How the design of cartel fines affects prices: Evidence from the lab*

Sindri Engilbertsson†, Sander Onderstal‡, and Leonard Treuren§

May 2026

Abstract

Antitrust authorities impose substantial penalties on firms engaging in illegal price-fixing. We examine how basing cartel fines on revenue, profit, or price overcharge influences prices, cartel incidence, and cartel stability. In an infinitely repeated Bertrand oligopoly game, we show that fines based on revenues/profits/overcharge incentivize firms to charge prices above/equal to/below the monopoly price. Cartels are stable for a smaller range of discount factors when fines are based on overcharges rather than other bases. We test these predictions in a laboratory experiment where subjects can form cartels, which allows them to discuss pricing at the risk of being detected and fined. By equalizing expected fines across treatments, we isolate the effect of the fine's base. We find that market prices are lowest under overcharge-based fines and highest under revenue-based fines. While these results align with the theoretical predictions, cartel incidence and cartel stability do not differ significantly across fining regimes. Our results suggest that antitrust authorities could improve enforcement by shifting from revenue-based fines to profit- or overcharge-based fines.

Keywords: Antitrust; Cartel; Collusion; Repeated game; Experiment
JEL Codes: C73; C92; D43; K21; L41

*For useful comments and discussions, we thank the editor, Dennis Carlton, several referees, Lisa Bruttel, Joe Harrington, Florian Heine, João Montez, Evgenia Motchenkova, Catherine Roux, Maarten Pieter Schinkel, Yossi Spiegel, Christian Thöni, and seminar participants at the University of Amsterdam, KU Leuven, LMU München, Sun Yat-Sen University Guangzhou, Padova University, and the 2024 MaCCI (Mannheim), 2024 IIOC (Boston), 2024 BECCLE (Bergen), 2024 M-BEES (Maastricht), 2024 CRESSE (Crete), 2024 EEA (Rotterdam), 2024 EARIE (Amsterdam), and 2024 ESA (Helsinki) conferences. Engilbertsson thanks the UvA Research Priority Area in Behavioral Economics (grant EB-1145). Treuren credits support from ERC Consolidator grant 816638 and KU Leuven internal funds. This experiment was preregistered (AEA RCT ID: AEARCTR-0012043). Opinions and errors remain ours.

†University of Amsterdam and Tinbergen Institute; s.engilbertsson@uva.nl.
‡University of Amsterdam and Tinbergen Institute; a.m.onderstal@uva.nl.
§KU Leuven; leonard.treuren@kuleuven.be.

1 Introduction

Antitrust authorities and courts regularly impose substantial penalties on firms that participate in illegal horizontal price-fixing agreements.[1] Across antitrust jurisdictions, such fines are based on cartel members' revenue.[2] However, theoretical literature has long established that revenue bases can increase cartel prices.[3] This discrepancy between theory and practice raises the question of the relative performance of different fining regimes.

This paper experimentally investigates how different fining bases affect prices, cartel incidence, and cartel stability. Specifically, we examine whether the theoretical concerns surrounding revenue-based fines – which are based on equilibrium selection assumptions – are justified. We also compare the effectiveness of revenue-based fines to viable alternatives. Since these questions are inherently empirical, but observational data offer limited insights, we use a laboratory experiment to address them. We compare revenue-based fines to two alternatives for which a legal basis exists in the U.S. fining guidelines: fines based on cartel members' profits, and fines based on cartel members' price overcharge with respect to the competitive price.[4]

We build our hypotheses on the analysis of infinitely repeated Bertrand games where firms can use trigger strategies to realize above one-shot equilibrium prices as part of a subgame-perfect Nash equilibrium. Coordinating on such trigger strategies leaves a paper trail that is detected with a fixed probability each period by the antitrust authority. Members of detected cartels are fined, and undiscovered cartels from previous periods can be detected. The basis of a cartel member's fine is either her revenue, her profit, or the price overcharge she sets. After discovery, cartels are reformed with a fixed probability in every subsequent period. This theoretical model isolates key factors of the antitrust policy under consideration.

To deal with the multiplicity of equilibria in infinitely repeated oligopoly games, we follow the theoretical literature and make assumptions on equilibrium selection. In particular, we assume that firms coordinate on the joint-profit-maximizing price by using trigger strategies.

[1] For instance, the European Commission imposed a €3.8 billion fine on truck manufacturers in the Trucks case (2016/2017). Other examples include the $2.5 dollars and €1.4 billion fines in the Foreign Exchange Market case by the United States Department of Justice (2015) and the European Commission (2019/2021), respectively, and the fine of ¥101 billion imposed by Japan's antitrust authority in 2023 on the electric power cartel.

[2] For instance, the guidelines on the method of setting fines by the European Commission (2006) state that: "In determining the basic amount of the fine to be imposed, the Commission will take the value of the undertaking's sales of goods or services to which the infringement directly or indirectly relates..." According to United States Sentencing Commission (2023), the base fine for bid-rigging, price-fixing, and market-allocation agreements is "20 percent of the volume of affected commerce" (p.311).

[3] See, e.g., Bageri et al. (2013) and Katsoulacos and Ulph (2013).

[4] For non-antitrust criminal organizations, fines are typically based on either the organization's pecuniary gain from the offense (profit) or the pecuniary loss inflicted by the offense (damages) (United States Sentencing Commission, 2023, p.526).

For each fining regime, we derive a critical discount factor, i.e., the lowest value of the firms' discount factors for which coordinating on the joint-profit-maximizing price is feasible. We then follow the literature by hypothesizing that both cartel incidence and cartel stability are lower the higher is the critical discount factor. Where our theoretical assumptions fail, a comparison of the fining regimes might deliver different results. In particular, revenue bases might not cause cartel prices to exceed the monopoly price.[5]

Our theoretical results on cartel prices align with the earlier literature (e.g., Bageri et al. (2013); Katsoulacos et al. (2015)). Revenue bases incentivize cartel members to charge a price above the no-antitrust monopoly price. In contrast, basing the fine on a cartel member's overcharge reduces the joint-profit-maximizing price compared to the monopoly price, while a profit base leaves the joint-profit-maximizing price unaffected. Intuitively, fines based on revenue serve as a tax pushing prices up; a profit-based fine does not affect the profit-maximizing price because the cartel's expected profits are a fraction of its profits without a fine; in an overcharge regime, the fine is strictly increasing in price, which gives the cartel an incentive to mitigate the price.

In contrast to earlier findings, our model suggests that, among the fining regimes studied, cartel incidence and cartel stability are lowest when cartel fines are based on the price overcharge. The reason is that, in the overcharge regime, defection to a lower price *decreases* the expected fine, in contrast to the revenue and profit regimes. Therefore, we point towards an additional theoretical benefit of an overcharge regime compared to the currently used fining regime. Central to this result is the assumption that defectors can be fined, which is consistent with antitrust practice (Buccirossi and Spagnolo, 2007).

We test the above predictions using a laboratory experiment.[6] While no laboratory experiment can fully replicate real-world cartels, our approach allows us to overcome significant challenges posed by field data. First, observing cartels in the field is inherently difficult because of their illegal nature. Moreover, discovered cartels likely form a non-representative sub-sample of the entire population of cartels. Second, laboratory control allows the researcher to obtain an apples-to-apples comparison of different fining regimes, which is challenging in the field as it is difficult to establish exogenous variation and to measure variables of interest like marginal costs and demand. Finally, experiments enhance internal validity by ensuring that theoretical assumptions are met as closely as possible under controlled conditions. For these reasons, laboratory experiments are widely employed as wind-tunnel tests

[5]For instance, Bageri et al. (2013, p.F550) remark that "of course, it could be argued that the practical significance of this distortion is likely to be small because it requires managers of firms involved in cartels to be well-informed and forward-looking, and to formulate strategic decisions at a level that may not be easily met in reality."

[6]We are not the first to study the effect of fining regimes on cartel behavior in a laboratory study. Fonseca et al. (2022) examine whether fining regimes that impose sanctions on managers involved in cartels have greater deterrence power than regimes that only levy corporate fines, allowing shareholders to determine the labor contracts of the firms' managers. They find less cartelization when managers can be prosecuted.

of theory-based policy recommendations (Falk and Heckman, 2009; List, 2020).

In the experiment, 279 participants compete in indefinitely repeated Bertrand markets that closely mirror our theoretical model. Subjects can opt into cartels by voting, which allows them to freely discuss pricing at the risk of being detected and fined. In the REVENUE, PROFIT, and OVERCHARGE treatments, the fine of a discovered cartel member is based on that individual's revenue, profit, or price overcharge, respectively. We equalize expected fines across treatments so that our results are not driven by behavioral responses to the size of the fine. Varying the treatments between participants allows us to identify the causal link between the three fining regimes and a host of outcomes of interest, including the market price, the price charged by cartels, the likelihood of cartel incidence, cartel formation, cartel recidivism, and cartel stability.

Our experimental findings on prices are in line with the theoretical predictions. While uncartelized markets yield prices close to the one-shot Nash equilibrium price in all three fining regimes, cartel prices are lowest when fines are based on the overcharge and highest when they are based on revenue. Indeed, when fines are based on revenue, both the price agreements that subjects form and the market prices that result from such agreements exceed the monopoly price. However, we find no significant differences in cartel incidence, cartel formation, cartel recidivism, and cartel stability across treatments. Therefore, our data suggest that market price differences across treatments are entirely determined by prices set by cartels. Our findings suggest benefits from antitrust authorities moving away from revenue bases towards profit or overcharge bases. We conclude by arguing that such a change is realistically implementable for a profit-based regime given current legal and institutional constraints.

We contribute to a strand of theoretical literature studying how cartel fining regimes influence market outcomes. In particular, revenue regimes have been shown to have the perverse effect of increasing cartel prices in Bageri et al. (2013) and Katsoulacos and Ulph (2013). Profit bases are studied in Block et al. (1981), Harrington (2004), and Harrington (2005), among others. The overcharge base is proposed as an attractive alternative to revenue and profit bases by Katsoulacos et al. (2015), the paper most closely related to our theoretical model as it compares the same fining regimes.[7] To generate unique equilibria, theoretical models of collusion based on infinitely repeated oligopoly games routinely make assumptions on equilibrium selection, for instance that firms coordinate on the joint-profit-maximizing price. Which equilibria players are most likely to end up in remains an empirical question, which we address using a laboratory experiment.[8]

[7] In contrast to our model, Katsoulacos et al. (2015) assume that defectors cannot be fined, that cartels immediately reform after detection, and that cartels can only be detected in the period in which they are formed. While these differences do not influence the ranking of cartel prices across fining regimes, deterrence is equal in all three regimes under the assumptions of Katsoulacos et al. (2015).

[8] As we compare commonly studied fining bases while holding fixed the level of the fine, we do not address

Our paper also contributes to the broader literature analyzing the impact of various competition policy instruments on cartel behavior. This literature has studied a wide range of policy questions including the effectiveness of leniency programs (see Marvão and Spagnolo (2018) and Hinloopen et al. (2023) for overviews), spillovers from legal cooperation in some markets to tacit collusion in others (e.g., Duso et al. (2014); Sovinsky (2022)), the role of communication in collusion (e.g., Kandori and Matsushima (1998); Awaya and Krishna (2016)), and the effect of market transparency programs on collusion (e.g., Vega-Redondo (1997); Byrne and De Roos (2019)). Theoretical predictions in these domains are routinely tested in laboratory experiments.[9]

Oligopoly laboratory experiments studying corporate leniency programs often compare treatments with fines to treatments with fines and a leniency program. Fines are either independent of firm conduct (e.g., Bigoni et al. (2012, 2015)), based on revenue (e.g., Apesteguia et al. (2007); Hinloopen and Soetevent (2008)), or based on profit (e.g., Andres et al. (2021)). Across treatments, the size of the fine varies but the base of the fine is fixed. In contrast, we hold the size of the expected fine fixed and vary the fining base, which allows us to study the effect of fining structure on cartel behavior. We also have chosen to abstract from leniency programs for the same reason and to limit the cognitive demands on our experimental subjects. We view the inclusion of leniency programs as an avenue for future work.

Finally, we contribute to the experimental literature on cooperation in indefinitely repeated games, surveyed by Dal Bó and Fréchette (2018). This literature finds the discount factor exceeding the critical discount factor to be a necessary, but insufficient, condition for cooperation to emerge in the absence of communication. Moreover, an important finding is that cooperation rates are increasing in the difference between the actual discount factor and the critical discount factor.[10] In contrast to earlier experiments on collusion, we cannot equalize critical discount factors across treatments, as the overcharge-based regime always has a higher critical discount factor than the other two regimes. Still, we do not find differences across fining regimes in any of our measures of collusion. In fact, we observe a tendency towards complete cartelization over time, suggesting that the results surveyed in Dal Bó and Fréchette (2018) do not extend to a setting where subjects can freely communicate.

The structure of this paper is as follows. In Section 2, we present our model and the

the optimal level or design of fines, which have been studied by Buccirossi and Spagnolo (2007), Katsoulacos and Ulph (2013), and Houba et al. (2018).

[9]Experiments show leniency programs having the desired effects on cartel formation, cartel discovery, and the price (e.g., Hinloopen and Soetevent (2008); Bigoni et al. (2012, 2015)); spillovers emerging from legal cooperation in one experimental market to tacit collusion in another (e.g., Normann et al. (2015); Hinloopen et al. (2024)); explicit communication about future conduct facilitating collusive prices (e.g., Fonseca and Normann (2012); Freitag et al. (2021); Gomez-Martinez et al. (2016); Andres et al. (2023)); transparency about competitors' actions increasing competition (e.g., Huck et al. (1999, 2000); Offerman et al. (2002)).

[10]Bruttel (2009) provides similar results in a sequential pricing duopoly experiment where an indefinitely repeated game is implemented in the lab as a finitely repeated game with many repetitions.

theoretical results on which we base our hypotheses. Section 3 contains our experimental design, experimental procedures, and hypotheses. Section 4 gives our experimental findings and their implications. Concluding remarks are in Section 5. Proofs of propositions are relegated to Appendix A.

2 Theoretical framework

Our theoretical framework is based on three main assumptions. First, collusive agreements must be self-enforcing due to the illegal nature of price-fixing. Second, communication is required to achieve collusion, and leaves a paper trail which can be detected by the antitrust authority. Third, firms internalize the possibility of price-fixing fines. To replicate key findings from the theoretical literature for the setting we implement in the laboratory, our model closely follows existing theory of antitrust penalties, where these assumptions are routinely made (e.g., Motta and Polo (2003); Aubert et al. (2006); Katsoulacos et al. (2015)). Readers who are mainly interested in the experimental results can proceed directly to Section 3, where we present the experimental design and the theoretical predictions for the parameters used in the experiment.

2.1 The model

Consider an infinitely repeated homogeneous-goods Bertrand game with $n \geq 2$ firms that maximize expected profit and that have a common discount factor $\delta \in (0, 1)$. Each period t, each firm i sets a price $p_{it} \in [0, \bar{p}]$, with $\bar{p} > 0$ the highest possible price. Market demand in period t, $q_t \equiv q(p_t)$, depends on the market price p_t, which is the lowest price set that period, i.e., $p_t \equiv \min_i p_{it}$, with $q(\bar{p}) = 0$. Firms produce at constant marginal cost $c \in (0, \bar{p})$, and average and marginal market revenue are assumed strictly decreasing in market quantity. The m firms that set the lowest price in a given period share the resulting market demand equally: $q_{it} = \frac{q_t}{m}$ if $p_{it} = p_t$. Firms that do not set the lowest price face no demand: $q_{it} = 0$ if $p_{it} > p_t$. Before deciding on the price, firms have the option to form a cartel. When firms choose not to form a cartel, we assume that the one-shot Bertrand Nash-equilibrium always occurs: $p_{it} = c$, implying that profits are zero. We denote the firm-period-level price and profits realized in this competitive benchmark by p^N and π^N.

When firms form a cartel, they open a channel of communication, and may be able to coordinate on market prices above p^N. Such collusive prices are supported by a grim trigger strategy whereby firms coordinate on a price and set it as long as all firms set that price in all previous periods since the inception of the cartel. Otherwise, firms revert to setting price p^N forever.[11] We assume that cartels coordinate on the joint-profit-maximizing price. Let

[11] Our theoretical analysis relies on the assumption that, without communication, firms do not coordinate on such grim-trigger strategies and instead play the one-shot Bertrand equilibrium. This is a common

$p^M(c)$ denote the monopoly price conditional on the firms' marginal costs c in the absence of antitrust. Note that our assumptions imply that $p^M(c)$ strictly increases with c.[12]

An antitrust authority can discover and punish firms' communication. Firms are labeled a cartel in a certain period if and only if they have accessed the chat in the current period or earlier. We assume that once a cartel has been formed, the antitrust authority can detect and punish it, regardless of the behavior of the cartel members. This implies that firms that defect from the cartel agreement and those on the punishment path of the grim trigger strategy profile can be convicted. This is in line with reality, as defectors do not face reduced fines in either the U.S. or the EU (Buccirossi and Spagnolo, 2007). In addition, the paper trail used to convict cartels typically originates years prior to detection and conviction (Kwoka and White, 2018). Each period after detection, cartels are reformed in uncartelized markets with fixed probability $\rho \in [0, 1]$, in line with evidence of limited recidivism (Marvão and Spagnolo, 2018). While our theoretical predictions do not depend on ρ, this approach nests the two assumptions commonly made in the theoretical literature: $\rho = 0$ or $\rho = 1$.

Each period, after prices are set and the market clears, the antitrust authority detects and convicts all active cartels with probability $\alpha \in (0, 1)$. Upon conviction in period t, each cartel member i pays fine $F_{it} = rB_{it}$, where r is the penalty rate and B_{it} is the penalty base. We focus on how different choices for B_{it} affect cartel pricing, cartel incidence, and cartel stability. In theory, cartels can be deterred entirely by ensuring that F_{it} is sufficiently large. For instance, in the spirit of Becker (1968), this can be done by imposing a penalty which ensures that the expected profit of forming a cartel is negative. The starting point of the literature on cartel fines is that complete deterrence is not feasible for several reasons. In particular, the legal principle of proportionality puts a general cap on fines, fines may be limited so as not to discourage efficiency-enhancing communication and cooperation between firms, and bankruptcy concerns put downward pressure on fines in particular instances (Buccirossi and Spagnolo, 2007; Houba et al., 2018).[13]

We follow the literature by assuming that the critical discount factor is a (negative) measure for both cartel incidence and cartel stability. In the absence of side payments, and due to the symmetric nature of firms, we focus on collusive agreements where all firms set

assumption in theoretical work on cartel fines. Indeed, there is ample support from earlier experiments that explicit communication facilitates collusion (e.g., Fonseca and Normann (2012); Freitag et al. (2021); Gomez-Martinez et al. (2016); Andres et al. (2023)). Ultimately, this is an empirical matter. In our experiment, we observe that when firms do not form a cartel, average market prices are close to the one-shot Nash price in all treatments (see Table 3).

[12] See Tirole (1988), pp. 66-67.

[13] To not further burden subjects in what is already a complicated experiment, fines depend on outcomes in the current period only. Note that we do introduce dynamic detection and that the incentives that our different fining bases provide in a static context also exist in a dynamic setting. Introducing a dynamic component to fines substantially complicates the theoretical analysis and is, therefore, typically ignored in the literature. Notable exceptions are in Harrington (2004; 2005). We find no evidence of subjects gaming the system by successfully predicting in which periods detection occurs.

the same price, and share demand. Let π^C denote a firm's single-period before-fine profit in a cartel whose members all set price p^C. F^C represents that firm's concomitant fine upon detection. A firm's expected present value of participating in the cartel and the competitive benchmark are, respectively, given by

$$V^C = \frac{(\pi^C - \alpha F^C) f(\alpha, \delta, \rho)}{1-\delta} \qquad \text{and} \qquad V^N = \frac{\pi^N}{1-\delta}, \tag{1}$$

where

$$f(\alpha, \delta, \rho) = \frac{1-\delta(1-\rho)}{1-\delta(1-\alpha)(1-\rho)} \ \in (0,1]. \tag{2}$$

With perfect recidivism (i.e., if $\rho = 1$), $f(\alpha, \delta, \rho) = 1$, so that $V^C = \frac{\pi^C - \alpha F^C}{1-\delta}$.

Let p^D denote the optimal defection of a cartel member if all other firms set p^C, and its resulting profit and fine π^D and F^D, respectively. If the cartel is convicted immediately after defection, all firms revert to playing the one-shot-Nash price forever. However, if the cartel is not immediately convicted after defection, firms select the price that maximizes their unilateral profit in the one-shot game (taking into account the possibility of being fined if an old cartel is detected), denoted by p^{PD}, in the next period(s) until detection. We denote the concomitant per-firm profit and fine by π^{PD} and F^{PD} respectively. After defection *and* detection, all firms set p^N again. Therefore, the expected present value of defection is given by

$$\begin{aligned} V^D &= \pi^D - \alpha F^D + \alpha\left(\delta\pi^N + \delta^2\pi^N + \ldots\right) \\ &\quad + (1-\alpha)\delta\Big[\pi^{PD} - \alpha F^{PD} + \alpha\left(\delta\pi^N + \delta^2\pi^N + \ldots\right) \\ &\quad + (1-\alpha)\delta\Big\{\pi^{PD} - \alpha F^{PD} + \alpha\left(\delta\pi^N + \delta^2\pi^N + \ldots\right) + \ldots\Big\}\Big] \\ &= \pi^D - \alpha \underbrace{\left(F^D - \frac{\delta}{1-\delta}\pi^N\right)}_{\text{Immediate detection}} + \frac{\delta(1-\alpha)}{1-\delta(1-\alpha)} \underbrace{\left(\pi^{PD} - \alpha\left(F^{PD} - \frac{\delta}{1-\delta}\pi^N\right)\right)}_{\text{Potential future detection}}. \end{aligned}$$

For all fining regimes that we consider, $\pi^{PD} - \alpha F^{PD} = 0$, so that $V^D = \pi^D - \alpha F^D$.

A stable cartel is one where all firms set cartel price p^C, in accordance with the collusive agreement. For stable cartels to be part of a subgame-perfect Nash equilibrium, two conditions must be met. First, the participation condition requires that $V^C \geq V^N$. This condition is always satisfied in our setting as $V^N = 0$, and caps on the maximum fine – such as the legal principle of proportionality – ensure that $\pi^C - \alpha F^C > 0$. Second, the stability condition requires that $V^C \geq V^D$. That is, defecting from the collusive agreement should not increase the expected present value of a firm's payoff stream. Cartel members, therefore, solve

$$\max_{p^C} \quad \pi^C - \alpha F^C \quad \text{s.t.} \quad V^C \geq V^D. \tag{3}$$

We next analyze how basing F^C on either revenue, profit, or the price overcharge (relative to the competitive price) influences cartel pricing and the critical discount factor.

2.2 Revenue-based fines

We implement revenue-based fines by setting the penalty base for a cartel member equal to that firm's revenue: $F_{it} = r_R p_{it} q_{it}$, where r_R is the exogenous penalty rate. We study a revenue base because it is the norm in practice (ICN, 2017). The European Commission, for instance, selects the most recent annual revenue of the product to which the infringement pertains as the fine base.[14] While U.S. guidelines base fines for organizations on the loss caused by the offense and the illegal gains, the guidelines mention that the volume of affected commerce – revenue – should be used instead for price-fixing, bid-rigging, and market allocation agreements (U.S. Sentencing Commission, 2023, p.311).

When fines are based on revenue, let p_R^C denote the price set by a stable cartel, and let δ_R^* denote the critical discount factor above which cartels are stable.

Proposition 1. *If fines are based on revenue,*

- ***i)*** *the price set by a stable cartel exceeds the monopoly price: $p_R^C > p^M(c)$;*
- ***ii)*** *with imperfect recidivism ($\rho \in [0, 1)$), the critical discount factor increases in the probability of discovery α: $\frac{\partial \delta_R^*}{\partial \alpha} > 0$;*
- ***iii)*** *with perfect recidivism ($\rho = 1$), the critical discount factor is independent of the probability of discovery α: $\delta_R^* = \frac{n-1}{n}$.*

Proposition 1 shows that revenue-based fines have a perverse price effect, which has also been established in the existing theoretical literature using similar models (e.g., Bageri et al. (2013); Katsoulacos and Ulph (2013)). The fine acts as a tax on revenue, reducing marginal revenue but leaving marginal cost unaffected. Specifically, as shown in the proof of Proposition 1, the cartel acts like a monopolist in the absence of antitrust, facing marginal cost $\frac{c}{1-\alpha r_R} > c$. Hence, cartel output decreases and the price increases above the monopoly price. While this reduces before-fine profit compared to the monopoly price, it increases expected profit by reducing the fine.

