch. Fig. 3 makes apparent that the observed average conflict rates are significantly different from the point predictions of our model (denoted by the short horizontal dashed lines; also see Table 1) except for one case. However, the expected directions and relative differences between conflict rates within as well as across experimental conditions are borne out rather well. In line with our hypothesis H3a, conflict rates at $t = 0$ are highest in *No Info* and *Opponent History* and lowest in the two experimental conditions in which subjects obtain information about their own ability from the start. Bonferroni corrected pairwise comparison tests (comparing conflict rates at $t = 0$ across conditions) produce only two statistically insignificant differences between the two highest and the two lowest rates at $t = 0$ (also see Fig. 3). In line with hypothesis H3b, conflict rates change within conditions over time. Conflict rates are statistically lower in the last three rounds than in the first three rounds in all but the *Own Ability* condition ($z = 3.45$, $p = 0.001$; $z = -0.14$, $p = 0.887$; $z = 5.42$, $p < 0.001$ and $z = 4.10$, $p < 0.001$).

These results provide clear evidence for our prediction that more information about subjects' abilities leads to less conflict. And if information about subjects' abilities cannot be increased, as in the *Own Ability* condition, conflict rates do not change.

The combination of our findings so far, that ability is positively associated with challenging and that conflict rate is low or decreases over time in all the conditions, implies that ability-based hierarchies typically emerge. In the *No Information* and *Opponent History* condition, the positive association between ability and challenging takes multiple rounds to appear and challenging rates decrease throughout the experiment. Conversely, in the *Own Ability* and *Own Ability + Opponent History* conditions, there is a clear positive relationship between ability and challenging already in the first round, and, conflict rates remain fairly stable throughout. This implies that hierarchy formation is slower in the *No Information* and *Opponent History* conditions than in the *Own Ability* and *Own Ability + Opponent History* conditions.

The decrease in conflict and the emergence of ability-based hierarchy may come at a price: higher ability subjects will exploit lower ability subjects. This is implied by our last hypothesis, which we test next.

Fig. 4 shows the predicted and observed average exploitation rates across experimental conditions in the first and the last three rounds. Exploitation too is a dyad-level outcome that occurs if one subject chooses Ch and the other subject chooses Bd in an

Fig. 2. Average challenge rates across experimental conditions, ability and time.

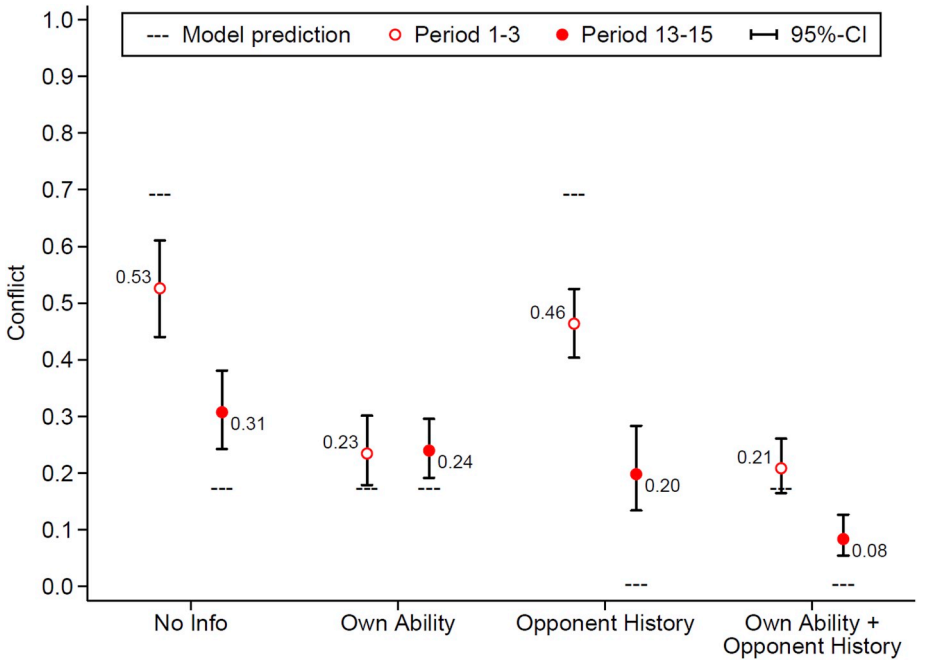

Fig. 3. Average conflict rates across experimental conditions time.