In our model, a firm's best response to all other firms setting p_R^C is not the monopoly price $p^M(c)$. Because defectors can be detected and fined, they set their price to maximize the sum of before-fine profit and the expected fine. Therefore, the optimal defection is to slightly undercut p_R^C and capture the entire market. This increases the defector's before-fine profit and fine n-fold compared to the cartel case. Because defection scales up expected

[14] The relevant annual sales are multiplied by a factor up to 0.3 based on the gravity of the infringement and then adjusted upward, primarily based on the duration of the infringement. Finally, the amount can be increased or decreased based on aggravating factors, mitigating factors, leniency applications, bankruptcy concerns, and out-of-court settlements (European Commission, 2006).

profit by n, this is the only relevant parameter for the critical discount factor in the case of perfect recidivism. With imperfect recidivism, increasing the rate of detection affects V^C more than V^D as, in addition to the higher expected fine, not all cartels immediately reform following detection.

An additional effect of the revenue base is that, following defection, prices remain above the competitive benchmark if previous cartel members can still be detected. Then, since setting p^N results in positive revenue but no profit, firms set price $p_R^{PD} = \frac{c}{1-\alpha r_R} > p^N$ until they are detected, after which they revert back to setting p^N. Finally, in Appendix B.1, we show that when the detection probability is increasing in the market price, a revenue base may still lead to cartel prices above the monopoly level, provided that the detection probability does not increase too strongly with the market price.

2.3 Fines based on profit

We implement profit-based fines by setting the penalty base for a cartel member equal to that firm's profit: $F_{it} = r_\pi(p_{it} - c)q_{it}$, where r_π is the exogenous penalty rate. We study this base as it is arguably a relatively straightforward-to-implement alternative to a revenue base as it has a legal basis – the US guidelines for non-antitrust offenses mention the incremental profit due to the offense ('pecuniary gain') as a base for the fine (U.S. Sentencing Commission, 2023, p.526).

Let p_π^C denote the price set by a stable cartel, and let δ_π^* denote the critical discount factor above which cartels are stable if fines are based on profit.

Proposition 2. *If fines are based on profit,*

- ***i)*** *the price set by a stable cartel equals the monopoly price: $p_\pi^C = p^M(c)$;*
- ***ii)*** *with imperfect recidivism ($\rho \in [0, 1)$) the critical discount factor increases in the probability of discovery α: $\frac{\partial \delta_\pi^*}{\partial \alpha} > 0$;*
- ***iii)*** *with perfect recidivism ($\rho = 1$) the critical discount factor is independent of the probability of discovery α: $\delta_\pi^* = \frac{n-1}{n}$.*

As a profit-based fine acts as a tax on profit, it does not affect the profit-maximizing price, so that the cartel sets the monopoly price. Similar results on cartel pricing in different Bertrand games are in Bageri et al. (2013) and Katsoulacos et al. (2015). The optimal defection, like in the revenue case, is to slightly undercut the cartel. This increases expected profit by a factor n, so the critical discount factor is identical to the revenue-based critical discount factor – only affected by antitrust when there is imperfect recidivism.

Note that in our Bertrand setting, incremental profit and profit bases are equivalent as profits are zero in the absence of a cartel. To investigate the generality of Proposition 2, consider what happens if incremental profit is benchmarked against a but-for price p^{BF} that

results in positive firm-level profit $\pi^{BF} > 0$. It can be shown that a cartel in an incremental profit-based regime still sets p_π^C, but that the critical discount factor is *below* δ_π^* as defecting scales up the fine more than the before-fine profit.[15] That is, when benchmark profits are positive, a profit base has more attractive properties than an incremental profit base as it results in the same cartel price and makes collusion sustainable for a smaller set of discount factors. This is reassuring, as a measure of total profit is relatively easy to obtain, while determining incremental profit requires an estimate of the counterfactual but-for quantity.

2.4 Overcharge-based fines

We implement overcharge-based fines by setting the penalty base for a cartel member equal to the difference between that firm's price and the competitive price: $F_{it} = r_O(p_{it} - p^N)rq$, where r_O is the exogenous penalty rate and rq denotes a reference quantity that the cartel members cannot influence through their behavior. We study this penalty base as Katsoulacos et al. (2015) show that it leads to cartel prices strictly below the monopoly price. Moreover, estimates of cartel overcharges are routinely made in damages litigation, suggesting that such a penalty base could be implementable in practice. Following Katsoulacos et al. (2015), we set the reference quantity equal to the one-shot Nash equilibrium quantity per firm, $q^N = \frac{q(p^N)}{n}$. However, the properties of the overcharge regime do not hinge on the choice of reference quantity, as long as this quantity does not depend on the firms' behavior. This observation is important for the practical applicability of the overcharge regime because calculating the one-shot Nash equilibrium quantity may be challenging in practice.[16]

We stress that overcharge-based fines do not correspond to damages claims. In the U.S., where damages claims are most prevalent, the standard formula for consumer damages in cartel cases is $(p^C - p^N)q^C$ (Harrington, 2014). That is, in practice, only damages for goods that were sold are taken into consideration. The overcharge base replaces the cartel output in the damages claim by some constant that the cartel cannot affect – in our case q^N. This distinction generates the attractive properties of an overcharge-based fine, as a price increase will put upward pressure on both overcharge-based fines and consumer damages on sold goods, but also put downward pressure on damages by reducing the collusive quantity.

Let p_O^C denote the price set by a stable cartel, π_O^C and F_O^C the corresponding firm-level profit and fine, and δ_O^* the critical discount factor above which cartels setting the joint-profit-maximizing price are stable if fines are based on the overcharge.

[15] The cartel price then solves $\arg\max_p (1 - \alpha r_\pi)\frac{(p-c)q(p)}{n} + \alpha r_\pi \pi^{BF}$, and the critical discount factor (for $\rho = 1$) is $\frac{(n-1)(1-\alpha r_\pi)\pi_\pi^C}{n(1-\alpha r_\pi)\pi_\pi^C + \alpha r_\pi \pi^{BF}}$.

[16] Although several jurisdictions mention the overcharge as relevant for determining fines, to our knowledge, neither an overcharge base, nor a fine that depends on q^N, have been implemented in practice. For example, the U.S. fining guidelines mention the overcharge substantially differing from 10 percent as one of the factors determining which fine is selected from the range of possible fines (U.S. Sentencing Commission, 2023, p.312).

Proposition 3. *If fines are based on the overcharge,*

i) *the price set by a stable cartel lies below the monopoly price:* $c \leq p_O^C < p^M(c)$*;*

ii) *the critical discount factor increases with the probability of discovery* α*, regardless of the level of recidivism:* $\frac{\partial \delta_O^*}{\partial \alpha} > 0 \ \forall \rho \in [0,1]$.

An overcharge-based fine reduces the cartel price compared to $p^M(c)$ as it directly targets the distortion created by the cartel: a price above p^N. The only possible way for a cartel to reduce the fine is to lower the cartel price. At $p^M(c)$, a price reduction decreases the expected fine by more than it decreases before-fine profit, thereby increasing expected profit. Proposition 3(i) extends the result of Katsoulacos et al. (2015) to a setting where recidivism is imperfect, defectors can be fined, and cartels formed in previous periods can be detected.

In contrast to the revenue and profit regimes, in the overcharge regime, defecting from the cartel agreement decreases the fine, while still increasing before-fine profit n-fold. This has two effects. First, even with perfect recidivism ($\rho = 1$), the critical discount factor is increasing in both the penalty rate and the detection probability. Second, defection is attractive compared to the other two fining regimes, where defection increases both the before-fine profit and the fine by a factor n.

For discount factors $\bar{\delta}_O < \delta < \delta_O^*$, stable cartels that set the joint-profit-maximizing price do not exist, but stable cartels that set a lower price can be part of a subgame-perfect Nash equilibrium. This price is always strictly below the joint-profit-maximizing price – as discussed in the proof of Proposition 3 – and $\bar{\delta}_O$ is always higher than the critical discount factors of the revenue and profit regimes. We compare joint-profit-maximizing prices and the corresponding critical discount factors across fining regimes in the next subsection.

2.5 Comparison of fining regimes

Propositions 1 to 3 allow us to obtain our main theoretical result:

Proposition 4. *Comparing across fining regimes shows that, regardless of the recidivism level* ρ*,*

i) *The price set by a stable cartel is highest if fines are based on revenue and lowest if fines are based on the overcharge:* $p_R^C > p_\pi^C > p_O^C$.

ii) *The critical discount factor is highest if fines are based on the overcharge:* $\delta_O^* > \delta_R^* = \delta_\pi^*$.

Revenue-based fines incentivize cartels to increase prices above the monopoly price as the slight reduction in before-fine profit is more than offset in expected profit by a lower penalty base. Overcharge-based fines reduce prices compared to the monopoly price, as they directly target the distortion created by the cartel: a price above the competitive price. Profit-based fines leave prices unaffected as they are essentially a proportional tax on firm profit.

Recall that a lower critical discount factor means that collusion is easier in the sense that coordination on the joint-profit-maximizing price is feasible for a larger set of firms' discount factors. Overcharge-based fines always increase the critical discount factor above which cartels are stable compared to the other fining regimes. This effect arises because defectors can be fined. As defecting from a collusive agreement increases before-fine profit and revenue n-fold, defecting increases the fine n-fold in revenue and profit-based regimes, but the fine does not increase in the overcharge regime. As a result, the critical discount factor is always highest in the overcharge regime, even though antitrust can still deter cartels under imperfect recidivism in the revenue and profit regimes.

Consider a non-degenerate distribution of discount factors δ over different markets. The average price – averaged over stable cartels and competitive markets – follows the same ranking as in Proposition 4i). This follows as overcharge-based fines result in the fewest number of stable cartels and the lowest cartel price of all fining regimes. While revenue and profit bases induce identical deterrence, prices of undeterred cartels are higher when fines are based on revenue.

Proposition 4 raises concerns about the current fining practice, as revenue bases are commonly encountered while overcharge bases have yet to be implemented. Why, then, is practice not more aligned with the theory? One potential reason is that the theoretical results are based on a host of assumptions that guarantee a unique subgame-perfect Nash equilibrium.[17] In particular, we assume – as does the broader literature – that firms coordinate on the joint-profit maximum, but coordination on an infinite amount of other prices is also possible.[18] Moreover, the cartel literature routinely assumes that subjects play grim trigger strategies, but in reality, strategies might be more forgiving. Therefore, we conduct a laboratory experiment to test whether subjects select the same equilibria as the theory presumes. In addition, an experiment allows us to randomize the fining regime and accurately track cartel behavior. In contrast, data on discovered cartels suffer sample selection bias, and identifying the cartel's duration and marginal costs is challenging.

Proposition 4's results on pricing extend beyond the cartel framing. To see this, consider an alternative scenario in which n firms in Bertrand competition merge into a monopoly and future entry is barred, with divestiture proceedings brought with probability α each subsequent period and accompanied by a fine of base B. The merged monopolist's pricing problem is identical to that of a stable cartel: under a revenue base, the firm acts as if its marginal cost were $c/(1-\alpha r_R)$ and prices above $p^M(c)$; under a profit base, the firm prices at $p^M(c)$; under an overcharge base, the firm prices below $p^M(c)$. What distinguishes the cartel from the merger case is the incentive-compatibility constraint that disciplines defection within the cartel–and this constraint shapes only the deterrence ranking, not the price ranking. The implication is that the threat of ex-post divestiture accompanied by a

[17] We discuss implementability in Section 5.

[18] "The multiplicity of equilibria is an embarrassment of riches," as famously noted by Tirole (1988).

profit- or overcharge-based fine could complement ex-ante merger review by deterring post-merger price increases.

All of the above has proceeded under the assumption that firms do not take into account the possibility of damages claims, which are an important financial risk of collusion, particularly in jurisdictions like the United States. An important motivation for this choice is keeping the experimental game manageable for our participants. In Appendix B.2, we extend the theoretical model by letting the parties anticipate that firms will have to pay consumers an expected share $r_D > 0$ of total damages $(p_{it} - p^N)q_{it} = (p_{it} - c)q_{it}$. We find that the spread of cartel prices across regimes may increase. Moreover, consumer anticipation of damages compensation puts upward pressure on posted prices, implying that the joint-profit-maximizing cartel price may exceed the monopoly price in all three fining regimes, rather than only under the revenue-based regime. Most importantly, the relative ranking of fining regimes in terms of cartel prices and critical discount factors remains unchanged. Hence, the central theoretical predictions motivating our laboratory experiment continue to hold in the presence of damages claims.[19]

3 Experimental design, procedures, and hypotheses

3.1 Experimental design and procedures

Our experiment tests how the different cartel fining regimes affect prices in the theoretical framework studied in Section 2. Subjects play infinitely repeated homogeneous Bertrand triopoly games.[20] Each period in each treatment follows the timeline displayed in Figure 1. Subjects first vote whether or not to form a cartel and communicate with one another[21] and then set their price unilaterally. Next, the market clears, and cartels are detected and punished with a fixed probability of 0.2. We vary fines across different treatments by basing them either on a firm's revenue, profit, or overcharge. Finally, subjects receive feedback on the prices, fines, and profits. With probability 0.9, the three matched subjects play another period, while with probability 0.1, each subject is re-matched with two new subjects before

[19]The price ranking of Proposition 4 is not specific to Bertrand. For instance, the cartel's joint-profit-maximization problem coincides with that of a symmetric Cournot competition. More generally, the intuition underlying the ranking extends beyond the particular market structure considered here. A full analysis of alternative oligopoly models nevertheless lies beyond the scope of the present paper and constitutes an interesting avenue for future work.

[20]Some authors speak of 'indefinitely' rather than 'infinitely' repeated games. We follow Dal Bó and Fréchette (2018) and use 'infinitely repeated' as a reference to the theoretical framework under consideration rather than a description of the implementation in the laboratory.

[21]Explicit communication is typical of uncovered cartel cases – even duopolies such as vitamin A500 USP and beta-carotene cartels (Marshall and Marx, 2012). Models of collusion routinely assume that communication is required for collusion to emerge (e.g., McCutcheon (1997); Aubert et al. (2006)).

playing the next period. We now explain all phases of a period in more detail.[22]

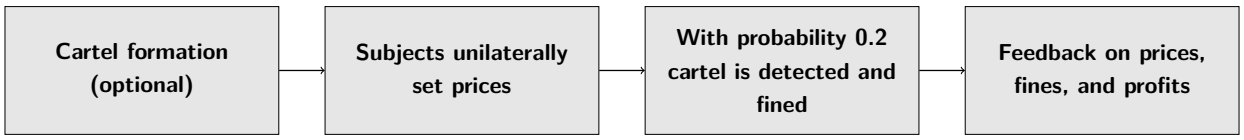


Figure 1: Timeline of a single period

After being matched with two other subjects, each subject unilaterally votes for or against cartel formation. Only if all three subjects vote in favor, a cartel is formed and a chat window becomes available to the subjects. This free-chat is available for 60 seconds in the cartel's first period and 30 seconds in all subsequent periods until the cartel is detected by the antitrust authority or subjects are re-matched.[23] If no cartel is formed, subjects start the next period by voting on cartel formation. We implement communication by using a chat as this facilitates coordination on the joint-profit-maximizing outcome and stable cartels to a much larger extent than restricted communication protocols such as suggesting prices (Cooper and Kühn (2014); Harrington et al. (2016)). In addition, unrestricted communication using natural language is a central feature of discovered hard-core cartels (Genesove and Mullin (2001); Harrington (2006)).

Market demand in period t of the Bertrand triopoly is given by $q(p_t) = 100 - p_t$, and marginal costs equal 47. We opt for a triopoly as tacit collusion is frequently observed in oligopoly experiments with two players (e.g., Huck et al. (2004); Fonseca and Normann (2012)), while without explicit communication, three-player experimental oligopolies tend to produce outcomes close to the one-shot Nash equilibrium (e.g., Chowdhury and Crede (2020); Freitag et al. (2021)). If subjects can earn more by tacitly colluding than by engaging in potentially costly communication, cartels will rarely form, making the study of their behavior challenging.[24] Indeed, in markets where cartels are not formed, market prices are close to the one-shot Nash equilibrium in our experiment.

We believe a homogeneous Bertrand game facilitates subjects' understanding relative to other oligopoly games, thereby limiting noise in the data due to erratic behavior. In addition, a Bertrand setting is standard in the existing theory on cartel fines and is widely used in

[22]We randomly drew supergame length and detection periods once before the first session and used that draw in all subsequent treatments and sessions. As realized supergame length and cartel detection affects behavior in subsequent (super)games (Mengel et al., 2022; Fonseca and Normann, 2014), our approach dampens within-treatment noise allowing us to isolate treatment effects with greater statistical power.

[23]We do not allow for partial cartels as this would further complicate the already challenging decision problem for subjects. Moreover, Clemens and Rau (2022) show that subjects are more likely to form complete than partial cartels when both are part of a Nash equilibrium.

[24]Indeed, even without antitrust Fonseca and Normann (2014) find that more cartels are formed in four-firm experimental oligopolies than in duopolies. The monetary gains from explicit communication appear lowest for Bertrand duopolies (Fonseca and Normann, 2012).

oligopoly experiments (e.g., Dufwenberg and Gneezy (2000); Fonseca and Normann (2012); Hinloopen et al. (2024)). Fines can vary substantially over periods in this setting as defectors capture the entire market. While extreme, a Bertrand game magnifies the incentives that also exist in Cournot games or differentiated goods games, thereby facilitating subjects' understanding. Several authors have instead employed differentiated goods price-setting duopolies when investigating antitrust in the laboratory (e.g., Bigoni et al. (2012, 2015)). While attractive in the duopoly case, differentiated goods price-setting games with more than two players are challenging to implement and place strong demands on experimental subjects.[25] To further aid subjects' understanding, an on-screen profit calculator was made available that displays profits net and gross of the fine after a subject enters hypothetical prices for all three subjects – a screenshot is in Appendix C.

After setting prices, the market clears, and members of active cartels – subjects with access to the chat that period – are discovered and fined with probability 0.2.[26] We follow the norm in the experimental literature by marking a market as a cartel if subjects have opted into the chat, rather than classifying cartels based on chat content. In Section 4.3 we analyze the chat contents to measure agreed-upon prices and cartel stability, and show that our results are not driven by how we classify cartels. After detection, the cartel is disbanded. Subjects can once again vote to reform the cartel from the next period onward.

We implement three treatments. In REVENUE, $F_{it} = p_{it}q_{it}$. In PROFIT, $F_{it} = 2.33(p_{it} - 47)q_{it}$. Finally, in OVERCHARGE, $F_{it} = 1.85(p_{it} - 47)\left(\frac{53}{3}\right)$.[27] Penalty rates are selected to equalize fines across treatments conditional on subjects coordinating on the joint-profit-maximizing price. This ensures that our results are not driven by behavioral responses to the size of the fine and align well with practice, where the principle of proportionality puts a cap on permissible changes of the total fine following penalty base adjustments. We refrain from including a treatment without antitrust: We are interested in studying how the cartel fining regime influences cartel behavior rather than comparing the behavior of legal (or unprosecuted) and illegal cartels. Our results are, therefore, informative for countries with antitrust authorities that enforce a cartel prohibition.[28]

After each period, with probability 0.9, subjects play another period against the same rivals. With probability 0.1, subjects are matched to different subjects before playing the next period. Such random termination, introduced by Roth and Murnighan (1978), is the standard way to implement an infinitely repeated game in the lab (Dal Bó and Fréchette

[25]For instance, Bigoni et al. (2012, 2015) restrict the action space and provide the subjects with payoff tables. Implementing payoff tables with a larger set of actions is difficult.

[26]Estimates of yearly cartel detection lie between 10 and 20 percent (Bryant and Eckard, 1991; Ormosi, 2014). Random draws prior to the first session determined detection, which was identical across sessions.

[27]Subjects see all numbers rounded to two decimal places – 32.69 in this case – and are informed about this rounding in the instructions.

[28]Europe, North America, and many countries in Africa, Asia, Oceania, and South America.

(2018)).[29] This implementation allows a subject to play multiple repeated games – 'supergames' – in one session. Random draws prior to the first session determined that each session consists of four supergames, with, respectively, eight, twelve, seven, and four periods.[30] Subjects could not be matched to the same subject in different supergames (perfect stranger matching), and their payment was based on all periods of play. A random continuation probability, together with a cumulative payment scheme, induces preferences that are theoretically equivalent to maximizing the discounted sum of utilities with discount factor $\delta = 0.9$.[31]

Table 1: Subject and observation count, by treatment

	REVENUE	PROFIT	OVERCHARGE	Total
Subjects	90	99	90	279
Markets	120	132	120	372
Market-periods	930	1,023	930	2,883
Observations	2,790	3,069	2,790	8,649

Notes: Count of subjects, markets, market-periods, and observations, by treatment.

The computerized experiment was conducted at the Center for Research in Experimental Economics and political Decision making (CREED) of the University of Amsterdam in September 2023 using oTree (Chen et al., 2016). Students were recruited by public announcement. In total, 279 students, mainly from the university's undergraduate population, participated across 21 sessions covering the three treatments. Each session had either 9 or 18 participants.[32] We employed a between-subject design – each subject participated in only one treatment. At the start of each session, matching groups of nine subjects were randomly formed. These groups did not change during the sessions. In each supergame, subjects were randomly re-matched to subjects they had never faced before in their matching group. Before the first supergame was played, subjects completed a test measuring their risk attitude, the outcome of which was communicated to them after the final supergame had finished (details are in Appendix D). Table 1 lists the number of subjects, supergames, and observations

[29]Random termination rules are commonly used in oligopoly experiments (e.g., Bigoni et al. (2012, 2015); Fonseca et al. (2022)). Alternatively, a fixed number of periods followed by a random termination rule has been used (e.g., Hinloopen et al. (2020)).

[30]Detection was similarly determined to occur in period six of the first supergame, periods two and ten of supergame two, periods five and six of supergame three, and never in the final supergame.

[31]This theoretical equivalence requires risk neutrality. However, Sherstyuk et al. (2013) provide evidence that subjects' behavior in infinitely repeated games does not change if the payoff scheme is altered to allow for deviations from risk neutrality.

[32]The share of sessions with only 9 subjects was equal across treatments.

across treatments.

Sessions took 70-90 minutes to complete. Excluding seat assignment and payment, sessions lasted on average 57:44 in OVERCHARGE, 1:00:26 in PROFIT, and 1:01:29 in REVENUE. Subjects earned points which were exchanged for euros at the end of the experiment at the rate of 300 points per euro. In addition, subjects received a show-up fee of 7 euros. In the rare occurrence of a loss, subjects were still paid the 7 euro show-up fee.[33] Average earnings were 16.1 euros per subject. To ensure that all subjects understood the experiment, they had to answer several test questions correctly before the experiment started. The instructions and test questions of all treatments are in Appendix C.

3.2 Hypotheses

Table 2 displays the theoretical predictions on prices and critical discount factors which are based on the model in Section 2, the parameters introduced in Section 3.1, and assuming perfect recidivism ($\rho = 1$).[34] Parameters were selected based on simplifying the presentation towards subjects while ensuring that (expected) fines are equalized across treatments and no focal prices emerge that could guide subject behavior.