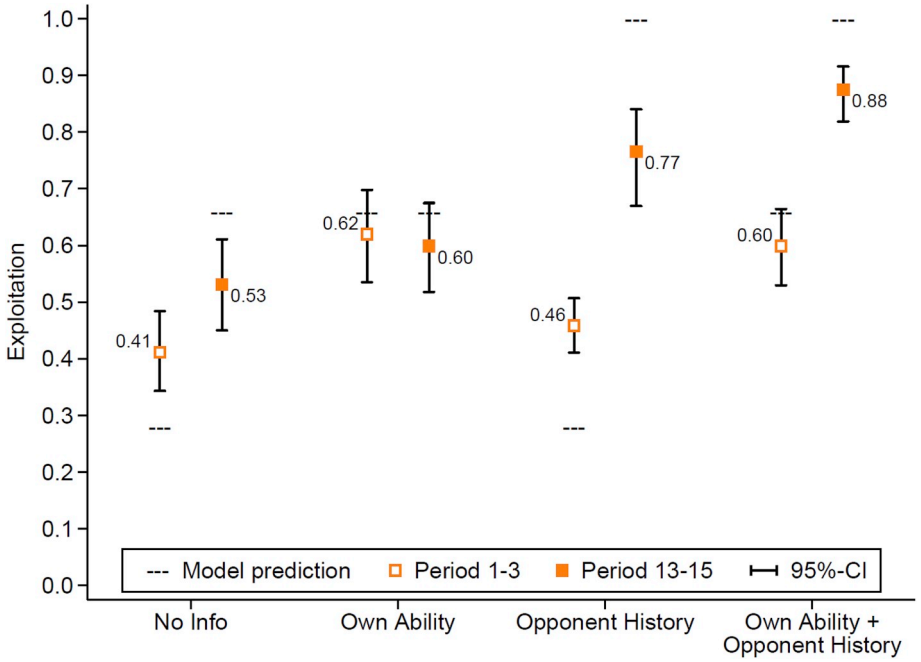

Fig. 4. Average exploitation rates across experimental conditions and time.

interaction. Fig. 4 is almost the mirror image of Fig. 3. Most important, in line with H4, the exploitation rate is highest where the conflict rate is lowest—in the last three periods in the *Own Ability* + *Opponent History* condition. The exploitation rate of 0.88 is significantly higher than the exploitation rates in the last three periods in conditions *No Info*, *Own Ability*, and *Opponent History* ($z = -5.61$, $p < 0.001$; $z = -4.86$, $p < 0.001$ and $z = -2.72$, $p = 0.007$, respectively).

Overall we find clear support for five of our six hypotheses, and hypothesis H2b finds partial support. While we find a close match between our model predictions and observations in qualitative terms, there are differences between our predictions and subjects' behaviour in quantitative terms. In the following section, we try to close the gap between predictions and behaviour as much as possible using psychological and demographic variables.

5.3. Explaining deviations from model predictions

Our formal model implies that actors are homogeneous with regard to social preferences and risk preferences. In fact we assume that actors are only self-regarding (i.e. only care about their own outcomes) and risk neutral (i.e. neither risk averse nor risk seeking). There are two types of deviations from model predictions: choosing to *back down* (Bd) when a subject should have chosen challenge (Ch), and choosing Ch when they should have chosen Bd. We use measures of gender, risk preferences (Binswanger, 1980; Eckel and Grossman, 2002; Holt and Laury, 2002), other-regarding preferences (Andreoni, 1990; Fehr and Schmidt, 1999), competitiveness, and dominant personalities to explain deviations from model predictions. Since this is a purely exploratory exercise, we do not state any hypotheses. Hence, results reported in this section should be interpreted with caution and corroborated in future studies.

We focus on the final five periods of each treatment because challenge rates become broadly stable by that point in the experiment (see Fig. 2) and the final five periods provides us with sufficient observations of deviations. We do not consider deviations from model predictions at the start of the experiment because there are few or they are difficult to identify.[5] By the end of the experiment, in the *No Info* and *Own Ability* conditions, subjects of low ability (1 and 2) should always choose Bd and subjects of high ability (3 and 4) should always choose Ch. Accordingly, Ch is conceived as deviation for the former, while Bd is conceived as deviation for the latter. In the *Opponent History* and the *Own Ability* + *Opponent History* conditions subjects should choose Ch when encountering an opponent with a lower ability and choose Bd when encountering an opponent with a higher ability (with random matching this means that ability 1 never choose Ch, ability 2 in 33% of cases, ability 3 in 66% of cases, and ability 4 always Ch). In these two conditions, we determine deviations based on subjects' actions, their abilities, and their opponents' abilities.