Table 2: Theoretical predictions, by treatment

	REVENUE	PROFIT	OVERCHARGE
p^C	$\frac{635}{8} \approx 79.38$	73.5	$\frac{482+\sqrt{\frac{1517}{2}}}{8} \approx 63.69$
δ^*	$\frac{2}{3}$	$\frac{2}{3}$	$\frac{2(318-\sqrt{\frac{1517}{2}})}{3(318-\sqrt{\frac{1517}{2}})-2(106-\sqrt{\frac{1517}{2}})} \approx 0.81$

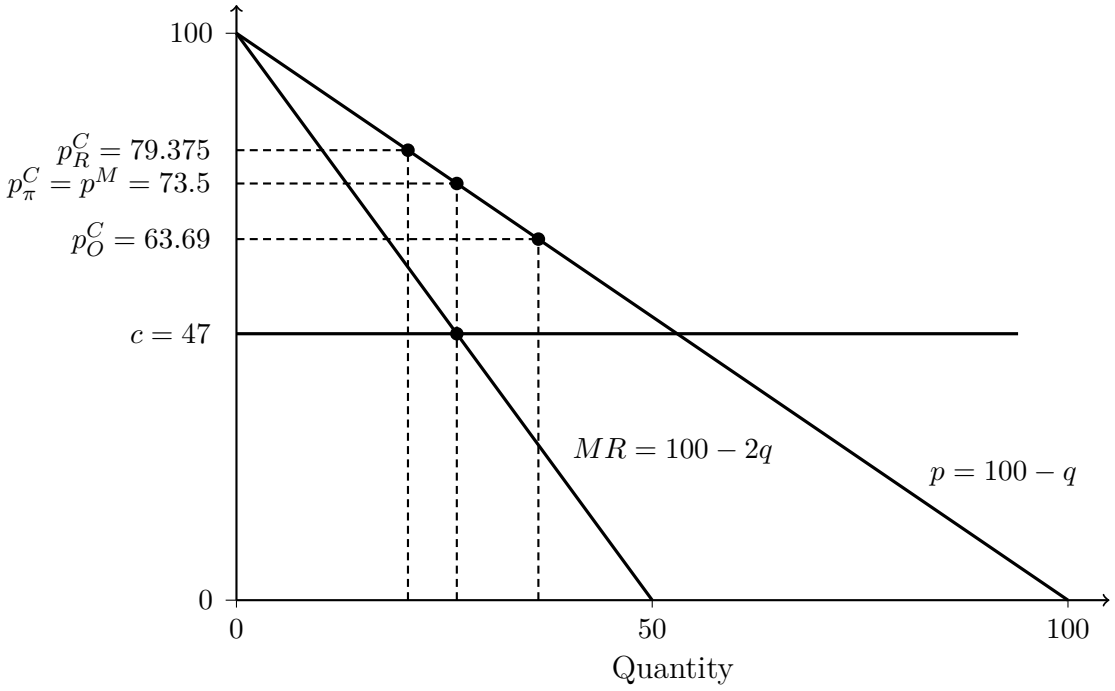
Notes: p^C is the joint-profit-maximizing price set by a stable cartel, and δ^* the critical discount factor above which this price can be part of a subgame-perfect Nash equilibrium, derived in Section 2.

To form intuition, Figure 2 displays the predicted cartel prices. Note that cartels maximize expected profit: before-fine profit minus the expected fine. In PROFIT, the expected fine is a fixed share of before-fine profit, so the price that maximizes expected profit coincides with the price that maximizes before-fine profit: the monopoly price $p^M = 73.5$. In REVENUE, the cartel price exceeds p^M. While this reduces before-fine profit, it also reduces revenue substantially, reducing the expected fine and increasing expected profit compared to setting the monopoly price. In OVERCHARGE, the cartel price lies below p^M. While this reduces before-fine profit, it also reduces the price-cost margin substantially, reducing the expected fine and increasing expected profit compared to setting the monopoly price.

[33] Out of 279 participants, this happened 11 times.

[34] This assumption is regularly made in related work (e.g., Motta and Polo (2003); Chen and Rey (2013)).

Figure 2: Cartel prices under different fining regimes



Notes: The figure plots cartel prices and quantities under the revenue, profit, and overcharge fining regimes for the parameters used in the laboratory experiment.

We test the following hypotheses against the null of no differences between treatments.

H1: Market prices are higher in REVENUE than in PROFIT, and higher in PROFIT than in OVERCHARGE.
H2: Stable cartels – cartels where all members set the agreed-upon price – are equally likely in REVENUE and PROFIT, and less likely in OVERCHARGE than in the other two treatments.

H1 and **H2** are derived from claims (i) and (ii) in Proposition 4, respectively. While the price of uncartelized markets is independent of the fining regime, the price of stable cartels ranks according to **H1**. Therefore, if in all treatments, stable cartels are formed in the same fraction of markets, our theoretical model predicts that market prices follow the ranking in **H1**. Notice that for all treatments, the continuation probability in the experiment – 0.9 – exceeds the critical discount factor, which implies that stable cartels can, in theory, be the norm regardless of fining base. As a result, our theoretical analysis gives rise to the null hypothesis of no differences in cartel incidence and cartel stability across treatments for perfect recidivism ($\rho = 1$). However, for the following reasons, we posit as an alternative hypothesis regarding cartelization that stable cartels are less likely to emerge in OVERCHARGE than in PROFIT and REVENUE. First, for sufficiently low recidivism rates cartels might be stable

in PROFIT and REVENUE, but not in OVERCHARGE. Second, the experimental literature on infinitely repeated games generally finds that the discount factor exceeding the critical discount factor is a necessary, but not sufficient, condition for coordination. Indeed, this literature suggests that subjects are more likely to cooperate as the discount factor increases beyond the critical discount factor (Dal Bó and Fréchette, 2018). **H2** follows as this difference is smallest in OVERCHARGE. Our final motivation behind **H2** is that in OVERCHARGE, the firms face higher strategic risk of cooperation than in the other regimes. The strategic risk is rooted in all firms paying a fine in the OVERCHARGE regime upon detection, even when selling zero quantity after a competitor defects, while only the defector pays a fine in the other two regimes. An established fact from the experimental literature on repeated prisoner's dilemma games (Dal Bó and Fréchette, 2011; Blonski et al., 2011; Dal Bó and Fréchette, 2018) is that increased strategic risk of cooperation negatively affects cooperation rates.

4 Experimental results

In this section, we analyze the data from the experiment. In Section 4.1, we compare REVENUE, PROFIT, and OVERCHARGE in terms of market prices and submitted prices. Section 4.2 presents the relative performance of the three fining regimes in terms of measures of cartel incidence.[35] We show that differences in prices across treatments follow our theoretical predictions and are driven by differences in prices of cartels rather than differences in the prevalence of cartels or prices in uncartelized markets. In Section 4.3, therefore, we use the communication data to show that our aggregate results on pricing originate in the pricing of stable cartels rather than differences in cartel stability.[36]

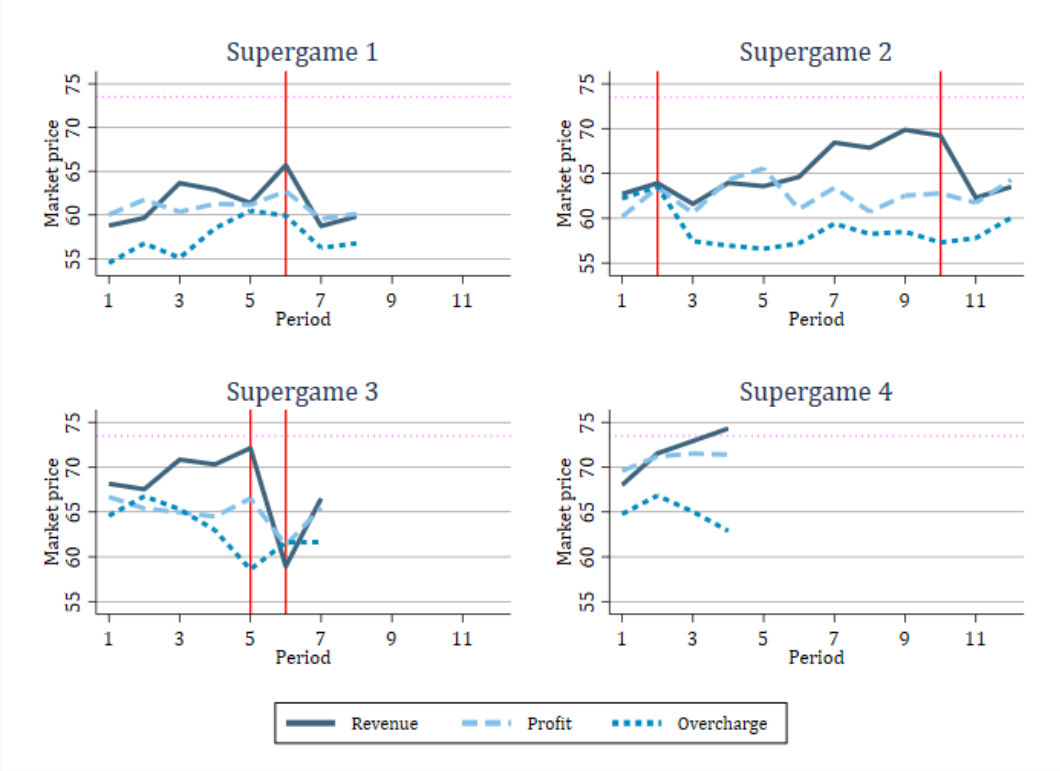
4.1 Prices

Figure 3 plots market prices over time by treatment and supergame, and Table 3 presents the aggregate results on prices across fining regimes. Market prices substantially exceed the one-shot Nash equilibrium price of 47 in all periods of all treatments. Market prices are typically highest in REVENUE (22 out of 31 periods) and lowest in OVERCHARGE (26 out

[35]Throughout this section, a cartel is said to exist in a market-period if the chat is active. This aligns with the experimental literature and legal practice, where explicit attempts to coordinate are typically of central importance (Motta, 2004).

[36]We use the Kruskal-Wallis (KW) test to compare all three treatments, Mann-Whitney U (MWU) test for pairwise comparisons across treatments, and the Wilcoxon signed-rank test for within-treatment comparisons. All tests are two-sided, with the average of a variable within a matching group taken as one independent observation in the non-parametric tests. All main results are robust to using less conservative approaches such as regressions with market, market-period or subject-period level data – depending on the outcome – while clustering the standard errors at the matching-group level.

Figure 3: Market price over time, by treatment and supergame



Notes: Average market price over time, by treatment and supergame. Market price = lowest submitted price in a market-period. Red vertical lines indicate a period at the end of which all cartels are detected. Dotted horizontal line indicates the monopoly price of 73.5.

of 31 periods). With experience, subjects learn to set higher prices. The range of average market prices across treatments at the start of each supergame steadily increases, rising from approximately 55-60 in the first supergame to approximately 65-70 by the final supergame. While market prices tend upward over time, they typically decrease to similar levels in all treatments in the period immediately following cartel detection.

Market prices in REVENUE (65.59) and PROFIT (63.74) are higher than market prices in OVERCHARGE (60.14) ($p = 0.005$ and $p = 0.072$, respectively). The concomitant submitted prices, 68.35, 66.25, and 62.33, compare similarly ($p < 0.001$ and $p = 0.030$, respectively). While both price measures are higher in REVENUE than in PROFIT, these differences are not significant at conventional significance levels. Market prices could differ across treatments due to differences in cartel incidence, cartel prices, and prices in uncartelized markets.

Cartel prices in REVENUE (71.50) exceed those in PROFIT (67.41) and OVERCHARGE (64.60) ($p = 0.030$ and $p < 0.001$, respectively). In line with the theoretical predictions, market prices exceeding the monopoly price are most common when fines are based on

Table 3: Prices, across treatments

	Market price (All markets)	Submitted price (All markets)	Market price (Cartels)	Market price (Competitive)
REVENUE	65.59 (12.97)	68.35 (12.08)	71.50 (9.08)	51.21 (9.21)
	$\vee$	$\vee$	$\vee^{**}$	$\vee$
PROFIT	63.74 (12.50)	66.25 (11.31)	67.41 (10.73)	50.88 (9.36)
	$\vee^{*}$	$\vee^{**}$	$\vee$	$\vee$
OVERCHARGE	60.14 (11.94)	62.33 (11.99)	64.60 (11.01)	48.82 (4.26)
	$\wedge^{***}$	$\wedge^{***}$	$\wedge^{***}$	$\wedge^{*}$
REVENUE	65.59 (12.97)	68.35 (12.08)	71.50 (9.08)	51.21 (9.21)
KW test	$p = 0.023$	$p = 0.004$	$p = 0.002$	$p = 0.197$

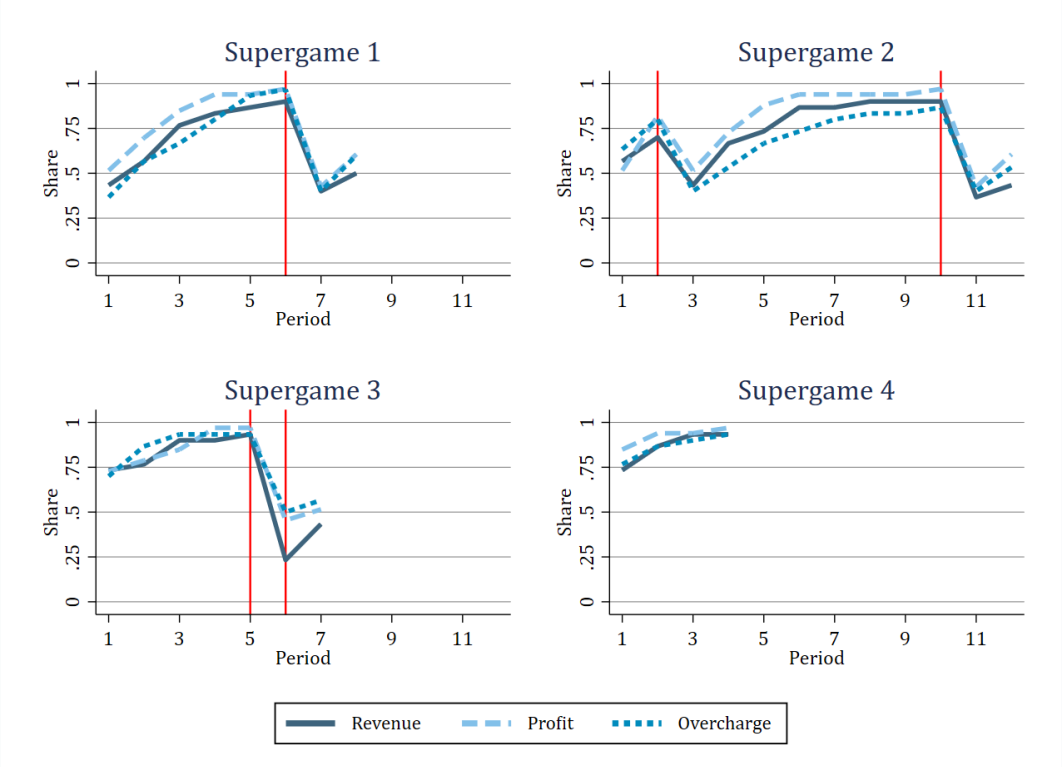
Notes: Average prices per treatment; Market price = lowest submitted price in a market-period; Submitted price = price submitted by a subject in a market-period; Cartels = market-periods with a cartel; Competitive = market-periods without a cartel; Standard deviation in brackets; ***, **, and * indicate statistical significance from pairwise, two-sided MWU tests at the 1%, 5%, and 10% level, respectively; Bottom row reports Kruskal-Wallis p-values across all three treatments.

revenue.[37] Market prices in uncartelized markets lie between 51.21 in REVENUE and 48.82 in OVERCHARGE ($p < 0.001$ for all treatments when compared to 47), only slightly above the one-shot Nash equilibrium price of 47. One implication of these findings is that a discount factor exceeding the critical threshold need not be sufficient for firms to coordinate on high prices. This pattern is consistent with previous work on repeated Bertrand experiments with more than two players, such as Dufwenberg and Gneezy (2000) and Fonseca and Normann (2012). The results are even more in line with the predictions as subjects gain experience. In the last two supergames, cartel prices differ significantly across treatments, whereas none of the differences in prices in uncartelized markets are significant.

Summing up, we conclude that the data are in line with hypothesis **H1**: market prices are highest in REVENUE and lowest in OVERCHARGE. While uncartelized markets yield prices close to the one-shot Nash equilibrium price in all three fining regimes, cartel prices are highest when fines are based on revenue and lowest when they are based on the overcharge. Moreover, market prices above the monopoly price are common when fines are based on revenue. We next turn to an additional factor that might contribute to the observed difference

[37] When fines are based on revenue, 33.98 percent of all market prices exceed the monopoly price of 73.5, significantly more often than 17.11 percent when fines are based on profit and 10.75 percent when fines are based on the overcharge ($p = 0.049$ and $p = 0.008$, respectively; $p = 0.564$ when comparing profit to overcharge bases).

Figure 4: Cartel incidence over time, by treatment and supergame



Notes: Average cartel incidence over time, by treatment and supergame. Cartel incidence = indicator for a cartel in a market-period. Red vertical lines indicate a period at the end of which all cartels are detected.

in market prices across treatments: differences in cartel incidence.

4.2 Cartel incidence

Figure 4 displays cartel incidence over time by treatment and supergame, and Table 4 presents the aggregate results on cartel incidence, as well as the closely related outcome measures cartel formation and cartel recidivism. Cartel incidence follows a near-identical trend over time in the three fining regimes. There is a tendency toward complete cartelization in all supergames – i.e., a tendency for all markets to contain a cartel.[38] As subjects gain experience, cartel incidence in the first period of a supergame increases in all treatments, from roughly 50 percent in the first supergame to about 75 percent in the final supergame.

[38]Recall that once subjects in a given market have agreed to form a cartel, that cartel remains active until it is detected, regardless of the subjects' behavior. This implies that cartel incidence can only decline over time following a period where all cartels are detected. In Section 4.3, we analyze the chat data to classify agreements and cartel stability and conclude that the incidence of agreements to set collusive prices does not significantly differ across treatments.

Detection of cartels causes cartel incidence to decline sharply, often below incidence in the first period of the supergame. However, cartel formation picks up again immediately after detection, suggesting that the effects of detection are short-lived.

There are no statistical differences between cartel incidence in REVENUE (0.71), PROFIT (0.78), and OVERCHARGE (0.72) (p-values lie between 0.433 and 0.986). This suggests that the likelihood that a cartel will be formed in an uncartelized market-period is equal across treatments. Similarly, cartel formation rates do not differ significantly when fines are based on revenue (0.40), profit (0.50), or the price overcharge (0.43) (p-values lie between 0.557 and 0.739). Hence, it is unsurprising that the probability with which a subject votes in favor of cartel formation is very similar across treatments – between 71 and 76 percent across the three fining regimes. These results are unchanged when focusing only on cartel formation in market-periods where detection shut down a cartel in the previous period. Such recidivism averages at 47 percent, ranging from 43 percent in REVENUE, to 46 percent in PROFIT, to 52 percent in OVERCHARGE (p-values lie between 0.243 and 0.959).

Forming a cartel comes with the risk of being fined, so failing to balance subjects' risk preferences across treatments might drive results rather than the fining regime. However, Figure D1 in Appendix D shows that the distribution of elicited risk preferences is highly similar across treatments. Indeed, the average of our risk measure across subjects in REVENUE, PROFIT, and OVERCHARGE does not differ significantly (p-values lie between 0.565 and 0.850). As all but one subject participates in a cartel at some point in the experiment, average differences between the risk preferences of cartel members are also absent. Finally, the average of elicited risk preferences over all cartel observations does not differ significantly between the three treatments (p-values between 0.512 and 0.971), suggesting that there are no between-treatment differences in *when* subjects with a particular attitude towards risk form a cartel. We conclude that risk-preference-based selection into cartels does not differ across treatments.

That our measures of cartel incidence do not differ significantly between treatments is surprising given the experimental literature on infinitely repeated games, as the difference between the discount factor implemented in the laboratory and the critical discount factor is found to be a strong predictor of cooperation (Dal Bó and Fréchette, 2018). However, the results surveyed by Dal Bó and Fréchette (2018) are based on infinitely repeated games without the opportunity to communicate, a setting where even experienced subjects rarely achieve sustained cooperation. Our results suggest that when subjects can use unrestricted communication, sustained cooperation is achieved regardless of the distance between the actual and critical discount factor.[39]

We interpret our results on cartel incidence as aligning with the null hypothesis of no differences rather than hypothesis **H2**: there are no significant treatment differences across treatments in terms of incidence. Together with the fact that market prices in uncartelized

[39]See Boczoń et al. (2025) for a similar finding in repeated multi-player prisoner's dilemmas.

Table 4: Measures of cartelization, across treatments

	Incidence	Formation	Voting	Recidivism
REVENUE	0.71 (0.45)	0.40 (0.49)	0.71 (0.45)	0.43 (0.50)
	$\wedge$	$\wedge$	$\wedge$	$\wedge$
PROFIT	0.78 (0.42)	0.50 (0.50)	0.76 (0.43)	0.46 (0.50)
	$\vee$	$\vee$	$\vee$	$\vee$
OVERCHARGE	0.72 (0.45)	0.43 (0.50)	0.71 (0.45)	0.52 (0.50)
	$\vee$	$\vee$	$=$	$\vee$
REVENUE	0.71 (0.45)	0.40 (0.49)	0.71 (0.45)	0.43 (0.50)
KW test	$p = 0.740$	$p = 0.777$	$p = 0.611$	$p = 0.434$

Notes: Incidence = indicator for a cartel in a market-period; Formation = indicator for cartel formation in a market-period; Voting = indicator for a vote in favor of a cartel in a market-period; Recidivism = indicator for formation of a cartel in a market the period after it has been detected; Standard deviation in brackets; ***, **, and * indicate statistical significance from pairwise, two-sided MWU tests at the 1%, 5%, and 10% level, respectively; Bottom row reports Kruskal-Wallis p-values across all three treatments.

markets do not differ significantly across fining regimes either, this suggests that much of the observed variation in market prices across treatments is rooted in differences between the market prices set by cartels. However, recall that a cartel is said to exist whenever the chat is active. Therefore, to determine the drivers of cartel prices and accurately classify cartel stability, we next turn to the contents of the discussions between cartel members. This allows us to determine whether differences in cartel incidence or cartel stability cause differences in cartel prices across treatments.

4.3 Price agreements and cartel stability

The fact that subjects form cartels and exchange chat messages does not imply that cartel members agree on which prices to set. We, therefore, create two classifications of cartel agreements based on chat data. An explicit price agreement to set price p in a given period is said to exist if at least one subject proposes price p and all other subjects explicitly agree *in that period* before any subject leaves the chat. We also employ a broader definition–'price agreements'–that adds implicit agreements where the context makes it clear that all subjects agree on a particular price. An example is one subject commenting "great! let's keep it going" after several periods of successful coordination, followed by "ok" from the other two subjects. We use the broader measure of price agreements in the main text. A motivation

of this decision as well as a detailed explanation of how chat data were classified, along with illustrative examples, are in Appendix E.[40]

Figure G2 in Appendix G reveals no apparent differences in price agreement incidence between treatments. Moreover, the shares of cartelized market-periods with price agreements in REVENUE (0.73), PROFIT (0.72), and OVERCHARGE (0.76) do not differ significantly, as is the case for the average risk aversion in market-periods with price agreements.[41] Therefore, differences in cartel pricing, which determines our aggregate results on market prices, are likely to be explained by the behavior of cartels with price agreements, and such cross-treatment differences are unlikely to be driven by selection on risk preference.

While price agreements are the norm, note that in cartelized market-periods without price agreements, market prices are higher when fines are based on revenue (64.33) than when based on profit (55.19) or the overcharge (54.88) ($p = 0.024$ and $p = 0.002$, respectively). Recall that, according to theory, when fines are based on revenue, firms that remain detectable but no longer coordinate with others set price $p_R^{PD} = \frac{c}{1-\alpha r_R} = 58.75$. Hence, one explanation for our aggregate results–albeit minor, given price agreements' prominence–is that revenue-based fines lead to higher prices even when cartels fail to reach an agreement.[42]

Figure 5 displays the prices that cartel members agree to set over time, by treatment and supergame. Table 5 presents our aggregate results on price agreements and cartel stability. Price agreements in REVENUE (75.05) are significantly higher than those in PROFIT (72.79) and OVERCHARGE (68.40) ($p = 0.005$ and $p < 0.001$, respectively), as well as the monopoly price of 73.5 ($p < 0.001$, one-sided t-test). When fines are based on profit, price agreements are higher than when fines are based on the overcharge ($p = 0.001$). The ranking of price agreements, therefore, is in accordance with the theoretical predictions.

The degree to which price agreements translate into market prices depends on cartel stability. A cartel is said to be stable if all three subjects set the agreed-upon price. The fraction of price agreements adhered to is high in all treatments, ranging from 0.80 in REVENUE to 0.89 in OVERCHARGE ($p = 0.052$). Over time, the fraction of cartels that are stable tends towards 1 (see Figure G3 in Appendix G). This suggests that our aggregate results on market prices are primarily driven by the prices set by stable cartels.[43] Indeed, market prices

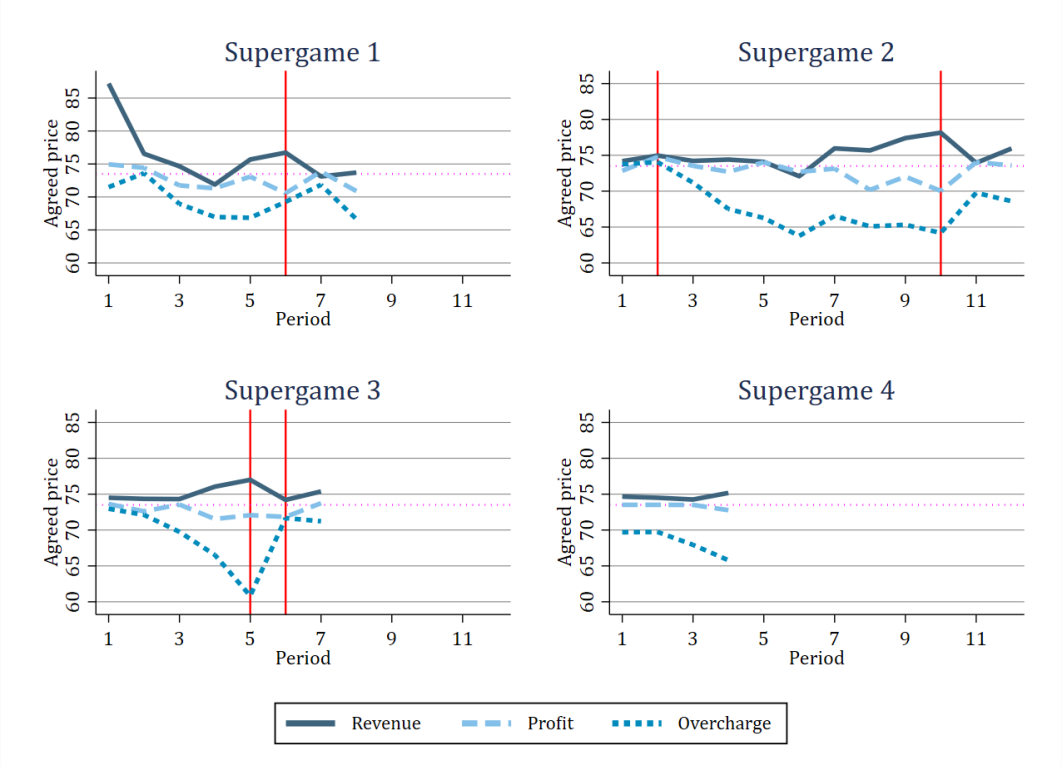
[40] An alternative recent approach is to employ statistical classification techniques to analyze chat data (Andres et al., 2023). We refrain from this as chat contents are often rather sparse, particularly after several periods of stable collusion.

[41] Comparisons of price agreement incidence result in p-values between 0.796 and 0.989, while p-values are between 0.314 and 0.912 for pairwise comparisons across treatments of average risk aversion in market-periods with price agreements.

[42] This is unlikely to result from our cartel agreement categorization being too conservative, as REVENUE cartels without an agreement only coordinate on prices above 47 in 10 of 124 market-periods (8.06 percent), less than the incidence of such coordination in the absence of cartels (10.78 percent). Learning effects are also unlikely, as the result persists in the final two supergames.

[43] Because defections become increasingly rare over time, we do not focus on them in the main text. Appendix F provides more details. Defections tend to occur in the first period in which a cartel agreement

Figure 5: Price agreements over time, by treatment and supergame



Notes: Average price agreement over time, by treatment and supergame. Price agreement = price that a cartel agrees to set in a market-period. Red vertical lines indicate a period at the end of which all cartels are detected. Dotted horizontal line indicates the monopoly price of 73.5.

in market-periods with a price agreement in place in REVENUE (74.20) are significantly higher than those in PROFIT (72.10) and OVERCHARGE (67.60) ($p = 0.013$ and $p < 0.001$, respectively), and market prices in PROFIT exceed those in OVERCHARGE ($p = 0.005$).

Price agreements above the monopoly price of 73.5 are common in REVENUE, covering 281 of 479 market-periods with a price agreement (58.7 percent). While price agreements and the fraction of above-monopoly agreements are stable over time, the standard deviation of agreements in REVENUE drops from 6.69 in the first two supergames to 3.40 in the final two. Moreover, 11.3 percent of price agreements in REVENUE fall between 79 and 80, a share stable over the supergames. Subjects appear to converge on a price between the monopoly price and the joint-profit-maximizing price.

Price agreements converge to the jointly optimal monopoly price in PROFIT. From Figure 5, it is clear that the average price agreement is stable in PROFIT and close to 73.5 in all supergames. This average masks significant learning over time. In the first two supergames,

is reached. While a sizable minority of subjects respond in line with a trigger strategy, price agreements are frequently renegotiated and re-established after defection.

Table 5: Price agreements and cartel stability, across treatments

	Agreement incidence (Cartels)	Price agreement (Cartels with a price agreement in place)	Cartel stability (Cartels with a price agreement in place)	Market price (Cartels with a price agreement in place)
REVENUE	0.73 (0.45)	75.05 (5.49)	0.83 (0.38)	74.20 (6.07)
	$\vee$	$\vee^{***}$	$\wedge$	$\vee^{**}$
PROFIT	0.72 (0.45)	72.79 (4.95)	0.86 (0.35)	72.10 (5.63)
	$\wedge$	$\vee^{***}$	$\wedge$	$\vee^{***}$
OVERCHARGE	0.76 (0.42)	68.40 (9.20)	0.89 (0.31)	67.60 (9.48)
	$\vee$	$\wedge^{***}$	$\vee$	$\wedge^{***}$
REVENUE	0.73 (0.45)	75.05 (5.49)	0.83 (0.38)	74.20 (6.07)
KW test	$p = 0.968$	$p = 0.000$	$p = 0.220$	$p = 0.000$

Notes: Agreement incidence = Indicator for a cartel with a price agreement in a market-period; Price agreement = Price that the cartel has agreed to set in a market-period; Cartel stability = Indicator for whether all three subjects in a cartel have set the agreed upon price in a market-period; Market price = Lowest submitted price in a market-period; Standard deviation in brackets; ***, **, and * indicate statistical significance from pairwise, two-sided MWU tests at the 1%, 5%, and 10% level, respectively; Bottom row reports Kruskal-Wallis p-values across all three treatments.

price agreements between 73 and 74 characterize 210 of 313 market-periods with an agreement (67.1 percent). This percentage increases to 94.2 in the final two supergames (247 of 262 instances), and the standard deviation of price agreements decreases from 5.94 in the first two supergames to 3.41 in the final two supergames.

While price agreements lie substantially below the monopoly price in OVERCHARGE, most are above the theoretical prediction of 63.69 (381 out of 510). Out of the 129 agreements below 64, 74 come in the first two supergames and only one is between 63 and 64. So, while low prices are not uncommon, we do not observe convergence to the theoretical prediction in overcharge. The standard deviation of price agreements is 9.66 in the first two supergames of OVERCHARGE, and 8.66 in the final two supergames – more than 2.5 times that of the other two treatments – indicating much more limited convergence than in REVENUE or PROFIT. Therefore, the theoretical predictions for OVERCHARGE hold only qualitatively.

Figure G5 in Appendix G displays the distributions of agreed-upon market prices in the fourth and final supergame for REVENUE and OVERCHARGE. For PROFIT, the distribution is not shown as convergence has occurred – in 111 of 114 cases is the market price between 73 and 74. In REVENUE, about 10 percent of the cases are centered around the joint-profit-maximum of 79.38, while the remainder of the mass is mostly close to, but above, 73.5. For

OVERCHARGE, agreed-upon market prices are much more spread out, although most of the mass is at 65 and above – the 25th percentile is 65. Although all these agreements can be rationalized as a subgame-perfect Nash equilibrium, it is perhaps surprising that a variety of agreements persists even in the fourth supergame, particularly in OVERCHARGE.

Figure 5 suggests one possible explanation for lack of full convergence in OVERCHARGE and REVENUE. As cartels go undetected in a given supergame, price agreements tend downward in OVERCHARGE and upward in REVENUE. As these patterns are observable in all supergames, learning is an unlikely explanation. Likewise, it seems unlikely that subjects can predict when detection occurs, for instance from talking to each other outside of the laboratory, as then price agreements would change sharply before a detection period.[44] To investigate these patterns, we run a number of regressions using data from the final three supergames (to exclude learning as an explanation), reported in Tables H1 and H2 in Appendix H.

Price agreements move closer to theoretical predictions the longer cartels remain undetected when fines are based on either revenue or overcharge. Moreover, greater risk appetite is associated with higher agreements in OVERCHARGE and lower agreements in REVENUE, although the latter correlation is not statistically significant. This suggests that some subjects initially under-assess detection risk, agree on prices near the monopoly level, and revise agreements toward theoretical predictions as non-detection persists. This may explain why full convergence is limited overall, despite appearing within cartels that remain undetected for several periods. Chat evidence supports this interpretation (Table E3 in Appendix E).

Our findings are robust to alternative cartel classifications. Table H3 in Appendix H shows similar results when using explicit price agreements rather than our broader measure. As an additional check, we exclude chat data and focus only on cartel-periods in which members coordinate on prices above marginal cost. Although this approach cannot define agreements and stability, it likely captures stable cartel behavior closely, since coordination is rare without a cartel and tacit collusion appears unlikely (see Figure G4 in Appendix G). Table H4 in Appendix H shows that such above-cost coordination occurs equally often across treatments, while cartel prices follow the same ranking as in Table 5.

5 Concluding remarks

Using a laboratory experiment, we have investigated the relative performance of three bases for cartel fines: revenue, profit, and price overcharge. While we observe no significant differences across treatments regarding cartel incidence and cartel stability, we find that average

[44]The 50 chat messages (out of 11,008) containing statements that anticipate detection are not unusually concentrated in the periods in which cartels are detected: only 6 of the 50 flagged messages occur in those periods, which is not significantly more than in other periods ($p = 0.610$, exact randomization test).

prices are lowest under overcharge-based fines and highest under revenue-based fines. Notably, subjects in the experiment frequently agree to set prices above the monopoly price when fines are based on revenue. The behavior of subjects in the lab aligns well with our theoretical assumptions on equilibrium selection.

The main policy implication of our study is to move away from revenue-based antitrust fines. While active antitrust enforcement can temporarily reduce the negative effects of revenue-based fines by dismantling cartels, our experiment suggests this effect might be short-lived as cartels tend to be re-established regardless of how often detection has previously occurred. The preferred alternative also depends on the feasibility of implementation. Indeed, a primary argument for using a revenue-based regime is ease of implementation.

The overcharge regime may be challenging to implement in legal practice because determining the overcharge requires an estimate of the counterfactual price.[45] A clear downside of such calculations is the lengthy duration of damage cases, which could substantially delay the first monetary penalty for firms engaged in price-fixing. Moreover, in complex cases where cartels coordinate on factors other than the price, and in jurisdictions where indirect purchasers have standing to sue for damages, it is not always clear that the price overcharge accurately reflects actual damages. An additional concern with the overcharge regime is that it may encourage firms to maintain elevated prices after detection, as these prices could influence the estimation of the but-for price on which the fine is based.

A profit-based fine is, therefore, a particularly attractive alternative. Such measures should be relatively straightforward to obtain in a reasonable timeframe – certainly compared to estimating the overcharge – and can also be applied when cartel members collude on factors other than prices. Since the overcharge-based and, particularly, profit-based regimes appear feasible alternatives, we conclude that there is little justification for continuing to base fines on revenue. These findings may have implications that extend beyond cartel enforcement, because the pricing effect of the fine base is not specific to cartels. For instance, if merged firms face a risk of divestiture proceedings, monetary sanctions accompanying such proceedings should be calibrated to the merged firm's overcharge or profits rather than to its revenues.

Policymakers may question whether our experimental results generalize to practice. Economic laboratory experiments frequently replicate in the field and with professional participants rather than students (Camerer, 2015; Fréchette, 2015). Notably, Dal Bó and Fréchette (2018, p.88) state: "*There is no robust evidence that risk aversion, economic training, altruism, gender, intelligence, patience, or psychological traits have a systematic effect on*

[45] Katsoulacos et al. (2019) argue that an overcharge-based regime is feasible as counterfactual price calculations are routinely made in cartel damage cases. If fines are accompanied by an estimate of the counterfactual price that can be used in future damage cases, the uncertainty and cost of such cases would substantially drop. Rüggeberg and Schinkel (2006) previously argued in favor of consolidating damage estimation for such reasons.

cooperation in infinitely repeated games . . . there is evidence consistent with the idea that the main motivation behind cooperation is strategic." Subjects' misunderstanding of probabilities might disappear if stakes were increased to match those in reality, although existing work suggests that cognitive biases persist even at high stakes (Enke et al., 2023). Lab experiments are particularly valuable in our case due to the limitations of observational data on cartels. We contribute to a long tradition of testing theory in the lab, the merits of which have been widely recognized (e.g., Falk and Heckman (2009); List (2020)).

Finally, we have abstracted from leniency programs and damages claims. Incorporating such factors would not eliminate the underlying forces driving our results, as long as fines are imposed with positive probability. In fact, under revenue-based fines, the possibility of future damages claims can further exacerbate the distortions created by revenue-based penalties (Katsoulacos et al., 2020), as we discussed in Appendix B.2. Integrating leniency programs and damages claims into our experiment is a promising avenue for future work, although it should be noted that even when holding the fining base fixed, incorporating damages claims and leniency results in a challenging-to-implement experiment (e.g., Bodnar et al. (2023)).

References

Andres, M., Bruttel, L., and Friedrichsen, J. (2021). The leniency rule revisited: Experiments on cartel formation with open communication. *International Journal of Industrial Organization*, 76:102728.

Andres, M., Bruttel, L., and Friedrichsen, J. (2023). How communication makes the difference between a cartel and tacit collusion: A machine learning approach. *European Economic Review*, 152:104331.

Apesteguia, J., Dufwenberg, M., and Selten, R. (2007). Blowing the whistle. *Economic Theory*, 31:143–166.

Aubert, C., Rey, P., and Kovacic, W. E. (2006). The impact of leniency and whistle-blowing programs on cartels. *International Journal of Industrial Organization*, 24(6):1241–1266.

Awaya, Y. and Krishna, V. (2016). On communication and collusion. *American Economic Review*, 106(2):285–315.

Bageri, V., Katsoulacos, Y., and Spagnolo, G. (2013). The distortive effects of antitrust fines based on revenue. *Economic Journal*, 123(572):F545–F557.

Becker, G. (1968). Crime and punishment: An economic approach. *Journal of Political Economy*, 75(2):169–217.

Bigoni, M., Fridolfsson, S.-O., Le Coq, C., and Spagnolo, G. (2012). Fines, leniency, and rewards in antitrust. *RAND Journal of Economics*, 43(2):368–390.

Bigoni, M., Fridolfsson, S.-O., Le Coq, C., and Spagnolo, G. (2015). Trust, leniency, and deterrence. *Journal of Law, Economics, and Organization*, 31(4):663–689.

Block, M. K., Nold, F. C., and Sidak, J. G. (1981). The deterrent effect of antitrust enforcement. *Journal of Political Economy*, 89(3):429–445.

Blonski, M., Ockenfels, P., and Spagnolo, G. (2011). Equilibrium selection in the repeated prisoner's dilemma: Axiomatic approach and experimental evidence. *American Economic Journal: Microeconomics*, 3(3):164–192.

Boczoń, M., Vespa, E., Weidman, T., and Wilson, A. J. (2025). Testing models of strategic uncertainty: Equilibrium selection in repeated games. *Journal of the European Economic Association*, 23(2):784–814.

Bodnar, O., Fremerey, M., Normann, H.-T., and Schad, J. (2023). The effects of private damage claims on cartel activity: Experimental evidence. *Journal of Law, Economics, and Organization*, 39(1):27–76.

Bruttel, L. V. (2009). The critical discount factor as a measure for cartel stability? *Journal of Economics*, 96:113–136.

Bryant, P. G. and Eckard, E. W. (1991). Price fixing: The probability of getting caught. *Review of Economics and Statistics*, 73(3):531–536.

Buccirossi, P. and Spagnolo, G. (2007). Optimal fines in the era whistleblowers: Should price fixers still go to prison? In Ghosal, V. and Stennek, J., editors, *The Political Economy of Antitrust*. Elsevier.

Byrne, D. P. and De Roos, N. (2019). Learning to coordinate: A study in retail gasoline. *American Economic Review*, 109(2):591–619.

Camerer, C. F. (2015). The promise and success of lab–field generalizability in experimental economics: A critical reply to Levitt and List. In Fréchette, G. and Schotter, A., editors, *Handbook of Experimental Economic Methodology*. Oxford University Press.

Chen, D. L., Schonger, M., and Wickens, C. (2016). oTree—An open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, 9:88–97.

Chen, Z. and Rey, P. (2013). On the design of leniency programs. *Journal of Law and Economics*, 56(4):917–957.

Chowdhury, S. M. and Crede, C. J. (2020). Post-cartel tacit collusion: Determinants, consequences, and prevention. *International Journal of Industrial Organization*, 70:102590.

Clemens, G. and Rau, H. A. (2022). Either with us or against us: Experimental evidence on partial cartels. *Theory and Decision*, 93(2):237–257.

Cooper, D. J. and Kühn, K.-U. (2014). Communication, renegotiation, and the scope for collusion. *American Economic Journal: Microeconomics*, 6(2):247–78.

Dal Bó, P. and Fréchette, G. R. (2011). The evolution of cooperation in infinitely repeated games: Experimental evidence. *American Economic Review*, 101(1):411–29.

Dal Bó, P. and Fréchette, G. R. (2018). On the determinants of cooperation in infinitely repeated games: A survey. *Journal of Economic Literature*, 56(1):60–114.

Dufwenberg, M. and Gneezy, U. (2000). Price competition and market concentration: An experimental study. *International Journal of Industrial Organization*, 18(1):7–22.

Duso, T., Röller, L.-H., and Seldeslachts, J. (2014). Collusion through joint R&D: An empirical assessment. *Review of Economics and Statistics*, 96(2):349–370.

Enke, B., Gneezy, U., Hall, B., Martin, D., Nelidov, V., Offerman, T., and Van De Ven, J. (2023). Cognitive biases: Mistakes or missing stakes? *Review of Economics and Statistics*, 105(4):818–832.

European Commission (2006). Guidelines on the method of setting fines imposed pursuant to Article 23(2)(a) of Regulation No 1/2003. *Official Journal of the European Union*, September 1.

Falk, A. and Heckman, J. J. (2009). Lab experiments are a major source of knowledge in the social sciences. *Science*, 326(5952):535–538.

Fonseca, M. A., Gonçalves, R., Pinho, J., and Tabacco, G. A. (2022). How do antitrust regimes impact on cartel formation and managers' labor market? An experiment. *Journal of Economic Behavior & Organization*, pages 643–662.

Fonseca, M. A. and Normann, H.-T. (2012). Explicit vs. tacit collusion—The impact of communication in oligopoly experiments. *European Economic Review*, 56(8):1759–1772.

Fonseca, M. A. and Normann, H.-T. (2014). Endogenous cartel formation: Experimental evidence. *Economics Letters*, 125(2):223–225.

Freitag, A., Roux, C., and Thöni, C. (2021). Communication and market sharing: An experiment on the exchange of soft and hard information. *International Economic Review*, 62(1):175–198.

Fréchette, G. (2015). Laboratory experiments: Professionals versus students. In Fréchette, G. and Schotter, A., editors, *Handbook of Experimental Economic Methodology*. Oxford University Press.

Genesove, D. and Mullin, W. P. (2001). Rules, communication, and collusion: Narrative evidence from the sugar institute case. *American Economic Review*, 91(3):379–398.

Gomez-Martinez, F., Onderstal, S., and Sonnemans, J. (2016). Firm-specific information and explicit collusion in experimental oligopolies. *European Economic Review*, 82:132–141.

Harrington Jr., J. E. (2004). Cartel pricing dynamics in the presence of an antitrust authority. *RAND Journal of Economics*, 35(4):651–673.

Harrington Jr., J. E. (2005). Optimal cartel pricing in the presence of an antitrust authority. *International Economic Review*, 46(1):145–169.

Harrington Jr, J. E. (2006). How do cartels operate? *Foundations and Trends in Microeconomics*, 2(1):1–105.

Harrington Jr., J. E. (2014). Penalties and the deterrence of unlawful collusion. *Economics Letters*, 124(1):33–36.

Harrington Jr, J. E., Gonzalez, R. H., and Kujal, P. (2016). The relative efficacy of price announcements and express communication for collusion: Experimental findings. *Journal of Economic Behavior & Organization*, 128:251–264.

Hinloopen, J., Martin, S., Onderstal, S., and Treuren, L. (2024). Spillovers from legal cooperation to non-competitive prices. *Tinbergen Institute Discussion Paper 2024-078/VII*.

Hinloopen, J., Onderstal, S., and Soetevent, A. (2023). Corporate leniency programs for antitrust: Past, present, and future. *Review of Industrial Organization*, 63(2):111–122.

Hinloopen, J., Onderstal, S., and Treuren, L. (2020). Cartel stability in experimental first-price sealed-bid and English auctions. *International Journal of Industrial Organization*, 71:102642.

Hinloopen, J. and Soetevent, A. R. (2008). Laboratory evidence on the effectiveness of corporate leniency programs. *RAND Journal of Economics*, 39(2):607–616.

Holt, C. A. and Laury, S. K. (2002). Risk aversion and incentive effects. *American Economic Review*, 92(5):1644–1655.

Houba, H., Motchenkova, E., and Wen, Q. (2018). Legal principles in antitrust enforcement. *Scandinavian Journal of Economics*, 120(3):859–893.

Huck, S., Normann, H.-T., and Oechssler, J. (1999). Learning in Cournot oligopoly–An experiment. *Economic Journal*, 109(454):80–95.

Huck, S., Normann, H.-T., and Oechssler, J. (2000). Does information about competitors' actions increase or decrease competition in experimental oligopoly markets? *International Journal of Industrial Organization*, 18(1):39–57.

Huck, S., Normann, H.-T., and Oechssler, J. (2004). Two are few and four are many: Number effects in experimental oligopolies. *Journal of Economic Behavior & Organization*, 53(4):435–446.

International Competition Network (2017). Setting of fines for cartels in ICN jurisdiction. *Report to the 16th International Competition Network Annual Conference*.

Kandori, M. and Matsushima, H. (1998). Private observation, communication and collusion. *Econometrica*, 66(3):627–652.

Katsoulacos, Y., Motchenkova, E., and Ulph, D. (2015). Penalizing cartels: The case for basing penalties on price overcharge. *International Journal of Industrial Organization*, 42:70–80.

Katsoulacos, Y., Motchenkova, E., and Ulph, D. (2019). Penalizing cartels—a spectrum of regimes. *Journal of Antitrust Enforcement*, 7(3):339–351.

Katsoulacos, Y., Motchenkova, E., and Ulph, D. (2020). Combining cartel penalties and private damage actions: The impact on cartel prices. *International Journal of Industrial Organization*, 73:102604.

Katsoulacos, Y. and Ulph, D. (2013). Antitrust penalties and the implications of empirical evidence on cartel overcharges. *Economic Journal*, 123(572):F558–F581.

Kwoka, J. E. and White, L. J., editors (2018). *The Antitrust Revolution: Economics, Competition, and Policy (7th edition)*. Oxford University Press.

List, J. A. (2020). Non est disputandum de generalizability? A glimpse into the external validity trial. NBER Working Paper 27535, National Bureau of Economic Research, Cambridge, MA.

Marshall, R. C. and Marx, L. M. (2012). *The Economics of Collusion: Cartels and Bidding Rings*. MIT Press.

Marvão, C. and Spagnolo, G. (2018). Cartels and leniency: Taking stock of what we learnt. In Corchón, L. and Marini, M., editors, *Handbook of Game Theory and Industrial Organization, Volume II*. Edward Elgar Publishing.

McCutcheon, B. (1997). Do meetings in smoke-filled rooms facilitate collusion? *Journal of Political Economy*, 105(2):330–350.

Mengel, F., Orlandi, L., and Weidenholzer, S. (2022). Match length realization and cooperation in indefinitely repeated games. *Journal of Economic Theory*, 200:105416.

Motta, M. (2004). *Competition Policy: Theory and Practice*. Cambridge University Press.

Motta, M. and Polo, M. (2003). Leniency programs and cartel prosecution. *International Journal of Industrial Organization*, 21(3):347–379.

Normann, H.-T., Rösch, J., and Schultz, L. M. (2015). Do buyer groups facilitate collusion? *Journal of Economic Behavior & Organization*, 109:72–84.