Out of the 2560 decisions taken by subjects in the final five periods of all conditions, only 17.7% constitute deviations from predictions (454/2560). Although the number of deviations is relatively low, we want to know why they occur. To address this

[5] We would only be able to analyse deviating behaviour in the *Own Ability* and *Own Ability* + *Opponent History*, because in the other two treatments any individual action is consistent with the mixed strategy equilibrium. Moreover, there are no deviations to Bd in the *Own Ability* + *Opponent History* treatment.

question, we estimate a multinomial logistic regression with the following categorical outcome variable: 0 = no deviation, 1 = deviation from Ch to Bd, 2 = deviation from Bd to Ch. There is some indication that conditions matter: only 10.9% deviations occur in *Own Ability* + *Opponent History* (70/640), while there are 16.1% (103/640) in *Own Ability*, 21.6% (138/640) in *No Info*, and 22.3% (143/640) in *Opponent History*. We thus control for experimental conditions in our statistical model. We also control for ability. We do this using a binary variable for 'low' and 'high' for which abilities 1 and 2 are classed as low while abilities 3 and 4 are high. We use this simplification to allow model stability as low ability subjects (1 and 2) deviate in comparable ways and high ability subjects (3 and 4) deviate in comparable ways.[6]

Table 2 shows the estimation results of our multinomial logit model. Our dominance and competitiveness scales do not predict deviations. However, SVO is positively associated with deviations to Bd (relative risk ratio = 1.037, $p = 0.004$) and risk preferences are positively associated with deviations to Ch (relative risk ratio = 1.351, $p = 0.001$). This means that subjects who are more prosocial, and consider others' payoffs, are more likely to back down than predicted by our model, while subjects who are more risk seeking are more likely to challenge than predicted. While these two factors help explain deviations from predictions, they are not particularly strong predictors. Moving from a low SVO of −10 to a high SVO of 30 only changes the probability of choosing Bd (instead as predicted Ch) from 1% to 6% (holding all other variables at their means). Similarly, increasing risk preferences from the lowest value of 1 to the highest value of 5 only increases the probability of choosing Ch (instead as predicted Bd) from 5% to 14%. We do not find any evidence that gender matters.

6. Discussion and conclusions

We examine experimentally how informal hierarchies form when people, arranged in small groups, repeatedly interact with each other in a conflict situation. Our subjects play the Hierarchy Contest Game (HCG) and we systematically manipulate the information that they have about their own ability and the ability of others and we consider the effect that such information has on subjects' actions and the outcomes that arise from their interactions. We derive our predictions from our game theoretic model and find that subjects' behaviours closely match the qualitative predictions of our model.

As anticipated by our model, at the start of the experiment, challenge rates are not associated with ability in the conditions in which subjects do not know their ability (*No Info* and *Opponent History*). Since they lack knowledge of their ability, these subjects cannot condition their actions based on it. Conversely, when subjects know their ability from the start (*Own Ability* and *Own Ability* + *Opponent History*) they pre-emptively adjust their behaviour such that the higher ability subjects challenge more frequently than lower ability subjects—this holds irrespective of whether or not they observe their partners' history.

As they progress through the experiment, however, subjects sometimes end up in conflict, and, this allows them to learn about their ability relative to others. Thus even subjects in the *No Info* and *Opponent History* can infer their own ability and adjust their actions accordingly (subjects in the other conditions already know their ability). This changes how subjects behave such that by the end of the experiment, subjects of a higher ability challenge more than those with a lower ability in every condition.

Whether or not subjects know their own ability is the strongest factor in influencing challenging rates and dynamics. This can be clearly seen by comparing the challenging rates in the *No Info* and *Opponent History* conditions to that in the *Own Ability* and *Own Ability* + *Opponent History* conditions (Fig. 2). Within these conditions, the challenging rates at the start and end of the experiment are similar as are the dynamics of challenging over time; conversely comparing across these conditions, both rates and dynamics are substantially different.

Despite this, opponent history also has an effect, albeit a more subtle one, on shaping challenging behaviour. Specifically, opponent history allows subjects with an intermediate ability to 'fine-tune' their challenging. Ability 2 and 3 subjects in *Opponent History* and *Own Ability* + *Opponent History* should, and do, challenge at intermediate rates—since they can pick more precisely who to challenge. In the *No Info* and *Own Ability* conditions they cannot do this and so even intermediate ability subjects challenge at very high or very low rates.

Turning to conflict, the most important of the outcomes, we find that average conflict in the experiment is negatively associated with information in two ways. First, it is reduced when subjects have more sources of information about themselves and their opponents. Conflict is highest in *No Info*, lower in *Own Ability* and the *Opponent History* and lowest of all in the *Own Ability* + *Opponent History*. Thus, the more information sources there are, the less conflict occurs over the course of the experiment. Second, conflict is reduced over time as subjects interact and generate knowledge about themselves and their opponents. As our model predicts, conflict decreases over time in every condition (except *Own Ability)*. Through repeated interactions, subjects reveal relevant information about themselves and, when possible, learn about their opponents' histories.