Offerman, T., Potters, J., and Sonnemans, J. (2002). Imitation and belief learning in an oligopoly experiment. *Review of Economic Studies*, 69(4):973–997.

Ormosi, P. L. (2014). A tip of the iceberg? The probability of catching cartels. *Journal of Applied Econometrics*, 29(4):549–566.

Roth, A. E. and Murnighan, J. K. (1978). Equilibrium behavior and repeated play of the prisoner's dilemma. *Journal of Mathematical Psychology*, 17(2):189–198.

Rüggeberg, J. and Schinkel, M. P. (2006). Consolidating antitrust damages in Europe: A proposal for standing in line with efficient private enforcement. *World Competition*, 29(3):295–420.

Sherstyuk, K., Tarui, N., and Saijo, T. (2013). Payment schemes in infinite-horizon experimental games. *Experimental Economics*, 16:125–153.

Sovinsky, M. (2022). Do research joint ventures serve a collusive function? *Journal of the European Economic Association*, 20(1):430–475.

Tirole, J. (1988). *The Theory of Industrial Organization*. MIT press.

United States Sentencing Commission (2023). *Guidelines Manual §2R1.1*. Washington, D.C.

Vega-Redondo, F. (1997). The evolution of Walrasian behavior. *Econometrica*, 65(2):375–384.

Appendices

A Proofs of propositions

Proof of Proposition 1

i) Assume that the stability condition holds. The cartel's price is then given by

$$p_R^C = \arg\max_p (1-\alpha r_R) q(p) \left(p - \frac{c}{1-\alpha r_R}\right) = p^M\left(\frac{c}{1-\alpha r_R}\right) > p^M(c).$$

The inequality in the above chain follows because p^M is a strictly increasing function. Notice that the cartel acts like a monopolist in the absence of antitrust facing marginal cost $\frac{c}{1-\alpha r_R}$ rather than c. We denote the corresponding firm-level profit and fine by π_R^C and F_R^C respectively.

ii)-iii) As firms facing a revenue-based fine act as if they have marginal cost $\frac{c}{1-\alpha r_R}$ and face no threat of fines, the optimal defection is to slightly undercut p_R^C and capture the entire market. This increases both the defector's before-fine profit π_R^D and its fine F_R^D n-fold compared to the cartel case: $\pi_R^D = n\pi_R^C$ and $F_R^D = nF_R^C$. Post-defection, if the cartel has not been convicted yet, the possibility of being fined implies that expected profit is negative if all firms set price equal to marginal cost. Instead, the unique symmetric pure-strategy one-shot Nash equilibrium market price equals $p_R^{PD} = \frac{c}{1-\alpha r_R}$. As before, the revenue-based fine incentivizes firms to act like firms facing marginal cost $\frac{c}{1-\alpha r_R}$ in the absence of antitrust enforcement. Expected profit is 0 if all firms set p_R^{PD}, $\pi_R^{PD} - \alpha F_R^{PD} = 0$, and as $\pi^N = 0$, the expected net present value of deviation equals $V_R^D = n\left(\pi_R^C - \alpha F_R^C\right)$.

The critical discount factor, δ_R^*, is the δ such that $\sigma_R(\alpha,\delta,\rho) = 1$, where $\sigma_R(\alpha,\delta,\rho) \equiv \frac{V_R^C}{V_R^D} = \frac{f(\alpha,\delta,\rho)}{n(1-\delta)}$, V_R^C is given in equation (1) and $f(\alpha,\delta,\rho)$ in equation (2). Let $\sigma_R^* \equiv \sigma_R(\alpha,\delta_R^*,\rho)$. By the implicit function theorem, we have

$$\frac{\partial \delta_R^*}{\partial \alpha} = -\frac{\partial \sigma_R^*/\partial \alpha}{\partial \sigma_R^*/\partial \delta}.$$

Since $\frac{\partial f(\alpha,\delta,\rho)}{\partial \delta} = \frac{-\alpha(1-\rho)}{(1-\delta(1-\alpha)(1-\rho))^2}$, $\frac{\partial f(\alpha,\delta,\rho)}{\partial \alpha} < 0$ if $\rho \in [0,1)$, and $\frac{\partial f(\alpha,\delta,\rho)}{\partial \alpha} = 0$ if $\rho = 1$, it follows that

$$\frac{\partial \sigma_R}{\partial \delta} = \frac{\partial f(\alpha,\delta,\rho)/\partial \delta}{n(1-\delta)} + \frac{f(\alpha,\delta,\rho)}{n(1-\delta)^2} > 0 \quad \text{and} \quad \frac{\partial \sigma_R}{\partial \alpha} = \frac{\partial f(\alpha,\delta,\rho)/\partial \alpha}{n(1-\delta)} \begin{cases} = 0 & \text{if } \rho = 1, \\ < 0 & \text{if } \rho \in [0,1). \end{cases}$$

■

Proof of Proposition 2

i) Assume that the stability condition holds. The cartel's price is then given by

$$p_\pi^C = \arg\max_p (1-\alpha r_\pi)(p-c) q(p) = p^M(c).$$

Let π_π^C and F_π^C denote the corresponding firm-level profit and fine respectively.

ii)-iii) As a profit-based fine does not alter a firm's incentives compared to the no-antitrust case, the optimal defection is to slightly undercut the monopoly price and capture the entire market. This increases the defector's before-fine profit and fine n-fold compared to the cartel case: $\pi_\pi^D = n\pi_\pi^C$ and $F_\pi^D = nF_\pi^C$. Post-defection, regardless of whether the cartel has been convicted, the firms revert to the Nash equilibrium of the static Bertrand game, so $p_\pi^{PD} = c$ and $\pi_\pi^{PD} = 0$. Together with $\pi^N = 0$ and the preceding paragraph, this implies that $V_\pi^D = n\left(\pi_\pi^C - \alpha F_\pi^C\right)$.

Denote the critical discount factor by δ_π^*. As $\frac{V_\pi^C}{V_\pi^D} = \frac{V_R^C}{V_R^D}$ it follows immediately from the proof of Proposition 1 that

$$\frac{\partial \delta_\pi^*}{\partial \alpha} \begin{cases} = 0 & \text{if } \rho = 1, \\ > 0 & \text{if } \rho \in [0, 1). \end{cases}$$

■

Proof of Proposition 3

i) Assume that the stability condition holds. For certain parameters, overcharge-based fines allow for stable cartels but constrain the cartel price. Consider first the behavior of a stable cartel whose pricing is not constrained by the stability condition in (3). This 'unconstrained' cartel price is then given by

$$p_O^{UC} = \arg\max_p (p - c)\, q(p) - \alpha r_O \left(p - p^N\right) q(p^N) < p^M(c).$$

The inequality follows from the first-order condition underlying the maximization problem: $(p - c)\frac{\partial q(p)}{\partial p} + q(p) - \alpha r_O q(p^N) = 0$. The first two terms define the monopoly price. The overcharge-based fine introduces the third term, which incentivizes the cartel to set a price below $p^M(c)$ to reduce the expected fine and increase expected profit. The difference between $p^M(c)$ and p_O^{UC} increases with the detection probability, the penalty rate, and the Nash-equilibrium quantity of the static Bertrand game, none of which are influenced by the cartel's pricing decision.

To see that the unconstrained cartel price is strictly above marginal cost, rewrite the cartel's maximization problem as

$$\max_q (p(q) - c)(q - \alpha r_O q(p^N)).$$

The associated first-order condition is $p(q) + \frac{\partial p(q)}{\partial q}(q - \alpha r_O q(p^N)) = c$. Note that the left-hand side of the first-order condition is equal to $p(q)$ if $q = \alpha r_O q(p^N)$ and lies below inverse demand $p(q)$ for higher values of q, implying that $p_O^{UC} > c$.

Note that an individual firm's fine is scaled by their share of the competitive output $q^N = \frac{q(p^N)}{n}$ instead of $q(p^N)$. A defector is, therefore, faced with the first-order condition

$(p-c)\frac{\partial q(p)}{\partial p}+q(p)-\alpha r_O\frac{q(p^N)}{n}=0$. Comparison to the first-order condition in the first paragraph of this Proof shows that a defector would ideally increase the price compared to the cartel. However, this would result in no demand, so the best a defector can do is to slightly undercut the cartel's price and capture the entire market. This increases the defector's before-fine profit n-fold while leaving the fine unchanged. Compared to the cartel case: $\pi_O^D=n\pi_O^C$ and $F_O^D=F_O^C$. Post-defection, regardless of whether the cartel has been convicted, firms revert to the Nash-equilibrium of the static Bertrand game, so $p_O^{PD}=c$ and $\pi_O^{PD}=0$. Together with $\pi^N=0$ and the preceding paragraph, this implies that $V_O^D=n\pi_O^C-\alpha F_O^C$.

Let $\sigma_O(\alpha,\delta,\rho,p)\equiv\frac{V_O^C}{V_O^D}=\frac{\left(q(p)/n-\alpha r_O q^N\right)f(\alpha,\delta,\rho)}{(q(p)-\alpha r_O q^N)(1-\delta)}$. Comparing σ_O to σ_R and σ_π, note that $\frac{\partial\sigma_R}{\partial p}=\frac{\partial\sigma_\pi}{\partial p}=0$, while

$$\frac{\partial\sigma_O}{\partial p}=\frac{\partial q(p)}{\partial p}\frac{(1-1/n)\alpha r_O q^N f(\alpha,\delta,\rho)}{\left(q(p)-\alpha r_O q^N\right)^2(1-\delta)}<0,$$

as demand strictly decreases with price. In addition, since $\frac{\partial f(\alpha,\delta,\rho)}{\partial\delta}=\frac{-\alpha(1-\rho)}{(1-\delta(1-\alpha)(1-\rho))^2}$

$$\frac{\partial\sigma_O}{\partial\delta}=\underbrace{P}_{>0}\left(\underbrace{\frac{\partial f(\alpha,\delta,\rho)/\partial\delta}{(1-\delta)}+\frac{f(\alpha,\delta,\rho)}{(1-\delta)^2}}_{>0}\right)>0,$$

where $P=\frac{q(p)/n-\alpha r_O q^N}{q(p)-\alpha r_O q^N}$. The critical discount factor, δ_O^*, is the δ such that $\sigma_O\left(\alpha,\delta,\rho,p_O^{UC}\right)=1$. Define $\bar{\delta}_O$ as the delta such that $\sigma_O\left(\alpha,\delta,\rho,c\right)=1$. As $p_O^{UC}>c$, $\delta_O^*>\bar{\delta}_O$. Consider parameter values such that $\sigma_O\left(\alpha,\delta,\rho,p_O^{UC}\right)=1$, and then decrease δ slightly. The above implies that p_O^{UC} can no longer be part of a subgame-perfect Nash equilibrium, but that $\sigma_O\left(\alpha,\delta,\rho,p\right)=1$ can be restored by the cartel selecting a 'constrained' cartel price p_O^{CC} that is strictly below p_O^{UC}. The price of stable cartels in the overcharge regime is, therefore, always strictly below the monopoly price and given by

$$p_O^C=\begin{cases}p_O^{UC} & \text{if } \delta_O^*\leq\delta<1,\\ p_O^{CC} & \text{if } \bar{\delta}_O<\delta<\delta_O^*.\end{cases}$$

ii) By the implicit function theorem, we have

$$\frac{\partial\delta_O^*}{\partial\alpha}=-\frac{\partial\sigma_O^*/\partial\alpha}{\partial\sigma_O^*/\partial\delta}.$$

As before, let $P=\frac{q(p)/n-\alpha r_O q^N}{q(p)-\alpha r_O q^N}$. As $\frac{\partial\sigma_O}{\partial\delta}>0$ by the proof of Proposition 3i), and as $\frac{\partial f(\alpha,\delta,\rho)}{\partial\alpha}\leq 0$, it follows that

$$\frac{\partial\sigma_O}{\partial\alpha}=\underbrace{\frac{\partial P}{\partial\alpha}}_{<0}\underbrace{\frac{f(\alpha,\delta,\rho)}{1-\delta}}_{>0}+\underbrace{\frac{P}{1-\delta}}_{>0}\underbrace{\frac{\partial f(\alpha,\delta,\rho)}{\partial\alpha}}_{\leq 0}<0,$$

so that $\frac{\partial \delta_O^*}{\partial \alpha} > 0$, for all $\rho \in [0,1]$.
■

Proof of Proposition 4
i) Follows directly from Propositions 1 to 3.
ii) By the proofs of Propositions 1 and 2, $\frac{V_\pi^C}{V_\pi^D} = \frac{V_R^C}{V_R^D}$, so that $\delta_R^* = \delta_\pi^*$. By the proofs of Propositions 2 and 3, showing that $\sigma_O(\alpha, \delta, \rho, p) < \sigma_R(\alpha, \delta, \rho)$ (defined in the same proofs) is sufficient to show that $\delta_O^* > \delta_R^*$ for all jointly-optimal cartel prices. Because $\sigma_O(\alpha, \delta, \rho, p) = \frac{\left(q(p)/n - \alpha r_O q^N\right) f(\alpha,\delta,\rho)}{(q(p) - \alpha r_O q^N)(1-\delta)}$ and $\sigma_R(\alpha, \delta, \rho) = \frac{f(\alpha,\delta,\rho)}{n(1-\delta)}$, we have

$$\sigma_O(\alpha, \delta, \rho, p) < \sigma_R(\alpha, \delta, \rho) \iff \frac{q(p)/n - \alpha r_O q^N}{q(p) - \alpha r_O q^N} < \frac{1}{n}. \tag{A1}$$

Note that $\sigma_O(\alpha, \delta, \rho, p) = \sigma_R(\alpha, \delta, \rho)$ if $\alpha r_O q^N = 0$ (i.e., if there is no antitrust enforcement). It follows that $\sigma_O(\alpha, \delta, \rho, p) < \sigma_R(\alpha, \delta, \rho)$ as $\alpha r_O q^N > 0$ and the left-hand side of the final inequality of equation (A1) is strictly decreasing in $\alpha r_O q^N$.
■

B Additional theoretical results

B.1 Price-dependent detection probability

This appendix studies an extension of the model in Section 2 in which the probability that an active cartel is detected at the end of a period depends on the market price it sets that period. The purpose of this exercise is to show that even if higher cartel prices make detection more likely, for instance by attracting the attention of the antitrust authority, a revenue-based fine can still induce a cartel to set a price above the no-antitrust monopoly price. We find that the above-monopoly result survives if the detection probability does not increase too strongly with the market price. We first establish this result for general market conditions and then provide critical slopes of the detection probability function for the experimental parameters.

Generalizing the model in the main text, we assume that the probability of detection is given by a function $\alpha(p) : [c, \bar{p}] \to [0, 1]$, where $\alpha(\cdot)$ is weakly increasing in the cartel price p and continuously differentiable. As marginal costs are fixed, this can be interpreted as detection increasing in the price-cost margin that the cartel sets. A natural alternative specification is one where the detection probability increases in the change in price, rather than the price level. Harrington (2004) shows that a stable cartel gradually raises price toward the unconstrained optimum level in this setting because larger price movements make detection more likely. Viewed through that lens, our baseline results can be read as describing the cartel's steady state, so that its price will increase over time and eventually

exceed the monopoly price. If detection depends on the level of the price, however, then the optimal unconstrained cartel price in the revenue regime can, but need not, be above the monopoly price, as we next show.

Compared to the value of colluding and the value of defection defined in the baseline model, the only change is that the detection probability now depends on the price so that $V^C = \frac{(\pi^C - \alpha^C F^C) f(\alpha^C, \delta, \rho)}{1-\delta}$ and $V^D = \pi^D - \alpha^D F^D$ with $\alpha^C = \alpha(p^C)$ and $\alpha^D = \alpha(p^D)$. With exogenous detection probability, f is a constant from the viewpoint of the cartel and the critical discount factor does not depend on the cartel price as the optimal defection always increases expected period profit by a factor n. As now f depends on p, this is no longer true, and as a consequence higher prices might reduce the set of all δ for which the cartel can be part of a subgame-perfect Nash equilibrium. Hence, in what follows, we need to account for the possibility that the unconstrained joint-profit-maximizing price is not feasible.

Proposition 5. *If fines are based on revenue and detection is given by a weakly increasing, continuously differentiable function $\alpha(p)$, then*

i) *if*

$$r_R(-q'(p^M))\left(\alpha(p^M)c - p^M\alpha'(p^M)(p^M - c)\right) > \left((1-\alpha(p^M)r_R)p^M - c\right) q(p^M)\frac{\delta(1-\rho)\alpha'(p^M)}{1-\delta(1-\alpha(p^M))(1-\rho)}, \tag{B1}$$

there exists $\varepsilon > 0$ such that $V_R^C(p) > V_R^C(p^M)$ for all $p \in (p^M, p^M + \varepsilon)$;

ii) *if*

$$\frac{\left((1-\alpha(p^M)r_R)p^M - c\right) q(p^M)}{n}\frac{f(\alpha(p^M), \delta, \rho)}{1-\delta} > \sup_{p<p^M}\left\{\left((1-\alpha(p)r_R)p - c\right) q(p)\right\}, \tag{B2}$$

there exists $\eta > 0$ such that $V_R^C(p) \geq V_R^D(p)$ for all $p \in [p^M, p^M + \eta]$.

Proof of Proposition 5

i) The goal is to show that the value of collusion is increasing in p at the monopoly price under the condition given in equation B1, which, together with continuity, proves claim this claim. In the revenue regime,

$$V_R^C(p) = \frac{\left((1-\alpha(p)r_R)p - c\right) q(p)}{n}\frac{f(\alpha(p), \delta, \rho)}{1-\delta}.$$

Differentiating with respect to p and evaluating at p^M gives

$$\frac{\partial V_R^C(p^M)}{\partial p} = \frac{f(\alpha(p^M), \delta, \rho)}{n(1-\delta)}\Bigg(r_R(-q'(p^M))\left(\alpha(p^M)c - p^M\alpha'(p^M)(p^M - c)\right) - \left((1-\alpha(p^M)r_R)p^M - c\right) q(p^M)\frac{\delta(1-\rho)\alpha'(p^M)}{1-\delta(1-\alpha(p^M))(1-\rho)}\Bigg),$$

where we used $q(p^M) + (p^M - c)q'(p^M) = 0$ and

$$\frac{\partial f(\alpha(p), \delta, \rho)}{\partial p} = -f(\alpha(p), \delta, \rho)\frac{\delta(1-\rho)\alpha'(p)}{1-\delta(1-\alpha(p))(1-\rho)}.$$

Condition (B1) ensures that the term inside the large parentheses is strictly positive, meaning $\frac{\partial V_R^C(p^M)}{\partial p} > 0$. Because $\alpha(p)$ and $q(p)$ are continuously differentiable, the derivative $\frac{\partial V_R^C(p)}{\partial p}$ is continuous at p^M. Since $\frac{\partial V_R^C(p^M)}{\partial p} > 0$, there exists $\varepsilon > 0$ such that $\frac{\partial V_R^C(p)}{\partial p} > 0$ for all $p \in (p^M, p^M + \varepsilon)$. Hence, $V_R^C(p) > V_R^C(p^M)$ for all $p \in (p^M, p^M + \varepsilon)$.

ii) The expected present value of defecting when the cartel price is p^C is

$$V_R^D(p^C) = \sup_{p<p^C} \left\{\left((1-\alpha(p)r_R)p - c\right) q(p)\right\}.$$

Because $p \mapsto \left((1-\alpha(p)r_R)p - c\right) q(p)$ is continuous on $[c, \bar{p}]$, we have

$$V_R^D(p^C) = \max_{p\in[c,p^C]} \left\{\left((1-\alpha(p)r_R)p - c\right) q(p)\right\}.$$

Hence, by Berge's Maximum Theorem, the function $p^C \mapsto V_R^D(p^C)$ is continuous at p^M. Since

$$V_R^C(p) = \frac{\left((1-\alpha(p)r_R)p - c\right) q(p)}{n}\frac{f(\alpha(p), \delta, \rho)}{1-\delta},$$

and $\alpha(\cdot)$ and $q(\cdot)$ are continuous, the function $p \mapsto V_R^C(p)$ is continuous at p^M as well. It follows that $V_R^C(\cdot) - V_R^D(\cdot)$ is continuous at p^M. Condition (B2) is exactly the statement that $V_R^C(p^M) > V_R^D(p^M)$. Therefore, by continuity, there exists $\eta > 0$ such that $V_R^C(p) \geq V_R^D(p)$ for all $p \in [p^M, p^M + \eta]$.

■

Proposition 5 gives conditions under which the cartel strictly prefers prices above the monopoly price to the monopoly price in the revenue regime, and under which such prices can be part of a subgame-perfect Nash equilibrium. To make the first condition more intuitive, we solve for the slope of the detection probability. To avoid dividing by zero, we assume that the net-of-expected-fine margin, and hence expected profit, at the monopoly price is positive ($(1-\alpha(p^M)r_R)p^M > c$), so that

$$\alpha'(p^M) < \alpha'_{crit}(p^M), \tag{B3}$$

where

$$\alpha'_{crit}(p^M) \equiv \frac{r_R(-q'(p^M))\alpha(p^M)c}{r_R(-q'(p^M))p^M(p^M-c) + \left((1-\alpha(p^M)r_R)p^M - c\right) q(p^M)\frac{\delta(1-\rho)}{1-\delta(1-\alpha(p^M))(1-\rho)}},$$

which shows that detection probability cannot increase too rapidly in p for the cartel price to exceed the monopoly price in the revenue regime. The baseline model explained that

a revenue regime puts upward pressure on prices, resulting in a price above the monopoly price as this increases expected profit by decreasing the expected fine. Making the detection probability depend on price puts downward pressure on price, and if this force is sufficiently strong because the detection probability is very responsive to prices, this new effect dominates and the cartel price no longer exceeds the monopoly price in the revenue regime.

To develop some sense of the quantitative implications of Proposition 5, we calibrate it using the demand system and parameter values from the experiment $q(p) = 100 - p, c = 47, r_R = 1, \delta = 0.9$. Recall that these values imply that $p^M = 73.5$. We normalize the level of the detection function at the monopoly price to the value used in the baseline model and experiment: $\alpha(p^M) = 0.2$. Under this calibration, condition (B3) becomes

$$\alpha'(p^M) < \frac{9.4}{1947.75 + 312.7\frac{0.9(1-\rho)}{1-0.72(1-\rho)}}. \tag{B4}$$

To interpret this threshold, we use a local linear approximation to the detection function around the monopoly price: $\alpha(p) \approx \alpha(p^M) + \alpha'_{\text{crit}}(p^M)(p - p^M)$. We evaluate it at the three cartel prices from Table 2: $p_O^C \approx 63.69, p_\pi^C = p^M = 73.5, p_R^C \approx 79.38$. The resulting values should therefore be interpreted as implied detection probabilities at those price points under the critical local slope, not as equilibrium prices of the present extension. Under the same local linear approximation, condition (B2) is satisfied for all parameter combinations reported in Table 6.

Table 6: Critical slope and implied detection probabilities when $\alpha(p^M) = 0.2$, by reformation probability ρ

ρ	$\alpha'_{\text{crit}}(p^M)$	$\alpha(p_O^C)$	$\alpha(p_\pi^C)$	$\alpha(p_R^C)$
1	0.005	0.153	0.200	0.228
0.5	0.004	0.157	0.200	0.225
0	0.003	0.169	0.200	0.219

Notes: The table uses the calibration $q(p) = 100 - p$, $c = 47$, $r_R = 1$, $\delta = 0.9$, and $\alpha(p^M) = 0.2$. The critical slope $\alpha'_{\text{crit}}(p^M)$ is computed from (B4). The next three columns use the local linear approximation in $\alpha(p) \approx \alpha(p^M) + \alpha'_{\text{crit}}(p^M)(p - p^M)$ evaluated at $p_O^C \approx 63.69$, $p_\pi^C = 73.5$, and $p_R^C \approx 79.38$. Numbers rounded to three decimal points.

Table 6 shows that, under our calibration, a revenue-based cartel price above the monopoly price remains possible provided that raising price from 73.5 to 79.38 increases the detection probability by no more than about 2.8 percentage points. Of course, it is difficult to form an empirically grounded opinion on whether such an increase is the norm or not, but at least the conditions under which the cartel price exceeds the monopoly price in the revenue regime do not appear particularly unreasonable even when detection probability depends positively on the price.

B.2 Damages claims

This appendix extends the model in Section 2 by allowing cartel members and consumers to internalize the possibility of a successful damages claim. We show that the key results from the baseline model continue to hold in that the relative rankings of the fining regimes in terms of cartel prices and critical discount factors are unaffected. Additionally, we find that anticipation of damages may push up the cartel price above the monopoly price in all regimes instead of only in the revenue regime. Our analysis proceeds in two steps. First, we examine the situation where only the firms anticipate damages claims. Then, we add the assumption that also consumers take future damages claims into account.

We extend our baseline model as follows. Suppose that, upon detection in period t, cartel member i also pays damages

$$H_{it} = r_D(p_{it} - p^N)q_{it},$$

where $r_D > 0$ measures the share of damages paid out. As $p^N = c$ in our Bertrand setting, this becomes $H_{it} = r_D(p_{it} - c)q_{it}$. Let $\tilde{p}_R^C$ and $\tilde{p}_\pi^C$ denote the counterparts of p_R^C and p_π^C in the presence of damages. In the overcharge regime, let $\tilde{p}_O^{UC}$ and $\tilde{p}_O^{CC}$ denote the counterparts of the unconstrained and constrained cartel prices p_O^{UC} and p_O^{CC}, respectively. Let $\tilde{\delta}_R^*$, $\tilde{\delta}_\pi^*$, and $\tilde{\delta}_O^*$ denote the critical discount factors corresponding to the revenue, profit, and overcharge regimes respectively. Recall that $q(p)$ denotes total market quantity and that $q^N \equiv q(p^N)/n$ is an individual firm's competitive quantity. In what follows, we assume $1 - \alpha(r_R + r_D) > 0$, a regularity condition ensuring that, in the revenue regime with damages, the cartel's objective can still be written as a positively scaled monopoly-profit problem with a well-defined effective marginal cost.

Proposition 6. *Suppose that cartel members internalize the possibility of a damages claim as given by H_{it},*

- ***i)*** *if fines are based on revenue, the price set by a stable cartel exceeds its counterpart in the baseline model: $\tilde{p}_R^C > p_R^C$. The critical discount factor is unchanged: $\tilde{\delta}_R^* = \delta_R^*$;*
- ***ii)*** *if fines are based on profit, the price set by a stable cartel and the critical discount factor are unchanged compared to the baseline model: $\tilde{p}_\pi^C = p_\pi^C = p^M(c)$ and $\tilde{\delta}_\pi^* = \delta_\pi^*$;*
- ***iii)*** *if fines are based on the overcharge, the unconstrained cartel price lies below its counterpart in the baseline model: $\tilde{p}_O^{UC} < p_O^{UC}$. When the stability condition binds, the constrained cartel price is given by*

$$\tilde{p}_O^{CC} = p\left(\frac{n\alpha r_O q^N \left(f(\alpha,\delta,\rho) - (1-\delta)\right)}{(1-\alpha r_D)\left(f(\alpha,\delta,\rho) - n(1-\delta)\right)}\right),$$

and satisfies $\tilde{p}_O^{CC} < p_O^{CC}$. If the unconstrained cartel price is stable in both models, then the price set by a stable cartel is strictly lower with damages. More generally, if

the stability condition binds, the price set by a stable cartel is weakly lower than in the baseline model. For any fixed cartel price $p > c$, the minimum discount factor required to sustain p is higher in the overcharge regime than in the revenue and profit regimes in the damages model, and is also higher than in the baseline overcharge model.

Proof of Proposition 6

i) In the revenue regime, the cartel solves

$$\tilde{p}_R^C = \arg\max_p \left((1-\alpha r_D)(p-c) - \alpha r_R p\right) q(p) = p^M\left(\frac{1-\alpha r_D}{1-\alpha(r_R+r_D)}c\right).$$

Compared to Proposition 1, the cartel acts like a monopolist (in the absence of antitrust) facing marginal cost $\frac{1-\alpha r_D}{1-\alpha(r_R+r_D)}c$ rather than $\frac{c}{1-\alpha r_R}$, the counterpart without damages. As

$$\frac{1-\alpha r_D}{1-\alpha(r_R+r_D)}c - \frac{c}{1-\alpha r_R} = \frac{\alpha^2 r_R r_D c}{(1-\alpha(r_R+r_D))(1-\alpha r_R)} > 0,$$

it follows that $\tilde{p}_R^C > p_R^C$ because $p^M(\cdot)$ is strictly increasing.

To establish the implied critical discount factor, let $\tilde{\pi}_R^C$, $\tilde{F}_R^C$, and $\tilde{H}_R^C$ denote a cartel member's profit, fine, and damages at $\tilde{p}_R^C$. A deviator slightly undercuts $\tilde{p}_R^C$ and captures the entire market, so that $\tilde{\pi}_R^D = n\tilde{\pi}_R^C$, $\tilde{F}_R^D = n\tilde{F}_R^C$, and $\tilde{H}_R^D = n\tilde{H}_R^C$. Post-defection, firms act as if they face marginal cost $\frac{1-\alpha r_D}{1-\alpha(r_R+r_D)}c$, so the unique symmetric pure-strategy one-shot Nash equilibrium market price equals $\frac{1-\alpha r_D}{1-\alpha(r_R+r_D)}c$, implying $\tilde{\pi}_R^{PD} - \alpha\tilde{F}_R^{PD} - \alpha\tilde{H}_R^{PD} = 0$. Hence,

$$\tilde{V}_R^C = \frac{(\tilde{\pi}_R^C - \alpha\tilde{F}_R^C - \alpha\tilde{H}_R^C) f(\alpha,\delta,\rho)}{1-\delta} \quad \text{and} \quad \tilde{V}_R^D = n(\tilde{\pi}_R^C - \alpha\tilde{F}_R^C - \alpha\tilde{H}_R^C).$$

So, anticipation of damages does not affect the cartel price. The ratio $\tilde{V}_R^C/\tilde{V}_R^D$ is, therefore, identical to V_R^C/V_R^D in the proof of Proposition 1, which implies $\tilde{\delta}_R^* = \delta_R^*$.

ii) In the profit regime, the cartel solves

$$\tilde{p}_\pi^C = \arg\max_p (1-\alpha(r_\pi + r_D))(p-c)q(p) = p^M(c) = p_\pi^C.$$

Let $\tilde{\pi}_\pi^C$, $\tilde{F}_\pi^C$, and $\tilde{H}_\pi^C$ denote a cartel member's profit, fine, and damages at $\tilde{p}_\pi^C$. As in the proof of Proposition 2, a deviator slightly undercuts $\tilde{p}_\pi^C$ and captures the entire market, so that $\tilde{\pi}_\pi^D = n\tilde{\pi}_\pi^C$, $\tilde{F}_\pi^D = n\tilde{F}_\pi^C$, and $\tilde{H}_\pi^D = n\tilde{H}_\pi^C$. Post-defection, firms revert to setting $p^N = c$, implying $\tilde{\pi}_\pi^{PD} - \alpha\tilde{F}_\pi^{PD} - \alpha\tilde{H}_\pi^{PD} = 0$. Hence,

$$\tilde{V}_\pi^C = \frac{(\tilde{\pi}_\pi^C - \alpha\tilde{F}_\pi^C - \alpha\tilde{H}_\pi^C) f(\alpha,\delta,\rho)}{1-\delta} \quad \text{and} \quad \tilde{V}_\pi^D = n(\tilde{\pi}_\pi^C - \alpha\tilde{F}_\pi^C - \alpha\tilde{H}_\pi^C).$$

The ratio $\tilde{V}_\pi^C/\tilde{V}_\pi^D$ is, therefore, identical to V_π^C/V_π^D in the proof of Proposition 2, implying $\tilde{\delta}_\pi^* = \delta_\pi^*$.

iii) In the overcharge regime, the unconstrained cartel solves $\tilde{p}_O^{UC} = \arg\max_p (1-\alpha r_D)(p-c)q(p) - \alpha r_O (p-c) q(p^N)$. The associated first-order condition is

$$(1-\alpha r_D)\left((p-c)\frac{\partial q(p)}{\partial p} + q(p)\right) - \alpha r_O q(p^N) = 0, \text{ or } (p-c)\frac{\partial q(p)}{\partial p} + q(p) = \frac{\alpha r_O q(p^N)}{1-\alpha r_D}.$$

Compared to Proposition 3, the right-hand side is strictly larger as $1-\alpha r_D < 1$. As the left-hand side is strictly decreasing in p, it follows that $\tilde{p}_O^{UC} < p_O^{UC}$.

To study stability, let $\pi(p) = \frac{(p-c)q(p)}{n}$ and $F_O(p) = r_O(p-c)q^N$. Then the analog of $\sigma_O(\alpha, \delta, \rho, p)$ from the proof of Proposition 3 is

$$\tilde{\sigma}_O(\alpha, \delta, \rho, p) = \frac{\left((1-\alpha r_D)\pi(p) - \alpha F_O(p)\right) f(\alpha, \delta, \rho)}{\left(n(1-\alpha r_D)\pi(p) - \alpha F_O(p)\right)(1-\delta)}.$$

Comparing this to $\sigma_R(\alpha, \delta, \rho) = \frac{f(\alpha,\delta,\rho)}{n(1-\delta)}$, we have

$$\tilde{\sigma}_O(\alpha, \delta, \rho, p) < \sigma_R(\alpha, \delta, \rho) \iff \frac{(1-\alpha r_D)\frac{q(p)}{n} - \alpha r_O q^N}{(1-\alpha r_D)q(p) - \alpha r_O q^N} < \frac{1}{n}.$$

This inequality holds if and only if $n\alpha r_O q^N > \alpha r_O q^N$, which is true as $n \geq 2$ and $\alpha r_O q^N > 0$. Therefore, for any given cartel price $p > c$, the critical discount factor is higher in the overcharge regime than in the revenue and profit regimes.

When the stability condition binds, the constrained cartel price is obtained by solving $\tilde{\sigma}_O(\alpha, \delta, \rho, p) \geq 1$ for $q(p)$, which yields

$$q(p) \geq \frac{n\alpha r_O q^N \left(f(\alpha, \delta, \rho) - (1-\delta)\right)}{(1-\alpha r_D)\left(f(\alpha, \delta, \rho) - n(1-\delta)\right)}, \text{ hence } \tilde{p}_O^{CC} = p\left(\frac{n\alpha r_O q^N \left(f(\alpha, \delta, \rho) - (1-\delta)\right)}{(1-\alpha r_D)\left(f(\alpha, \delta, \rho) - n(1-\delta)\right)}\right).$$

Compared to the constrained cartel price in the baseline model, the only new term is $1-\alpha r_D$. As inverse demand is strictly decreasing in output, it follows that $\tilde{p}_O^{CC} < p_O^{CC}$.

Comparing this to $\sigma_O(\alpha, \delta, \rho, p)$ from the proof of Proposition 3, we obtain

$$\sigma_O(\alpha, \delta, \rho, p) - \tilde{\sigma}_O(\alpha, \delta, \rho, p) = \frac{\alpha^2 r_D F_O(p)(n-1)\pi(p) f(\alpha, \delta, \rho)}{\left(n\pi(p) - \alpha F_O(p)\right)\left(n(1-\alpha r_D)\pi(p) - \alpha F_O(p)\right)(1-\delta)} > 0$$

for all $p > c$. Therefore, damages tighten the stability condition for any given cartel price. In addition,

$$\frac{\partial \tilde{\sigma}_O}{\partial p} = \frac{(n-1)\alpha r_O q^N (1-\alpha r_D)(p-c)^2 \frac{\partial q(p)}{\partial p} f(\alpha, \delta, \rho)}{n\left(n(1-\alpha r_D)\pi(p) - \alpha F_O(p)\right)^2 (1-\delta)} < 0,$$

so lower prices relax the stability condition. Hence, if the unconstrained cartel price is stable in both models, the price set by a stable cartel is strictly lower with damages. If the stability condition binds, the cartel can only restore stability by lowering the price, so the price set by a stable cartel is weakly lower than in the baseline model.

Finally, because $\tilde{\sigma}_O(\alpha, \delta, \rho, p) < \sigma_O(\alpha, \delta, \rho, p)$ for any given $p > c$, the minimum discount factor required to support that price is higher with damages than in the baseline model. ■

Next, suppose that consumers understand that, upon detection, cartel member i pays damages $H_{it} = r_D(p_{it} - p^N)q_{it}$ and, therefore, consumers expect to recoup a share αr_D of the overcharge. A consumer facing posted price p_{it} then behaves as if the relevant price were $x_{it} \equiv p_{it} - \alpha r_D(p_{it} - p^N)$. As $p^N = c$, we have $x_{it} = (1 - \alpha r_D)p_{it} + \alpha r_D c$. Demand is, therefore, given by $q(x_{it})$ rather than $q(p_{it})$. Consumer anticipation changes the relationship between posted prices and demand: for any posted price above marginal cost, the expected net price lies below the posted price, which increases demand.

Let $\hat{p}_R^C$ and $\hat{p}_\pi^C$ denote the counterparts of $\tilde{p}_R^C$ and $\tilde{p}_\pi^C$ when consumers anticipate damages. In the overcharge regime, let $\hat{p}_O^{UC}$ and $\hat{p}_O^{CC}$ denote the counterparts of $\tilde{p}_O^{UC}$ and $\tilde{p}_O^{CC}$, respectively, and let $\hat{p}_O^C$ denote the price set by a stable cartel, i.e., $\hat{p}_O^C = \hat{p}_O^{UC}$ if the unconstrained cartel price is stable and $\hat{p}_O^C = \hat{p}_O^{CC}$ otherwise. Finally, let $\hat{\delta}_R^*$, $\hat{\delta}_\pi^*$, and $\hat{\delta}_O^*$ denote the corresponding critical discount factors.

Proposition 7. *Suppose that cartel members internalize the possibility of a damages claim and that consumers anticipate these damages as described above,*

- ***i)*** *the price set by a stable cartel is highest if fines are based on revenue and lowest if fines are based on the overcharge: $\hat{p}_R^C > \hat{p}_\pi^C > \hat{p}_O^C$;*
- ***ii)*** *the critical discount factor is highest if fines are based on the overcharge: $\hat{\delta}_O^* > \hat{\delta}_R^* = \hat{\delta}_\pi^*$.*
- ***iii)*** *the posted prices in the revenue and profit regimes strictly exceed the monopoly price from the main text: $\hat{p}_R^C > p^M(c)$ and $\hat{p}_\pi^C > p^M(c)$. In the overcharge regime, $\hat{p}_O^C > p^M(c)$ if and only if the corresponding net cartel price exceeds $p^M(c) - \alpha r_D(p^M(c) - c)$. In particular, if the unconstrained cartel price is stable, then $\hat{p}_O^C > p^M(c)$ for sufficiently small $r_O > 0$.*

Proof of Proposition 7

i) As $x = (1-\alpha r_D)p + \alpha r_D c$, we have $p = c + \frac{x-c}{1-\alpha r_D}$, so posted prices are strictly increasing in net prices and the order of posted prices and net prices is identical. In the revenue regime, the cartel solves $\max_p \left((1 - \alpha r_D)(p - c) - \alpha r_R p\right) q(x)$, which is equivalent to $\max_x \frac{1-\alpha(r_R+r_D)}{1-\alpha r_D}(x - \hat{c}_R)q(x)$, where $\hat{c}_R \equiv c\frac{1-\alpha r_D - \alpha^2 r_R r_D}{1-\alpha(r_R+r_D)} > c$. Hence, the chosen net price exceeds $p^M(c)$. In the profit regime, the cartel solves $\max_p(1 - \alpha(r_\pi + r_D))(p - c)q(x)$, which is equivalent to $\max_x \frac{1-\alpha(r_\pi+r_D)}{1-\alpha r_D}(x - c)q(x)$, so the chosen net price equals $p^M(c)$. In the overcharge regime, the cartel solves $\max_p(1 - \alpha r_D)(p - c)q(x) - \alpha r_O(p - c)q(p^N)$, which is equivalent to $\max_x(x - c)\left(q(x) - \frac{\alpha r_O q(p^N)}{1-\alpha r_D}\right)$. Thus, the unconstrained net cartel price lies below $p^M(c)$, and the constrained cartel price, when relevant, is lower still. As posted prices are strictly increasing in net prices, it follows that $\hat{p}_R^C > \hat{p}_\pi^C > \hat{p}_O^C$.

ii) In the revenue and profit regimes, a deviator slightly undercuts the cartel and captures the entire market, so before-fine profit, the fine, and damages all increase n-fold, while the post-defection equilibrium yields zero expected profit. Hence, as in Propositions 1, 2, and 6, the stability ratio is $\frac{f(\alpha,\delta,\rho)}{n(1-\delta)}$ in both regimes, implying $\hat{\delta}_R^* = \hat{\delta}_\pi^*$. In the overcharge regime, the fine does not scale up with the deviator's quantity sold, so deviation is more attractive than in the other two regimes. Equivalently, the stability ratio is lower in the overcharge regime for every cartel price above c. Therefore, $\hat{\delta}_O^* > \hat{\delta}_R^* = \hat{\delta}_\pi^*$.

iii) In the revenue regime, the chosen net price exceeds $p^M(c)$ by the proof of part *i)*. As $p = c + \frac{x-c}{1-\alpha r_D}$ and $1 - \alpha r_D < 1$, any posted price that implements a net price above $p^M(c)$ must itself exceed $p^M(c)$. Hence, $\hat{p}_R^C > p^M(c)$. In the profit regime, the chosen net price equals $p^M(c)$, so $\hat{p}_\pi^C = c + \frac{p^M(c)-c}{1-\alpha r_D} > p^M(c)$.

Let x_O^C denote the corresponding net cartel price in the overcharge regime. Then $\hat{p}_O^C = c + \frac{x_O^C - c}{1-\alpha r_D}$, so

$$\hat{p}_O^C > p^M(c) \iff x_O^C > p^M(c) - \alpha r_D(p^M(c) - c).$$

This gives the necessary and sufficient condition for the posted overcharge price to exceed the monopoly price from the main text. As $r_O \downarrow 0$, any unconstrained maximizer $x_O^{UC}(r_O)$ converges to $p^M(c)$. Since $p^M(c) - \alpha r_D(p^M(c) - c) < p^M(c)$, there exists $\bar{r}_O > 0$ such that $x_O^{UC}(r_O) > p^M(c) - \alpha r_D(p^M(c) - c)$ for all $r_O \in (0, \bar{r}_O)$. Therefore, $\hat{p}_O^{UC} > p^M(c)$ for all such r_O.

■

C Instructions

Subjects could read through the computerized instructions at their own pace. All test questions needed to be answered correctly for the subject to progress to the experiment. Most of the instructions were identical across treatments. For the parts that differ, we include all three versions for comparison. We also provide a screenshot of the on-screen calculator from the `REVENUE` treatment and the risk preference test – discussed in Appendix D.

Introduction

We ask that you do not talk to other people during the experiment. Please refrain from verbally reacting to events that occur during the experiment. The use of mobile phones is not allowed. If you have any questions, or need assistance of any kind, please notify the experimenter by raising your hand.

Please comply with these rules, otherwise you will be asked to leave and you will not be paid.

Your earnings will depend on your decisions, and the decisions of other participants: your rivals. You will be paid privately and in cash at the end of the experiment.

Description of the experiment

In this experiment you will play a game four times. Each game consists of several periods. At the end of each period, there is a 90% chance that another period will be played and a 10% chance that the game ends.

In all periods of a game you will be matched to the same two participants: your rivals. In different games you will have different rivals. You always face the same rivals in different periods of the same game. You never face the same rivals in different games.

In each period of a game you and your rivals pick prices. Before picking prices, you can vote to form a cartel. If you and your two rivals vote in favour of a cartel, a cartel is formed, and you can chat about prices before setting your price. If no cartel is formed, next period you can vote again. Cartels are illegal and there is a 20% chance each period that all active cartels are detected. If your cartel is detected, you will pay a fine. The next period you can vote to form a new cartel. If your cartel is not detected, it is automatically active the next period. The market is described in detail on the next page. All numbers are rounded to two decimal points.

The market

The price you set must be between 0.01 and 99.99 (all inputs are rounded to two decimals). The quantity you sell from setting price p is:

$q = 0$	if you do not set the lowest price
$q = \frac{100-p}{n}$	if n firms set the lowest price

Your before-fine profit from setting price p is:

Before-fine profit $= 0$	if you do not set the lowest price.
Before-fine profit $= (p - 47)\frac{100-p}{n}$	if n firms set the lowest price

Note that setting a price below 47 will result in a loss.

When choosing your price, an on-screen calculator is available, as well as information on the history of the game.

Example 1: Firm 1 and 2 set price 50 and firm 3 sets price 61. Firm 3 did not set the lowest price and makes 0 profit this period. Firms 1 and 2 both set the lowest price so both get a before-fine profit equal to $(50 - 47)\frac{100-50}{2} = 75$.

Example 2: All three firms set price 70. All three firms set the lowest price and so get a before-fine profit equal to $(70 - 47)\frac{100-70}{3} = 230$.

The next page will give you more information about forming cartels.

Cartels

Before choosing prices, you and your rivals vote to form a cartel. Remember that a cartel is only formed if all three firms vote in favor. A cartel gives you access to a chat. In the first period of a cartel, you can chat for 1 minute before setting prices. In other periods you can chat for 30 seconds.

After chatting with your rivals, you will still need to set a price independently. Chatting about anything that can identify you in or outside of the lab will result in you not being paid for this experiment.

Forming a cartel is illegal and there is a 20% chance in each period that all active cartels are detected. If your cartel is not detected, it is automatically active next period. If it is detected, you pay a fine and the cartel ends. You then vote again on forming a new cartel. Your fine depends on the price, p, you set that period, as follows:

[OVERCHARGE]:

$Fine = 0$ if your price is 47.

$Fine = 32.69 * (p - 47)$ if your price is above 47.

Note: The fine increases by 32.69 points for each point the price exceeds 47, and you can be fined despite making no profit.

[PROFIT]:

$Fine = 2.33 * \textit{Before-fine profit}$

Note that you cannot be fined if you do not make a profit, and that a higher price will not always result in a higher fine.

[REVENUE]:

$Fine = 0$ if you did not set the lowest price.

$Fine = p * \left(\frac{100-p}{n}\right)$ if n firms set the lowest price.

Note that you cannot be fined if you do not set the lowest price, and that a higher price will not always result in a higher fine. The fine depends on your revenue.

Example 3: All three firms vote to form a cartel. A cartel is formed, and the firms discuss prices. All three firms set a price of 91.00 and get:

before-fine profit $= (91 - 47)\frac{(100-91)}{3} = 132$ points.

The cartel is not detected this period, so total profit is 132. The cartel and chat are automatically active next period.

Example 4: All three firms vote to form a cartel. A cartel is formed, and the firms discuss prices. Firms 1 and 2 set a price of 55. Firm 3 sets a price of 50.99. Firm 1 and 2 have before-fine profit equal to 0 as they did not set the lowest price. Firm 3 has before-fine profit equal to:

before-fine profit $= (50.99 - 47)\frac{100-50.99}{1} = 195.55$ points.

[OVERCHARGE]:
The cartel is detected this period. Firms 1 and 2 both pay the following fine:

$Fine = 32.69 * (55 - 47) = 261.52$ points.

Their profit for this period is -261.52.

Firm 3 pays the following fine:

$Fine = 32.69 * (50.99 - 47) = 130.43$ points.

Firm 3's profit for this period is 195.55 - 130.43 = 65.12 points. The next period starts with a new vote to form a cartel.

[PROFIT]:
The cartel is detected this period. Firms 1 and 2 pay no fine as they have no before-fine profit. Their profit for this period is 0.

Firm 3 pays the following fine:

$Fine = 2.33 * 195.55 = 455.63$ points.

Firm 3's profit for this period is 195.55 - 455.63 = -260.08 points. The next period starts with a new vote to form a cartel.

[REVENUE]:
The cartel is detected this period. Firms 1 and 2 pay no fine as they did not set the lowest price and therefore earned no revenue. Their profit for this period is 0.

Firm 3 pays the following fine:

$Fine = 50.99 * \left(\frac{100-50.99}{1}\right) = 2499.02$.

Firm 3's profit for this period is 195.55 - 2499.02 = -2303.47 points. The next period starts with a new vote to form a cartel.

Payment

During the experiment you will earn points. 3 points equal 1 eurocent. You will be paid based on all points earned in all four games, plus the 7 euro show-up fee.

In the unlikely event that you will make a loss in the experiment, you will still receive the 7 euro show-up fee. You will be paid privately and in cash at the end of the experiment.

You will now have to answer some questions to show that you understand the instructions. The first game begins when everyone has answered all questions correctly.

Question 1

How many games will you play, and against how many other people will you play?