That conflict does not decrease in the *Own Ability* condition provides further support for our model and postulated mechanism. Only in this condition do subjects not learn additional information over the course of the experiment. Already in the beginning they know

[6] Ability 1 always deviates to Ch while ability 2 almost always deviates to Ch; ability 4 always deviates to Bd while ability 3 almost always does the same. These patterns of deviating are, to a large extent, pre-determined by model parameters and our experimental design. Ability 1 subjects can only ever deviate to Ch (since they are always predicted to Bd) and Ability 4 subjects can only ever deviate to Bd (since they should always Ch). In the *No Info* and *Own Ability* conditions, abilities 2 and 3 are similarly restricted: ability 2 can only deviate to Ch and ability 3 to Bd. The only possibility for abilities 2 and 3 to deviate to both Bd and Ch are in the *Opponent History* and *Own Ability* + *Opponent History* conditions. However, here too deviation is restricted: ability 2 can mostly deviate to Ch (they tend to encounter subjects higher than them in ability and so should typically Bd) while ability 3 can mostly deviate to Bd (they tend to encounter subjects lower than them and so should typically Ch).

Table 2
Multinomial logit of deviations from model predictions in the final 5 periods.

	Deviation from Ch to Bd		Deviation from Bd to Ch	
	exp(Coef.)	SE	exp(Coef.)	SE
Experimental design features				
High ability (3 & 4)	6.409***	2.488	0.092***	0.026
Own Ability condition	0.905	0.354	0.554*	0.154
Opponent History condition	1.180	0.491	0.942	0.251
Own Ability + Opponent History condition	0.514	0.219	0.378***	0.109
Individual-level characteristics				
Female	0.636	0.196	1.083	0.266
SVO angle	1.037**	0.013	0.994	0.008
Risk preference	0.849	0.122	1.351***	0.118
Dominance scale	0.990	0.256	1.031	0.206
Competitiveness scale	0.879	0.204	0.848	0.124
pseudo R^2	0.152			
$\chi^2_{(18)}$	145.7			
N_1 (decisions)	2560			
N_2 (clusters)	128			

Notes: The table lists exponentiated coefficients estimates from a multinomial logit and cluster-robust standard errors (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$, for two-sided tests). The categorical outcome variable has three categories: 0 = no deviation, 1 = deviation from Ch to Bd, 2 = deviation from Bd to Ch. The first category constitutes the baseline. N_1 indicates the number of decisions and N_2 the number of clusters.

their own ability and since they cannot observe others' history, there is nothing left for them to discover. In the other conditions subjects discover their own ability and/or learn the ability of their opponents by these opponents' histories of actions and interaction outcomes.

Both of these findings support Gambetta's (2009) argument of prisoners fighting, in part, because of a lack of credible information, and that were these to be available, conflict would be reduced. Similar to our subjects who do not know their opponents' histories, new entrants to a prison typically fight more frequently than those who are already established in an institution. And, similar to our subjects who lack information about their abilities, prisoners who have less experience with violence—including those who are younger and women—fight at high rates. Our experiment provides clear causal evidence that information shapes conflict and this supports the theory that information matters in prison conflict.

Some previous literature posits that when people have the opportunity to create and build reputations about their abilities, they have increased incentives to compete (Benard, 2013, 2015). This leads to increased aggression—challenging in our case. The perspective we present here is that information reduces the need for conflict. Consistent with our model, findings decisively resolve this question in favour of the latter possibility: past history helps solve conflict and reduces challenging. The differences in design between our experiment and that used by Benard may account for the different results; in contrast to Benard we consider repeated interactions and opponent history contains information on both challenging *and* interaction outcomes.

Our findings concerning conflict also match those reported by Szekely and Gambetta (2019) who show that signs and signals of toughness reduce aggression and, based on simulations, resulting fights.

Overall, subjects' actions and the outcomes that occur are close to the predictions of our model. Our model captures well the qualitative time and treatment effects on challenging, conflict, and exploitation. We show that game theoretic models of strategic interdependence can, at least in some settings, successfully represent human behaviour.

Of course, our model does not perfectly predict behaviour and there is some deviation in subjects' actions. One reason might be that we assume backward looking agents and do not take reputation formation through forward looking agents into account. Although we doubt that this could explain the deviations (see footnote 3), such a theoretical extension would be worthwhile pursuing in future research.

We examine possible explanations for these deviations using individual differences between subjects. We find that social value orientation (SVO) and risk preferences are the only two of our psychological measures that are predictive of deviations. SVO is positively related to backing down—an intuitive result in that people who care more about others' outcomes are less likely to challenge. Risk preferences are associated with more challenging; this is intuitive too, as people who are more willing to take risks choose the action associated with the greater variation in outcome. It is also broadly consistent with the experimental results in the economic contest literature (Dechenaux et al., 2015). Subjects' gender has no bearing on whether they deviate from model predictions. However, these results are based on a purely exploratory analysis and therefore should be interpreted with caution.