-1 game, against 2 people

-1 game, against 8 people

-4 games, against the same 2 people each game, in total 2 people

-4 games, against 2 different people each game, in total 8 people

---

Question 2

Do all 4 games have the same number of periods?

-Yes

-No

-We can't be sure, after each period there is another period with 90% chance

---

Question 3

Firm 1 and 3 set price 80, firm 2 sets price 90. What is the before-fine profit of firm 3?

Question 4

Firm 1 and firm 2 vote in favor of a cartel, firm 3 votes against a cartel. Is there a cartel? Can firm 1 and 2 be fined?

-Yes and yes: Firm 1 and firm 2 form a cartel together and can therefore be fined

-No and no: No cartel is established and firms can only be fined when they are in a cartel

-No and yes: No cartel is established but since they voted for a cartel they can be fined

---

Question 5

Each period, there is a 20% chance that active cartels are detected and firms are fined. Does this mean that cartels will be discovered once every 5 periods?

-Yes, a 20% chance means once every 5 periods

-No, there is a 20% chance each period, but there could be many periods without detection

---

Question 6

Firms 1, 2 and 3 agreed in the chat to set a certain price, but all three firms set a lower price. Can the cartel still be detected and fined?

-No, the cartel members did not stick to the agreement

-Yes, once a cartel has been formed it can be detected, regardless of the firms' actions

---

Question 7

You are in a cartel. Which of these prices will lead to the highest fine?

-30

-50

-60

-90

Question 8

Firm 3 is part of a cartel and sets a price of 50. Firms 1 and 2 set a price of 40. Firm 3's before-fine profit is, therefore, 0. The cartel is detected. Does firm 3 need to pay a fine?

-No, because the lowest price is 40

-No, because Firm 3 sells nothing

-Yes, because Firm 3's fine depends on Firm 3's price

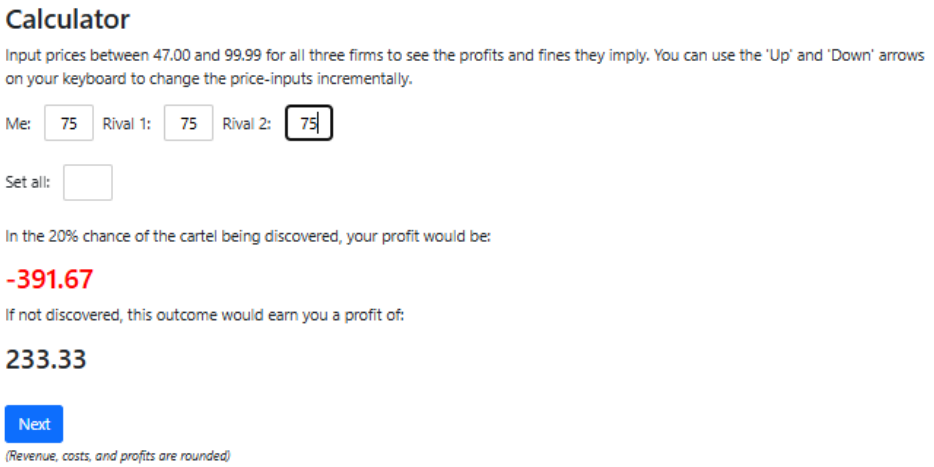


Figure C1: Screenshot of the on-screen calculator available to participants on the chat- and price setting pages (from REVENUE).

Instructions risk preference test

Below you see a table with four columns and multiple rows. For each row, you must make a choice between participating in a risky lottery, where there is a 20% chance of a low outcome and an 80% chance of a high outcome, or not participating, in which case you earn 0 points.

During the experiment you will earn points. 3 points equal 1 euro cent. You will be paid based on the outcome of this lottery choice, your performance in the rest of the experiment, plus a 7 euro show-up fee. You will be paid privately and in cash at the end of the experiment.

You must make a choice for every row, but one row has been randomly selected for payment.

When you go to the next page, all your choices are confirmed. The selected row is revealed at the end of the experiment.
If you chose 'Play Lottery' for the selected row, the lottery is played and you either receive the high or the low outcome.
If you chose 'No Lottery' for the selected row, the lottery is not played, and your payoff will not be affected.

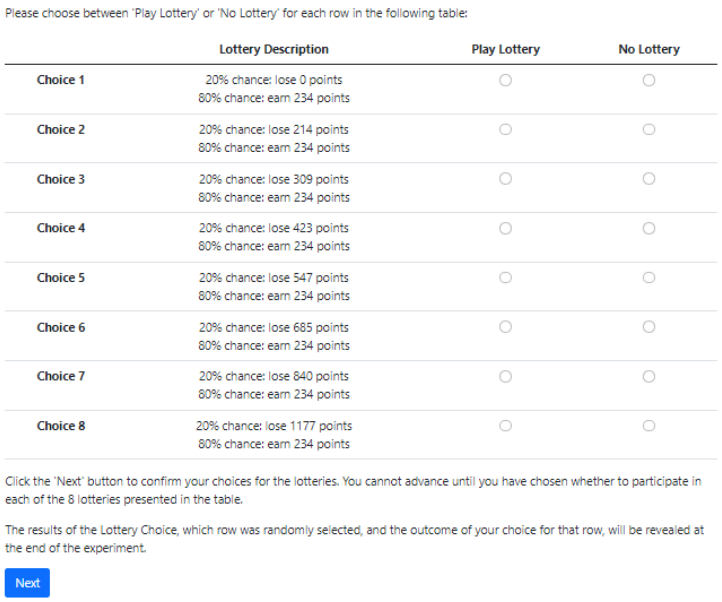
D Risk preference test

Joining a cartel and coordinating prices is risky, as collusion is detectable and punishable in our experiment. We, therefore, measured subjects' risk preferences. Before reading the instructions for and taking part in the repeated Bertrand games, each subject participated in a risk elicitation task based on Holt and Laury (2002), with outcomes chosen to mirror

the payoffs in the game that participants would subsequently play. The outcome of this test was communicated to the subjects at the very end of the session, after the conclusion of the Bertrand games.

Each subject needed to indicate for eight lotteries whether she wanted to participate or not. Figure D1 displays the lotteries as seen by the participants, and Appendix C includes the instructions. For each lottery, the chance of 'winning' was fixed at 80% (equal to the chance of cartels *not* being detected) and the rewards for winning were 234 points – the single-period before-fine profit of a subject in a cartel that coordinates on the monopoly price. However, the cost of 'losing' increased with each lottery. Subjects were paid based on one lottery, drawn randomly before the first session. If a subject had opted to play the randomly chosen lottery, a random draw determined whether any points were added or subtracted to the total earned in the four supergames.

Figure D1: Risk preference test



We construct as a measure of risk preferences the first row that a subject opts to not play the lottery. This measure ranges from 1 (subject does not play the lottery in row 1) to 9 (subject plays all eight lotteries), with higher values indicating a higher appetite for risk. Table D1 describes this measure by treatment. Our measure of risk preferences is distinctly balanced across treatments. Including this measure as a control variable in regressions that compare outcomes across treatments barely affects point estimates, and leads to at best a modest increase in efficiency. In the main text we, therefore, refrain from such analyses and mainly use information on risk preferences to argue that selection into cartels is similar in the three treatments.

Table D1: Risk preferences, by treatment

	REVENUE	PROFIT	OVERCHARGE
25th percentile	3	3	3
50th percentile	4	4	5
75th percentile	6	6	6
Mean	4.53	4.44	4.63
Standard deviation	1.84	1.99	1.84
Observations	90	99	90

Notes: Descriptive statistics on our measure of risk preferences, number equals last lottery that a subject opted to play.

E Classification of cartel agreements

This appendix provides a description of how we use the communication data to determine whether cartels coordinate on a particular price. We employed a third party with knowledge of competition policy to classify all chat messages to determine whether subjects had agreed on a specific price. We utilize two definitions of cartel agreements in our analysis. Explicit price agreements are classified purely based on the content of the chat.

Explicit price agreements An explicit price agreement to set price p in a given period is said to exist if at least one subject proposes price p, and all other subjects explicitly agree or reaffirm before any subject leaves the chat.

Explicit price agreements are present in 862 of all 2,122 market-periods with a cartel (40.62 percent). When an explicit agreement is in place, the average market price is 70.48 and subjects successfully coordinate on a price above the one-shot Nash equilibrium price of 47 in 716 market-periods (83.06 percent of all explicit price agreements), suggesting our measure captures cartel agreements. However, cartels without explicit price agreements still coordinate on prices above 47 in 55.63 percent of market-periods (701 of 1,260), compared to only 10.78 percent in uncartelized market-periods (82 of 761). Unsurprisingly, the average price in cartelized markets without explicit agreements (65.97) substantially exceeds that of uncartelized markets (50.29). Moreover, Figure G1 in Appendix G shows that the fraction of explicit price agreements tends downward within supergames, while the incidence of successful price coordination does not. These findings suggest our measure of explicit price agreements substantially underestimates the frequency of cartel agreements.

We construct a broader measure of price agreements based on the content of the chat

and on the past behavior of the cartel. This is necessary because stable cartels typically reduce communication significantly after successfully coordinating prices, up to the extreme cases where stable cartels at some point require no communication whatsoever but continue coordinating on the previous period's price.

Price agreements A price agreement to set price p in a given period is said to exist if an explicit price agreement is in place, or if an implicit agreement to set price p is in place. An implicit agreement to set price p in a given period is said to exist if i) at least one subject proposes price p, and all other subjects explicitly agree or reaffirm but at least one subject has left the chat before all non-proposers agree or reaffirm, ii) at least one subject suggests to do the same as the previous period without explicitly suggesting a price, and all other subjects explicitly agree or reaffirm, or if iii) at least one subject proposes price p or suggests to do the same as the previous period without explicitly suggesting a price, none of the non-proposers explicitly disagree but at least one non-proposer does not agree or reaffirm, iv) no price is proposed and no suggestions to follow past behavior are made, but coordination on an agreed-upon price p was achieved in the previous period and none of the subjects voice disagreement with past behavior.

Categories i) to iii) in the definition of price agreements only rely on chat data. Category i) exists because sometimes subjects leave before all subjects have explicitly agreed, so these subjects cannot be certain whether an agreement was reached. We construct Category ii) because subjects commonly suggest which price to set based on the previous period (e.g., "same" or "again?"). Category iii) captures cases where some subjects stop responding over time, an extreme case of which is captured by Category iv). We construct this final category because, in stable cartels, subjects occasionally stop communication altogether. However, lack of communication also occurs when subjects cease attempts to coordinate after unsuccessful previous attempts. Therefore, we resort to defining such cases based on past behavior. Table E1 and Table E2 give examples where lack of communication was classified as a price agreement and not as a price agreement, respectively.

Price agreements are present in 73.7 percent of all market-periods with a cartel (1,564 of 2,122 instances). The average market price of cartelized markets with price agreements is 71.28, while market prices of cartelized markets without price agreements (58.05) are now much closer to those of uncartelized markets (50.29). Figure G1 in Appendix G shows that the incidence of price agreements tracks the movement of successful coordination over time, while the level is higher, suggesting that our broader measure is more appropriate. Hence, we use the broad measure of price agreements for our main results and show that our main results are robust to alternative approaches.

Table E1: Lack of communication classified as price agreement

Sg - Period	Firm 1	Firm 2	Firm 3	Category
2-1	85? okay *Firm 1 exits chat*	75? 73.5 it is then *Firm 2 exits chat*	The max profit is at 73.5 Shall we do that? Cool *Firm 3 exits chat*	Explicit price agreement
2-1	price: *73.5*	price: *73.5*	price: *73.5*	market price: *73.5*
2-2	again? okay *Firm 1 exits chat*	73.5 *Firm 2 exits chat*	nice Yes *Firm 3 exits chat*	Explicit price agreement
2-2	price: *73.5*	price: *73.5*	price: *73.5*	market price: *73.5*
2-3	73.5 *Firm 1 exits chat*	73.5 *Firm 2 exits chat*	Yep *Firm 3 exits chat*	Explicit price agreement
2-3	price: *73.5*	price: *73.5*	price: *73.5*	market price: *73.5*
2-4	*Firm 1 exits chat*	*Firm 2 exits chat*	*Firm 3 exits chat*	Price agreement
2-4	price: *73.5*	price: *73.5*	price: *73.5*	market price: *73.5*

Table E2: Lack of communication not classified as price agreement

Sg - Period	Firm 1	Firm 2	Firm 3	Category
2-7	will we 90 uis bad 75 is bigger win *Firm 1 exits chat*	*Firm 2 exits chat*	90? do 90 *Firm 3 exits chat*	No agreement
2-7	price: *74.9*	price: *88*	price: *90*	market price: *74.9*
2-8	*Firm 1 exits chat*	*Firm 2 exits chat*	*Firm 3 exits chat*	No agreement
2-8	price: *73*	price: *74.9*	price: *69.99*	market price: *69.69*
2-9	*Firm 1 exits chat*	*Firm 2 exits chat*	*Firm 3 exits chat*	No agreement
2-9	price: *75*	price: *66*	price: *59.69*	market price: *59.69*

Table E3: Decreasing price agreement as cartel goes undetected in OVERCHARGE

Sg - Period	Firm 1	Firm 2	Firm 3
3-3	*Firm 1 exits chat*	last time 72 once more then we all switch to 47 *Firm 2 exits chat*	72 it is *Firm 3 exits chat*
3-3	price: *72*	price: *72*	price: *72*
3-4	*Firm 1 exits chat*	NOW 47 47 47 because we are in the 4th round *Firm 2 exits chat*	now what yeah lets go *Firm 3 exits chat*
3-4	price: *47*	price: *47*	price: *47*
3-5	Lets keep 72 *Firm 1 exits chat*	47 again it's last round we hae profit and 47 allows us to keep it play it safe 47 go *Firm 2 exits chat*	so 72? oka 47 *Firm 3 exits chat*
3-5	price: *47*	price: *47*	price: *47*

F Defection and cartel response

This appendix describes defections from price agreements and the responses that follow them. Recall that, in the experiment, defecting from a price agreement does not close the chat. Subjects can continue to communicate until detection or the end of the supergame–we denote this sequence of uninterrupted cartel periods a 'cartel spell'. Therefore, within a cartel spell we can potentially observe defection, renegotiation, and formation of a new agreement. Overall, defections are rare in our experiment, with the lowest rates observed in supergames 3 and 4. The bulk of undercutting happens in the first period that a price agreement is in place. Such defections reduce chances of future agreements and cartel formation in the same supergame, but only a minority of the subjects appear to follow a trigger strategy and renegotiation is common. Due to the limited observations on defections, we mostly refrain from making statistical comparisons and are careful to interpret between-treatment differences.

Table F1 shows that defections are rare and downward in the vast majority of all cases. Defections occur in 89 of 1,437 agreement subject-periods in REVENUE (6.19 percent), 93 of 1,725 in PROFIT (5.39 percent), and 67 of 1,530 in OVERCHARGE (4.38 percent). As can be seen from Figure G3, defections primarily occur in the first supergame and have all but been eliminated by the final supergame. Defections are downward in between 86.52 percent of all cases (in REVENUE) and 92.47 percent of all cases (in PROFIT). Investigating the chat suggests that the rare upward defections are due to mistakes, fear of fines, and attempts to signal cooperativeness after earlier undercuts. The remaining tables and figures in this appendix focus on downward defections.

Table F1: Defections at the subject-period and cartel-spell levels, by treatment

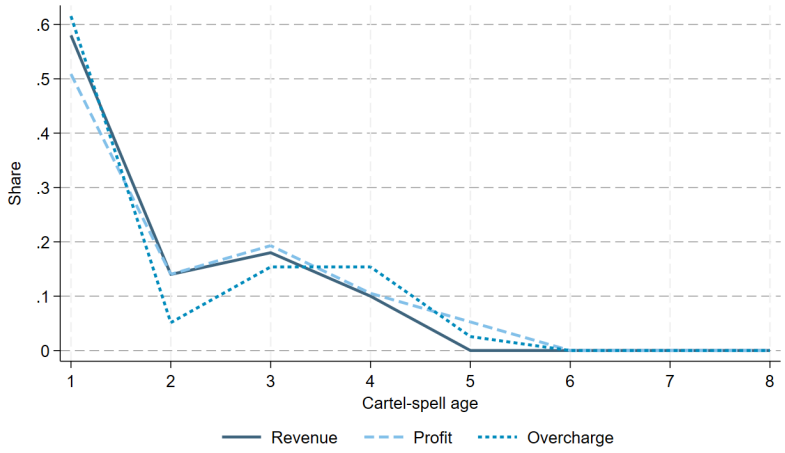
	Subject-period level			Cartel-spell level			
	Agreements count	Total defections	Downward defections	Spell count	Defection share	Second chance count	Second chance defection share
REVENUE	1,437	89	77	179	0.35	32	0.44
PROFIT	1,725	93	86	227	0.31	18	0.56
OVERCHARGE	1,530	67	61	201	0.24	16	0.44

Notes: Agreement count = subject-periods with a price agreement; Total defections = subject-periods in which submitted price differs from the price agreement; Downward defections = subject-periods in which submitted price is below the price agreement; Spell count = count of uninterrupted cartel periods ending at detection or supergame end; Defection share = share of spells with at least one defection; Second chance count = count of spells where a price agreement is formed after defection in the same spell; Second chance defection share = share of second chance spells with at least one defection in a later period than the first defection.

The share of cartel spells with at least one defection ranges from 0.24 in OVERCHARGE to 0.35 in REVENUE (Table F1). The share of cartel spells where subjects get a "second chance" in the sense that an agreement is formed after at least one subject previously defected from an agreement lies between 0.08 in PROFIT and 0.18 in REVENUE. A single defection, therefore, shuts down the cartel permanently more often than not. Moreover, second chances are less successful than the initial agreements, as roughly 50 percent of second chances experience another defection in all treatments.

Figure F1 shows that the majority of first downward defections occur in the first period in a cartel spell with a price agreement. Over time, the incidence of such defections drops and by six periods is zero regardless of supergame or treatment. This suggests that defections are initiated not so much due to cartel members who were initially happy to collude switching to defection, but rather are due to subjects who did not intend to collude in the first place joining a cartel with the sole intent of obtaining the defection profits.

Figure F1: Timing of first downward defection, by treatment



Notes: The share of all first downward defections that occur at each period, counting from the first period of a cartel spell with a price agreement.

We next focus on reactions to the first downward defection in a cartel spell. We further restrict the sample to initial defections that occur before the final period in a cartel spell, so that a reaction can be observed, and define the first period as the first instance of a price agreement within a cartel. As, by Figure F1, the majority of initial defections occurs in the first period, we do not exclude them. This implies that the sample is unbalanced when we compare pre- and post-defection behavior. Qualitatively, however, our insights also hold in the balanced panel.

Table F2 reports prices before, during, and after the first downward defection. Price agreements in defection and pre-defection periods are very similar, but market prices in

defection periods are substantially lower. The average market-price-determining initial defection is on average 91.59 percent of the average price agreement in REVENUE, 92.91 percent of the average price agreement in PROFIT, and 87.17 percent of the average price agreement in OVERCHARGE. Post-defection, market prices remain below the pre-defection agreement price and the defection-period market price in all treatments, especially in PROFIT and OVERCHARGE. The more limited decline in REVENUE can be explained by the possibility of detection putting upward pressure on prices even after defection, and, by Table F1, that subjects are most likely to form another and successful price agreement in that regime. Overall, Table F2 tells us that the first defection permanently reduces market prices, even below defection levels.

Table F2: Prices around the first downward defection, by treatment

	Pre-defection periods	Defection period		Post-defection periods
	Price agreement = Market price	Price agreement	Market price	Market price
REVENUE	74.42 (8.69)	74.99 (6.97)	68.68 (7.88)	67.41 (12.53)
PROFIT	74.17 (5.35)	73.52 (4.54)	68.31 (8.92)	56.46 (11.10)
OVERCHARGE	71.20 (5.75)	69.75 (7.97)	60.80 (9.83)	54.80 (10.73)
Observations	93	126	126	386

Notes: Average price agreements and market prices with standard deviations in brackets; The sample consists of all periods from the first price agreement on in cartel spells with a downward defection before the final period of the spell; Observations count market-periods.

We next investigate the extent to which subjects act in accordance with a grim trigger strategy. Our theory, like the broader literature on collusion, assumes that firms collude by using such strategies. However, the broader experimental literature on cooperation has found that subjects often base cooperation on other strategies (Dal Bó and Fréchette, 2018). Moreover, that our aggregate results are in line with our theoretical predictions is neither a necessary nor sufficient condition for subjects to act in accordance with a grim trigger strategy. As in the main text, let p^{PD} denote price on the punishment path of grim trigger: the unilaterally optimal post-defection but pre-detection price that subjects revert to for the remainder of the cartel spell after any defection from the price agreement if they use a grim trigger strategy: 58.75 in REVENUE and 47 in PROFIT and OVERCHARGE. We refer to the subject who commits the first downward defection in a cartel spell as the defector, and to the other subjects in that cartel as the suckers.

Table F3 reports the share of suckers and defectors who set a price below $1.1p^{PD}$ after the first downward defection, as well as the share of market prices below this threshold. We focus on $1.1p^{PD}$ as pricing equal to marginal cost rarely happens absent a cartel. The evidence

suggests grim-trigger for a sizable minority of subjects. The next-period shares of suckers acting in line with grim trigger lie between 0.25 in REVENUE and 0.45 in OVERCHARGE, while the corresponding shares for defectors range from 0.19 in PROFIT to 0.60 in OVERCHARGE. The shares drop further when we require pricing in line with grim trigger in all later periods of the cartel spell, ranging from 0.08 to 0.40 for suckers and from 0.10 to 0.53 for defectors. Pricing in accordance with grim trigger is more common in OVERCHARGE than in the other two treatments. The share of market prices below $1.1p^{PD}$ in REVENUE and PROFIT is considerably higher than when focusing on individual subjects, suggesting that low post-defection market prices are typically driven by a single subject undercutting rather than by coordinated reversion to p^{PD} within those treatments.

Table F3: Trigger pricing after the first downward defection, by treatment

	Share of suckers setting $p < 1.1p^{PD}$			Share of defectors setting $p < 1.1p^{PD}$			Share of market prices below $1.1p^{PD}$		
	Next period	All later periods	Obs.	Next period	All later periods	Obs.	Next period	All later periods	Obs.
REVENUE	0.25	0.08	85	0.28	0.11	47	0.39	0.23	44
PROFIT	0.28	0.15	89	0.19	0.10	52	0.49	0.38	47
OVERCHARGE	0.45	0.40	62	0.60	0.53	43	0.51	0.46	35

Notes: Sucker = Subject who does not commit the first downward-defection; Defector = subject who commits the first downward defection; p^{PD} is the price on the grim trigger punishment path; Next period is the first period in the cartel spell following the first downward defection and all later periods refers to all remaining periods in that spell, including the next period.

Table F4: Example of a trigger strategy from the PROFIT chat

Sg - Period	**Firm 1**	**Firm 2**	**Firm 3**
3-1	hi guys lets do 70 ok and whoever break it we just punish each other with47 73.5	 let do 73.5 73.5 got max profit	 hey ok i agree
3-1	Price: *73.5*	Price: *73.5*	Price: *73.5*

Table F4 shows anecdotal evidence from the chat supporting the suggestion that some subjects might use a grim trigger strategy. After forming a cartel, the subjects agree to set the monopoly price and punish each other by setting price equal to marginal cost in case someone defects.

While the previous evidence suggests that some subjects act in accordance with a trigger strategy, the data also indicates that many do not. Indeed, defection often leads to renegotiation instead of permanently breaking down a cartel. Table F5 compares the share of cartel periods with a price agreement pre- and post-defection, and also compares the share of cartels that reform after detection depending on whether a previous cartel spell in the market experienced a defection. The share of post-defection periods with a price agreement (between 0.17 in PROFIT and 0.33 in REVENUE) is significantly lower than the corresponding share before the first downward defection (between 0.73 in PROFIT and 0.89 in REVENUE). Still, after the first downward defection, between one in six and one in three cartel periods features a price agreement, indicating that subjects regularly continue to coordinate on a price within the same cartel spell.