Informal, ability-based hierarchies emerge in every experimental condition. Even though we allocate abilities to subjects from the outset of our experiment, ability is an individual-level property, and, that it is subsequently reflected at the group-level as a hierarchy, is far from certain (Manzo and Baldassarri, 2015). Yet, we found this to be the case. By the end of the interaction sequence, subjects who are higher in ability challenge much more than subjects lower in ability. Conflict too remains either low or decreases through the experiment. Both results imply that when people enter a new environment and they know only their own ability (and the possible range of abilities in their group) then that is sufficient to create a ranking highly correlated with ability. It is not necessary to know others' history for hierarchy formation. And when past history on others is available, subjects with lower ability defer to higher ability subjects in a more selective way. While we have no doubt that arbitrary hierarchies can emerge, as past work shows (Lynn et al., 2009;

Manzo and Baldassarri, 2015), we demonstrate here that they can also be tightly linked to abilities.

Ultimately, our experiment demonstrates that information availability has a fundamental role in shaping conflict and the emergence of hierarchies. Although ability hierarchies emerge in every condition, the extent to which hierarchies form and the amount of conflict that arises during and following their formation, depends on information about own and others' abilities. While our experiment addresses some fundamental questions, there are many future directions in which to proceed with this research. One particularly interesting question is what happens with the includes of equal ability individuals. As Gould (2003) and Gambetta (2009) argue, it is likely that for these people information does not reduce conflict since it cannot settle disputes.

Appendix A

HCG 1: Neither player knows their ability and the other player's ability

The fact that neither actor knows their own ability nor the ability of their opponent turns the HCG into the chicken game shown in Fig. A. Since, in expectation, an actor has higher ability in half of the interactions and lower ability in the other half, the payoffs if both players choose Ch are now $0.5(w - c) + 0.5(-c) = 0.5w - c$ for each player. The other parts of the payoff matrix do not change as compared to the HCG, as players' ability is only relevant if both choose Ch. The game shown in Fig. A has two pure strategy Nash equilibria, one in which Player A chooses Bd and Player B chooses Ch, and one in which Player A chooses Ch and Player B chooses Sh. This can be verified by showing that in equilibrium no player has an incentive to change their strategy unilaterally (recall that we assume $c > 0.5w$).

		Player B: Ch	Player B: Bd
Player A	Ch	$0.5w - c$ $0.5w - c$	w 0
	Bd	0 w	$0.5w$ $0.5w$

Fig. A. Chicken game.

The chicken game also has a mixed strategy Nash equilibrium (MSE). Denote the probability of Player A to choose Ch with p and the probability of Player B to choose Ch with q. The expected utility of Player A from challenging is $U_A(\text{Ch}) = q(0.5w - c) + (1 - q)w$, and Player A's expected utility from backing down is $U_A(\text{Bd}) = q0 + (1 - q)0.5w$. Since the game is symmetric, Player B's utilities for either action can be derived correspondingly. In the MSE both players must be indifferent between choosing Ch and Bd. That is, it must hold that $U_A(\text{Ch}) = U_A(\text{Bd})$ and $U_B(\text{Ch}) = U_B(\text{Bd})$. Solving the two equations for the two unknowns p and q gives the probability with which either player chooses Ch in the MSE:

$$p^* = \quad q^* = 0.5\frac{w}{c} \tag{A1}$$

HCG 2: Both players know their ability but not the other player's ability

The HCG can now be conceptualized as one with actors having private information about their ability (Harsanyi, 1967). While actors know the probability distribution of abilities and thus their rank in the group, they do not know their interaction partners' abilities. However, given group size N and their ability, they can derive the probability of being the player with higher ability in an interaction:

$$p_i = \Pr\left(A_i > A_j | A_i, N\right) = \frac{A_i - 1}{N - 1} \tag{A2}$$

Hence, an actor i's expected utility from choosing Ch is:

$$U_i(Ch|A_i) = p_i\left[q_j(w - c) + \left(1 - q_j\right)w\right] + (1 - p_i)\left[r_j(-c) + \left(1 - r_j\right)w\right] \tag{A3}$$

In equation (A3), q_j is the challenging probability of a weaker player j and r_j is the challenging probability of a stronger player j. Player i's expected utility from backing down is:

$$U_i(Bd|A_i) = p_i\left(1 - q_j\right)\frac{w}{2} + (1 - p_i)\left(1 - r_j\right)\frac{w}{2} \tag{A4}$$

Actor i chooses Ch if $U_i(\text{Ch}|A_i) > U_i(\text{Bd}|A_i)$, that is, if:

$$w - p_i q_j(2c - w) - (1 - p_i)r_j(2c + w) > 0 \tag{A5}$$

For the actor with highest ability (i.e., $A_i = N$) $p_i = 1$. Therefore, based on equation (A5), $w/(2c - w) > q_j$ must hold for the actor with highest ability to choose Ch. Since $w > c$, it follows that $w/(2c - w) > 1$, which means that the actor with highest ability always challenges. Consequently, for the actor with second highest ability $r_j = 1$. Now we can show that there is a threshold τ_u such that every actor i with $A_i > \tau_u$ chooses Ch if the player with higher ability challenges with certainty. That is, assuming $r_j = 1$, we can rearrange equation (A5) to:

$$\frac{p_i(2c + w) - 2c}{p_i(2c - w)} > q_j \tag{A6}$$

Substituting p_i for equation (A2), we get:

$$\frac{(A_i - 1)(2c + w) - 2c(N - 1)}{(A_i - 1)(2c - w)} > q_j \tag{A7}$$

As long as the left-hand side of equation (A7) is larger than 1, actor i with ability A_i challenges irrespective of the proportion of players with lower ability who challenge. Thus, solving for A_i gives:

$$A_i > \frac{c(N - 1) + w}{w} = \tau_u \tag{A8}$$

In the same vein, we now calculate a lower threshold τ_l such that every actor i with $\tau_l > A_i$ never challenges if the actor with higher ability challenges with certainty (i.e., $r_j = 1$). This corresponds to checking for which highest value A_i the left-hand side of equation (A7) is still negative. After rearranging the numerator of equation (A7) accordingly, we get:

$$\tau_l = \frac{2cN + w}{2c + w} > A_i \tag{A9}$$

Note that we are still assuming all actors with higher ability than the actor with highest ability for whom equation (A9) holds challenge with certainty. However, there is a gap between τ_u and τ_l and for all players i with $\tau_u > A_i > \tau_l$ it matters how large the challenging probabilities of higher and lower ability players are. Subtracting τ_l from τ_u gives $c(N - 1)(2c - w)/w(2c + w)$ which can be shown to be smaller than 1 if:

$$N < \frac{w^2 + c(2c + w)}{c(2c - w)} \tag{A10}$$

That is, given some reasonable values for w and c (e.g., $w = 5$ and $c = 3$), N must be smaller than 19.3 for there to be less than one actor i with $\tau_u > A_i > \tau_l$.[7] Hence, for groups that are not too large, the game outlined above has a Nash equilibrium in which the strongest $N - \lfloor \tau_u \rfloor$ actors challenge and the other actors back down with certainty.

Appendix B

Table B
Descriptive statistics

	N	Range / %	Mean	SD
Dependent variables				
initial challenge rate (Ch = 1)	1536	61.9%		
ultimate challenge rate (Ch = 1)	1536	55.3%		
initial conflict rate (Fi = 1)	768	35.8%		
ultimate conflict rate (Fi = 1)	768	20.7%		
initial sharing rate (Sh = 1)	768	11.98%		
ultimate sharing rate (Sh = 1)	768	10.03%		
deviations from predictions	2560			
(no deviation = 0)		82.27%		
(deviation from Ch to Bd = 1)		5.78%		
(deviation from Bd to Ch = 2)		11.95%		
Independent variables				
ability		1–4		
knows own ability (yes = 1)		50%		
knows interaction partner's history (yes = 1)		50%		
experimental condition		1–4		
sex (female = 1)	128	60.94%		

(*continued on next page*)

[7] If we restrict the analysis to values of N, c and w such that there is at most one actor i with $\tau_u > S_i > \tau_l$, then it remains to show that for actor i challenging is the dominant strategy. From the perspective of actor i, all actors with higher ability challenge whereas all actors with lower ability back down. Given this constellation, it remains to show that it does not pay off for actor i to back down. This is the case if equation (A3) > equation (A4) given $r_j = 1$ and $q_j = 0$. Thus, it must hold that $A_i > (2\ cN + w)/(2c + w)$ which holds as $A_i > \tau_l$.

Table B (*continued*)

	N	Range / %	Mean	SD
risk preference	128	1–5	2.422	1.240
SVO	128	−16.26 - 45	13.373	14.249
competitiveness	128	1.4–5	3.479	.753
dominant personality	128	1.88–4.65	3.321	.616
Control variables				
age	128	17–60	24.125	6.175
ability 1 (yes = 1)		25%		
ability 2 (yes = 1)		25%		
ability 3 (yes = 1)		25%		
ability 4 (yes = 1)		25%		
ability 2 × knows history (yes = 1, yes = 1)		12.5%		
ability 3 × knows history (yes = 1, yes = 1)		12.5%		

Appendix C. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ssresearch.2019.102393.