Table F5: Collusion after the first downward defection, by treatment

	Price agreement share (Within cartel spell)			Recidivism share (Within supergame)		
	Pre defection		Post defection	No prior defection		Prior defection(s)
REVENUE	0.89 (0.32)	$>^{*}$	0.33 (0.36)	0.71 (0.46)	$>$	0.50 (0.51)
PROFIT	0.73 (0.45)	$>^{**}$	0.17 (0.30)	0.76 (0.43)	$>$	0.68 (0.47)
OVERCHARGE	0.85 (0.36)	$>^{***}$	0.22 (0.36)	0.82 (0.39)	$>^{***}$	0.57 (0.50)
Observations	114		386	217		109

Notes: Price agreement and recidivism shares; Pre defection = share of pre-defection cartel periods with a price agreement, in spells that later experience a downward defection; Post defection = share of post-first-downward-defection periods with a price agreement in a spell; No prior defection = share of markets with a cartel and price agreement at detection and no prior downward defection that form a new cartel immediately following detection; Prior defections = share of markets with a cartel and a prior downward defection at detection that form a new cartel immediately following detection; Observations count market-periods; Standard deviation in brackets; ***, **, and * indicate statistical significance from two-sided Wilcoxon signed-rank tests on matching-group averages at the 1%, 5%, and 10% level, respectively.

Recidivism after detection follows a similar pattern. Cartel spells ending in detection without a prior downward defection immediately re-form between 71 (REVENUE) and 82 (OVERCHARGE) percent of the time, while the corresponding percentages for spells ending in detection after a prior downward defection lie between 50 and 68. The decline is most

pronounced and statistically significant in OVERCHARGE, while differences in REVENUE and PROFIT do not reach significance, in part because of the small number of qualifying spells. Combined with the cartel stability data discussed in Section 4.3, which show that the share of stable cartels approaches one by the final supergame, the picture that emerges is one in which defection lowers but does not eliminate the prospect of further coordination, and renegotiated agreements are commonly successful.

Finally, Tables F6 and F7 provide anecdotal evidence on the different ways that subjects reform agreements following a defection. In Table F6, Firm 2 has just undercut an agreement to set 65 by setting 55. The next period, Firm 3, who acted as a sucker in the previous period, expresses dissatisfaction with the defection in strong terms, and suggests setting a price of 55. After Firm 2 adheres to that agreement and apologizes, a new agreement to set 65 is reached and adhered to.

Table F6: Example of an immediate renegotiation from the REVENUE chat

Sg - Period	**Firm 1**	**Firm 2**	**Firm 3**
2-5	*Firm 1 exits chat*	sorry i thought people were gonna stab in the back	firm 2 your mother doesnt love you i hope you know that 55 55
2-5	Price: *60*	Price: *55*	Price: *55*
2-6		sorry about that so what now alright	alright its okay your mother is starting to love you again 65 65
2-6	Price: *65*	Price: *65*	Price: *65*

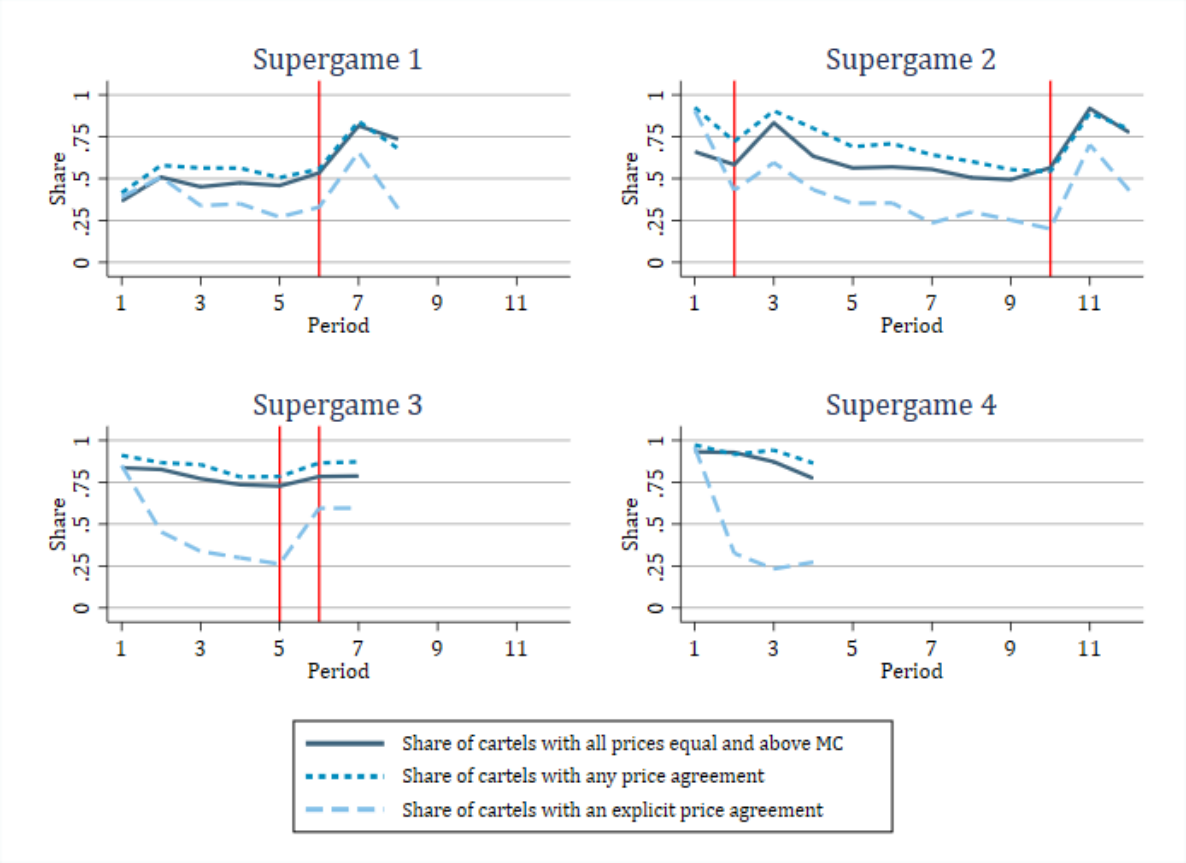
The chat in Table F7 reveals a different pattern: after a series of defections, first by Firm 1 and after that by Firm 3, the cartel members agree on a strategy in which the firm that has thus far acted as a sucker, Firm 2, sets a price that captures the entire market while the other two firms set a higher price. This way, all members have captured the market once at a price of 73, and after that the agreement to all set 74 is adhered to.

Table F7: Example of an agreement to compensate a sucker from the REVENUE chat

Sg - Period	**Firm 1**	**Firm 2**	**Firm 3**
4-2	sorry i didnt know if i can trust you last round they were messing me up ... 74	is there a problem man cmon ?? 	nice game lol nowo hhow can we rust you
4-2	Price: *74*	Price: *74*	Price: *73*
4-3	suggeation the one who didnt make anything yet puts 73 we put 74 put 73 bro we put 74	i havent ok i put 73	i good with that
4-3	Price: *74*	Price: *73*	Price: *74*
4-4	Price: *74*	Price: *74*	Price: *74*

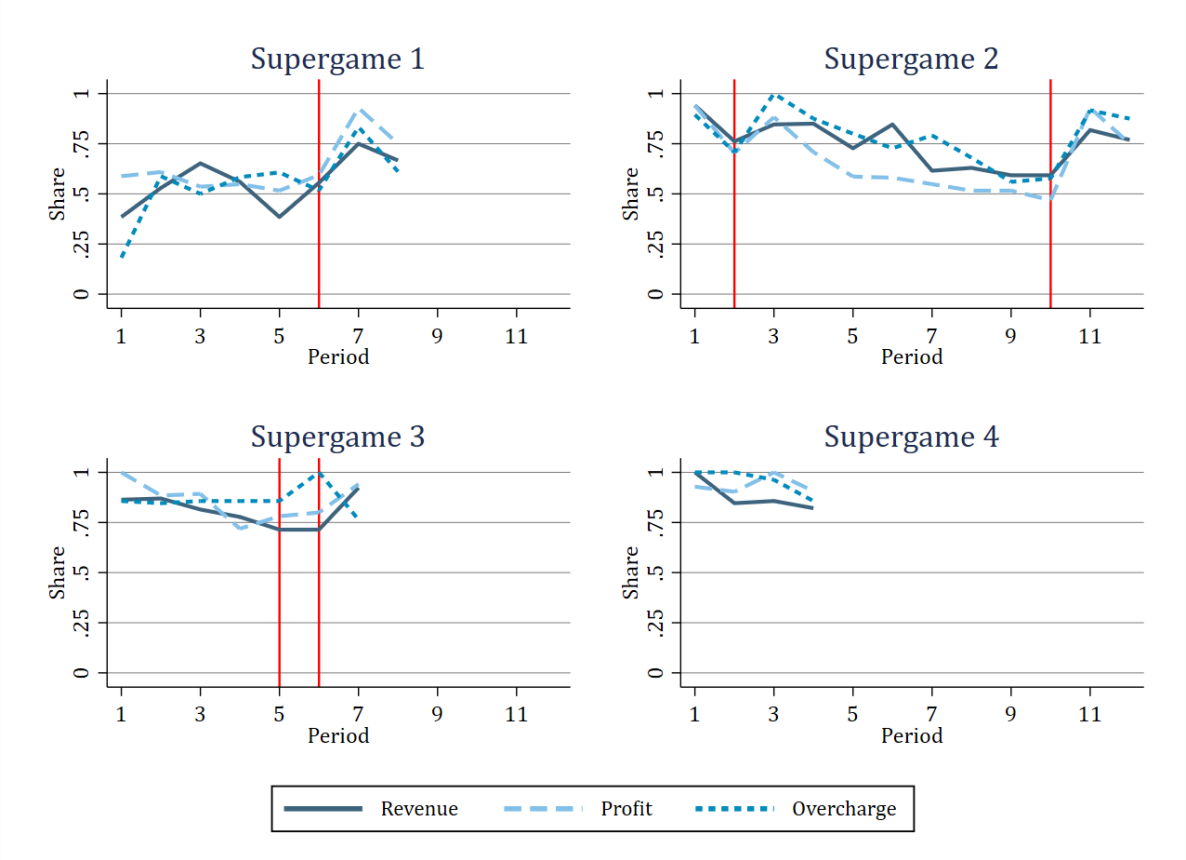
G Additional figures

Figure G1: Coordination and agreement incidence over time, by supergame



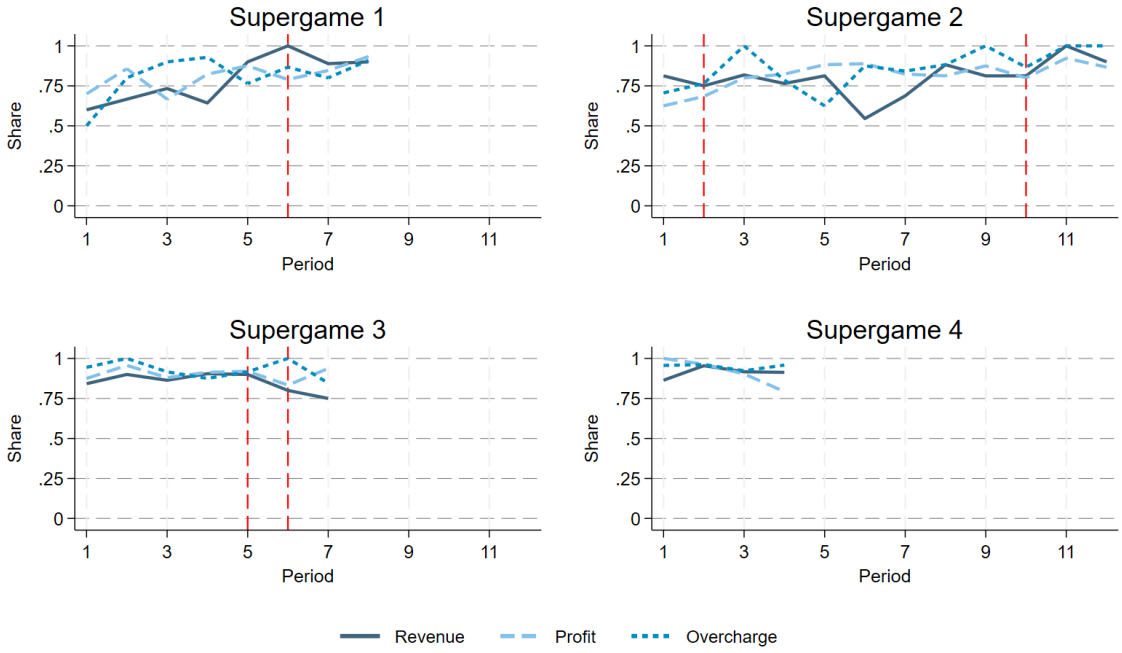
Notes: Share of cartels where all firms set equal prices above marginal cost, share of cartels with an explicit price agreement, and share of cartels with a price agreement in place, over time and by supergame. Red vertical lines indicate a period at the end of which all cartels are detected.

Figure G2: Price agreement incidence over time, by treatment and supergame



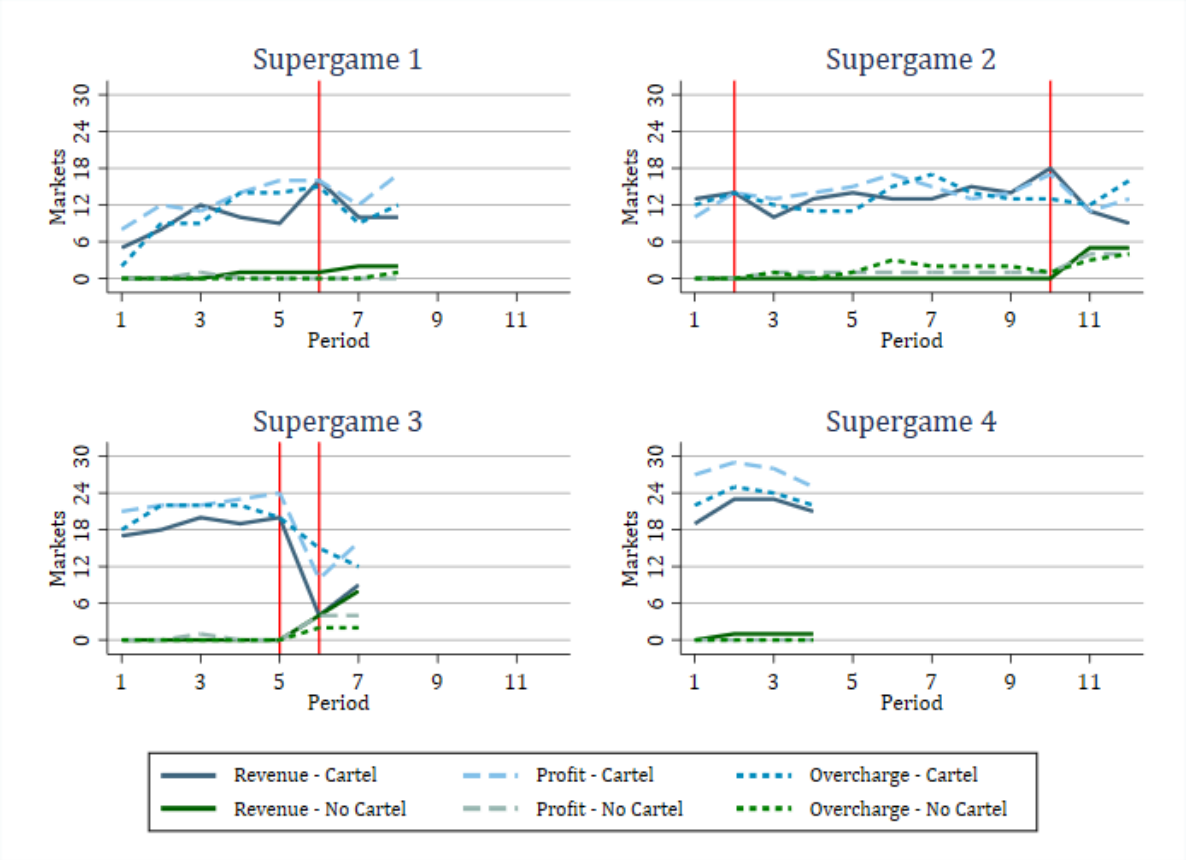
Notes: Average price agreement incidence over time, by treatment and supergame. Price agreement incidence = indicator for a price agreement (explicit or implicit) in a cartelized market-period. Red vertical lines indicate a period at the end of which all cartels are detected.

Figure G3: Cartel stability over time, by treatment and supergame



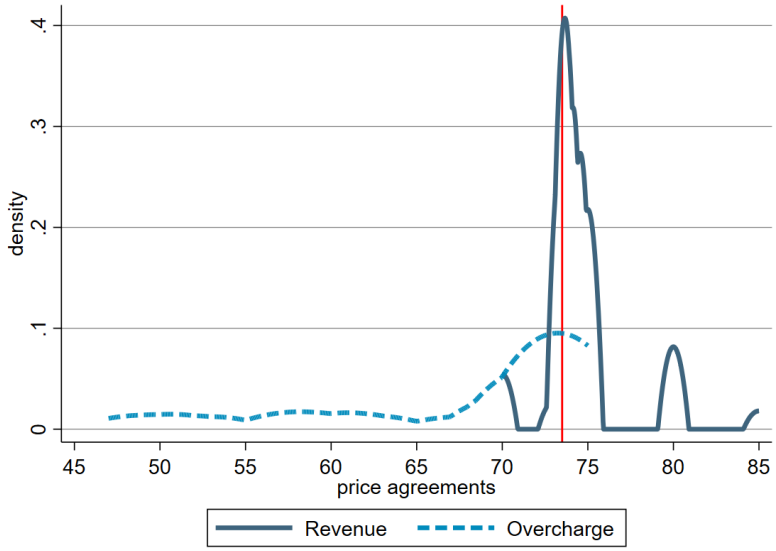
Notes: Average cartel stability over time, by treatment and supergame. Cartel stability = indicator for a cartel where all subjects set the agreed upon price in a market-period. Red vertical lines indicate a period at the end of which all cartels are detected.

Figure G4: Markets with $p_1 = p_2 = p_3 > 47$ over time, by treatment, cartelization, and supergame



Notes: Number of markets where all subjects set the same price (>47) over time, by treatment, supergame, and whether a cartel is active or not. Total number of markets per treatment and supergame is 30 for REVENUE and OVERCHARGE, and 33 for PROFIT. Red vertical lines indicate a period at the end of which all cartels are detected.

Figure G5: Distribution of agreed-upon market prices in the final supergame of REVENUE and OVERCHARGE



Notes: Distribution of agreed-upon market prices in the fourth and final supergame of REVENUE and OVERCHARGE. Red vertical lines indicate the joint-profit-maximizing price in the absence of cartel fines.

H Additional tables

Table H1: Regressions of price agreements on number of consecutive periods with agreement without detection in REVENUE and OVERCHARGE

	REVENUE		OVERCHARGE	
Detectionless periods	0.89	1.53*	-3.65***	-3.47**
	(0.52)	(0.74)	(0.87)	(1.11)
(Detectionless periods)^2	-0.06	-0.13	0.31*	0.31**
	(0.07)	(0.07)	(0.14)	(0.12)
Betrayal last period		-2.12**		-1.63
		(0.90)		(3.69)
Messages sent		0.18*		0.00
		(0.09)		(0.18)
Reformed cartel		0.16		0.48
		(0.67)		(1.63)
Constant	73.33***	71.49***	74.46***	74.12***
	(0.84)	(1.65)	(1.25)	(3.07)
Observations	392	392	421	421
R^2	0.03	0.06	0.12	0.12
Market Fixed Effects	Yes	Yes	Yes	Yes

Notes: Dependent variable = Price agreement; Detectionless periods = Number of consecutive periods with an agreement leading up to the current period without cartel detection; Betrayal last period = Indicator equal to one if at least one cartel member defected from the price agreement in the previous period; Messages sent = Mean number of messages sent in the chat this cartel-period; Reformed cartel = Indicator equal to one if the current cartel was formed after a previous cartel had been detected; Based on the final three supergames; Standard error clustered at the matching group in parentheses; ***, **, and * indicate statistical significance at the 1%, 5%, and 10% level, respectively.

Table H2: Regressions of price agreements on cartel risk appetite in REVENUE and OVERCHARGE

	REVENUE		OVERCHARGE	
Cartel risk appetite	-0.38	-0.35	2.11*	2.01*
	(0.38)	(0.53)	(1.04)	(0.93)
Cartel irrationality		-0.32		0.94
		(0.94)		(1.72)
Majority male		0.21		-1.21
		(0.90)		(2.05)
Detectionless periods		1.02		-3.89**
		(0.73)		(1.27)
(Detectionless periods)^2		-0.10		0.36**
		(0.09)		(0.14)
Betrayal last period		-1.86*		-0.63
		(0.84)		(3.27)
Messages sent		0.08		-0.01
		(0.10)		(0.14)
Reformed cartel		0.20		-1.30
		(0.55)		(1.63)
Constant	76.76***	74.60***	58.56***	65.53***
	(1.79)	(2.44)	(5.66)	(5.07)
Observations	392	392	421	421
R^2	0.01	0.03	0.06	0.13
Supergame Fixed Effects	No	Yes	No	Yes

Notes: Dependent variable = Price agreement; Cartel risk appetite = Mean (across cartel members) row of the Holt and Laury risk preference test where higher values indicate riskier choices; Cartel irrationality = Indicator that equals one if any cartel subject switched multiple times in the Holt and Laury risk preference test; Majority male = Indicator if at least two of the three cartel members are male; Detectionless periods = Number of consecutive periods with an agreement leading up to the current period without cartel detection; Betrayal last period = Indicator equal to one if at least one cartel member defected from the price agreement in the previous period; Messages sent = Number of messages sent in the chat; Reformed cartel = Indicator equal to one if the current cartel was formed after a previous cartel had been detected; Based on the final three supergames; Standard error clustered at the matching group in parentheses; ***, **, and * indicate statistical significance at the 1%, 5%, and 10% level, respectively.

Table H3: Price agreements and cartel stability using explicit agreements, across treatments

	Agreement incidence (Cartels)	Price agreement (Cartels with an explicit price agreement in place)	Cartel Stability (Cartels with an explicit price agreement in place)	Market price (Cartels with an explicit price agreement in place)
REVENUE	0.38 (0.49)	74.80 (5.84)	0.79 (0.41)	73.53 (6.74)
	$\wedge$	$\vee^{**}$	$\wedge$	$\vee^{*}$
PROFIT	0.36 (0.48)	72.35 (5.63)	0.83 (0.38)	71.52 (6.33)
	$\wedge^{**}$	$\vee^{***}$	$\wedge$	$\vee^{**}$
OVERCHARGE	0.48 (0.50)	68.09 (9.34)	0.88 (0.32)	67.12 (9.66)
	$\vee^{**}$	$\wedge^{***}$	$\vee^{*}$	$\wedge^{***}$
REVENUE	0.38 (0.49)	74.80 (5.84)	0.79 (0.41)	73.53 (6.74)
KW test	$p = 0.042$	$p = 0.000$	$p = 0.167$	$p = 0.002$

Notes: Agreement incidence = Indicator for an explicit price agreement in a market-period; Price agreement = Price that the cartel has explicitly agreed to set in a market-period; Cartel stability = Indicator for whether all three subjects in a cartel have set the agreed-upon price in a market-period; Market price = Lowest submitted price in a market-period; Standard deviation in brackets; ***, **, and * indicate statistical significance from pairwise, two-sided MWU tests at the 1%, 5%, and 10% level, respectively; Bottom row reports Kruskal-Wallis p-values across all three treatments.

Table H4: Measures of markets where $p_1 = p_2 = p_3 > 47$, across treatments

	Share (All markets)	Share (No cartels)	Share (Cartels)	Market price (Cartels)
REVENUE	0.50 (0.50)	0.12 (0.32)	0.65 (0.48)	75.05 (5.19)
	$\wedge$	$\vee$	$\vee$	$\vee^{***}$
PROFIT	0.53 (0.50)	0.11 (0.32)	0.65 (0.48)	72.52 (5.11)
	$\wedge$	$\vee$	$\wedge$	$\vee^{**}$
OVERCHARGE	0.53 (0.50)	0.09 (0.29)	0.70 (0.46)	68.51 (8.94)
	$\vee$	$\vee$	$\vee$	$\wedge^{***}$
REVENUE	0.50 (0.50)	0.12 (0.32)	0.65 (0.48)	75.05 (5.19)
KW test	$p = 0.823$	$p = 0.933$	$p = 0.869$	$p = 0.000$

Notes: The fraction of markets where $p_1 = p_2 = p_3 > 47$, and market prices of cartels, across treatments; Standard deviation in brackets; ***, **, and * indicate statistical significance from pairwise, two-sided MWU tests at the 1%, 5%, and 10% level, respectively; Bottom row reports Kruskal-Wallis p-values across all three treatments.