References

Andreoni, James, 1990. Impure Altruism and donations to public goods: a theory of warm-glow giving. Econ. J. 100 (401), 464–477.
Benard, Stephen, 2013. Reputation systems, aggression, and deterrence in social interaction. Soc. Sci. Res. 42 (1), 230–245.
Benard, Stephen, 2015. The value of vengefulness: reputational incentives for initiating versus reciprocating aggression. Ration. Soc. 27 (2), 129–160.
Binswanger, Hans P., 1980. Attitudes toward risk: experimental measurement in rural India. Am. J. Agric. Econ. 62, 395–407.
Blau, Peter Michael, 1964. Exchange and Power in Social Life. John Wiley & Sons, New York, NY.
Boehm, Christopher, 1999. Hierarchy in the Forest: the Evolution of Egalitarian Behavior. Harvard University Press, Cambridge, MA.
Bolle, Friedel, Kaehler, Jessica, 2007. Introducing a signaling institution: an experimental investigation. J. Inst. Theoretical Econ. JITE 163 (3), 428–447.
Buskens, V., Werner, R., van der Veer, J., 2010. Trust in triads: an experimental study. Soc. Netw. 32 (4), 301–312.
Carrillo, Juan D., Palfrey, Thomas R., 2009. The compromise game: two-sided adverse selection in the laboratory. Am. Econ. J. Microecon. 1 (1), 151–181.
Chambers, John R., De Dreu, Carsten K.W., 2014. Egocentrism drives misunderstanding in conflict and negotiation. J. Exp. Soc. Psychol. 51, 15–26.
Chase, I.D., Lindquist, W.B., 2009. Dominance hierarchies. In: Hedström, P., Bearman, P. (Eds.), Oxford Handbook of Analytical Sociology. Oxford University Press, Oxford, pp. 566–591.
Chase, Ivan D., Seitz, Kristine, 2011. Self-structuring properties of dominance hierarchies: a new perspective. Adv. Genet. 75, 51–81.
De Dreu, C.K.W., Gross, J., Méder, Z., Giffin, M., Prochazkova, E., Krikeb, J., Columbus, S., 2016. In-group defense, out-group aggression, and coordination failures in intergroup conflict. Proc. Natl. Acad. Sci. U. S. A 113 (38), 10524–10529. https://doi.org/10.1073/pnas.1605115113.
Crosetto, P., Weisel, O., Winter, F., 2019. A flexible z-Tree and oTree implementation of the social value orientation slider measure. J. Behav. Exp. Financ. 23, 46–53. https://doi.org/10.1016/j.jbef.2019.04.003.
Dechenaux, E., Kovenock, D., Sheremeta, R.M., 2015. A survey of experimental research on contests, all-pay auctions and tournaments. Exp. Econ. 18 (4), 609–669.
DeScioli, Peter, Wilson, Bart J., 2011. The territorial foundations of human property. Evol. Hum. Behav. 32 (5), 297–304.
Dixit, Avinash, 1987. Strategic behavior in contests. Am. Econ. Rev. 77 (5), 891–898.
Eckel, Catherine C., Grossman, Philip J., 2002. Sex differences and statistical stereotyping in attitudes toward financial risk. Evol. Hum. Behav. 4 (23), 281–295.
Edgar, Kimmett, Martin, Carol, 2001. Conflicts and Violence in Prison. Report for the Economic and Social Research Council.
Enquist, Magnus, 1985. Communication during aggressive interactions with particular reference to variation in choice of behaviour. Anim. Behav. 33 (4), 1152–1161.
Fearon, James D., 1995. Rationalist explanations for war. Int. Organ. 49 (3), 379–414.
Fehr, Ernst, Schmidt, Klaus M., 1999. A theory of fairness, competition, and cooperation. Q. J. Econ. 114 (3), 817–868.
Fischbacher, Urs, 2007. Z-tree: Zurich toolbox for ready-made economic experiments. Exp. Econ. 10 (2), 171–178.
Fitzmaurice, Garrett M., Laird, Nan M., Ware, James H., 2004. Applied Longitudinal Analysis. John Wiley & Sons, Hoboken, NJ.
Gambetta, D., 2009. In: Codes of the Underworld: How Criminals Communicate. Princeton University Press, Princeton, NJ.
Gould, Roger V., 2003. Collision of Wills: How Ambiguity about Social Rank Breeds Conflict. University Of Chicago Press, Chicago, IL.
Grafen, Alan, 1990. Biological signals as handicaps. J. Theor. Biol. 144 (4), 517–546.
Greiner, Ben, 2015. Subject pool recruitment procedures: organizing experiments with ORSEE. J. Econ. Sci. Assoc. 1 (1), 114–125.
Harsanyi, John C., 1967. "Games with incomplete information played by 'Bayesian' players, I–III, Part I. The basic model. Manag. Sci. 14 (3), 159–182.
Helmreich, Robert L., Spence, J.T., 1978. Work and family orientation questionnaire: an objective instrument to assess components of achievement motivation and attitudes toward family and career. JSAS Catalog Sel. Documents Psychol. 8.
Holt, Charles A., Laury, Susan K., 2002. Risk aversion and incentive effects. Am. Econ. Rev. 92 (5), 1644–1655.
Homans, G.C., 1961. Social Behavior: its Elementary Forms. Harcourt Brace & World, New York, NY.
Kaminski, Marek, 2004. Games Prisoners Play: the Tragicomic Worlds of Polish Prison. Princeton University Press, Princeton, NJ.
Kübler, Dorothea, Müller, Wieland, Normann, Hans-Theo, 2008. Job-market signaling and screening: an experimental comparison. Games Econ. Behav. 64 (1), 219–236.
Long, J. Scott, 1997. Regression Models for Categorical and Limited Dependent Variables. Sage Publications, Thousand Oaks, CA.
Lynn, Freda B., Podolny, Joel M., Lin, Tao, 2009. A sociological (De)Construction of the relationship between status and quality. Am. J. Sociol. 115 (3), 755–804.
Magee, Joe C., Galinsky, Adam D., 2008. Social hierarchy: the self-reinforcing nature of power and status. Acad. Manag. Ann. 2 (1), 351–398.
Manzo, Gianluca, Baldassarri, Delia, 2015. Heuristics, interactions, and status hierarchies: an agent-based model of deference exchange. Sociol. Methods Res. 44 (2), 329–387.
Maynard Smith, John, Harper, David, 2003. Animal Signals. Oxford University Press, Oxford.
Milinski, Manfred, 2016. Reputation, a universal currency for human social interactions. Phil. Trans. R. Soc. B 371 (1687), 20150100.
Murphy, Ryan O., Ackerman, Kurt A., Handgraaf, Michel J.J., 2011. Measuring social value orientation. Judgment Decis. Mak. 6 (8), 771–781.
O'Keeffe, M., Viscusi, W.K., Zeckhauser, R.J., 1984. Economic contests: comparative reward schemes. J. Labor Econ. 2 (1), 27–56.
Podolny, Joel Marc, Lynn, Freda, 2009. Status. In: Hedström, P., Bearman, P. (Eds.), The Oxford Handbook of Analytical Sociology. Oxford University Press, Oxford.
Powell, Robert, 2004. Bargaining and learning while fighting. Am. J. Pol. Sci. 48 (2), 344–361.

Przepiorka, W., Berger, J., 2017. Signaling theory evolving: signals and signs of trustworthiness in social exchange. In: Jann, B., Przepiorka, W. (Eds.), Social Dilemmas, Institutions, and the Evolution of Cooperation. De Gruyter Oldenbourg, Berlin, pp. 373–392.
Przepiorka, Wojtek, Diekmann, Andreas, 2013. Temporal embeddedness and signals of trustworthiness: experimental tests of a game theoretic model in the United Kingdom, Russia, and Switzerland. Eur. Sociol. Rev. 29 (5), 1010–1023.
Przepiorka, Wojtek, Norbutas, Lukas, Corten, Rense, 2017. Order without law: reputation promotes cooperation in a cryptomarket for illegal drugs. Eur. Sociol. Rev. 33 (6), 752–764.
Rapoport, Anatol, Chammah, Albert M., 1966. The game of chicken. Am. Behav. Sci. 10 (3), 10–14.
Ray, John J., 1981. Authoritarianism, dominance and assertiveness. J. Personal. Assess. 45 (4), 390–397.
Searcy, W.A., Nowicki, S., 2005. The Evolution of Animal Communication: Reliability and Deception in Signaling Systems. Princeton University Press, Princeton, NJ.
Silverman, Dan, 2004. Street crime and street culture. Int. Econ. Rev. 45 (3), 761–786.
Simpson, B., Willer, R., Ridgeway, C.L., 2012. Status hierarchies and the organization of collective action. Sociol. Theory 30 (3), 149–166.
Snijders, Tom A.B., Bosker, Roel, 2012. Multilevel Analysis: an Introduction to Basic and Advanced Multilevel Modeling, second ed. Sage Publications, London.
Spence, Michael A., 1973. Job market signaling. Q. J. Econ. 87 (3), 355–374.
Spence, Michael A., 1974. Market Signaling: Informational Transfer in Hiring and Related Screening Processes. Harvard University Press, Cambridge, MA.
Sykes, Gresham M., 1958. The Society of Captives: A Study of a Maximum Security Prison. Princeton University Press, Princeton, NJ.
Szekely, A., Gambetta, D., 2019. Does information about toughness decrease fighting? Experimental evidence. PLOS ONE, forthcoming.