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

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

March 2024

Abstract

Competition authorities impose substantial penalties on firms that participate in illegal horizontal price-fixing agreements. We investigate how basing cartel fines on either revenue, profit, or price overcharge influences cartel and market prices, as well as cartel incidence and stability. In an infinitely repeated Bertrand oligopoly game, we show that revenue bases incentivize firms to increase prices above the monopoly price, while only overcharge bases incentivize price reductions. Cartels are stable for a smaller range of discount factors when fines are based on overcharges than on other bases. We test these predictions in a laboratory experiment where subjects can opt into cartels, which allows them to discuss pricing at the risk of being detected and fined. We equalize excepted fines across treatments so that our results originate in the base of the fine, not the size of the fine. Variation in market prices across treatments is determined by cartel prices, which follow the theoretical predictions, but cartel incidence is equal across fining regimes. Our results suggest benefits from authorities moving away from revenue bases towards profit or overcharge bases.

Keywords: Antitrust; Cartel; Collusion; Repeated game; Experiment

JEL Codes: C73; C90; D43; K21; L41

*We are grateful to Joe Harrington, Evgenia Motchenkova, Maarten Pieter Schinkel, Yossi Spiegel, and seminar participants at the University of Amsterdam, KU Leuven, and the 2024 MaCCI (Mannheim) conference for useful comments and discussions. Engilbertsson thanks the University of Amsterdam Research Priority Area in Behavioral Economics (grant EB-1145). Treuren gratefully acknowledges support from ERC Consolidator grant 816638. This experiment was preregistered with the AEA RCT Registry (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

Competition authorities and courts regularly impose substantial penalties on firms participating in illegal horizontal price-fixing agreements.[1] Throughout antitrust jurisdictions, such fines are based on cartel members' revenue.[2] However, it is an established fact in the theoretical literature that revenue bases can increase cartel prices (e.g., Bageri et al. (2013); Katsoulacos and Ulph (2013)). This discrepancy between theory and practice raises the question of the relative performance of different fining regimes.

The primary aim of this paper is to determine how different fining bases influence cartel and market prices. In particular, we are interested in assessing whether the theoretical concerns of revenue-based fines are empirically warranted, and comparing the performance of revenue-based fines to that of viable alternatives. While these questions are empirical at heart, observational data is of limited help. Therefore, we address these question using a theoretical model and a laboratory experiment. We compare revenue-based fines to two alternatives for which a legal basis exists in the US fining guidelines: fines based on the (incremental) profit of cartel members, and fines based on cartel members' price overcharge with respect to the competitive price.[3]

We first consider an infinitely repeated Bertrand game where firms can use trigger strategies to support prices above the one-shot Nash equilibrium price 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, regardless of their behavior, 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. This theoretical model isolates key factors of the antitrust policy under consideration and closely follows existing theoretical work.

Our theoretical results on cartel prices align with the broader literature (e.g., Bageri et al. (2013); Katsoulacos et al. (2015)). Revenue bases incentivize cartel members to increase their price above the no-antitrust monopoly price. In contrast, basing the fine on a cartel member's

[1] For instance, the European Commission awarded a 3.8 billion euro fine to truck producers in the Trucks case (2016/2017). Further examples include the 2.5 billion dollars and 1.4 billion euros 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 yen imposed by Japan's competition 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 (2021), the base fine for bid-rigging, price-fixing, and market-allocation agreements is "20 percent of the volume of affected commerce" (p.311).

[3] For non-antitrust criminal purpose organizations, the pecuniary gain to the organization from the offense – profit – or the pecuniary loss from the offense caused by the organization – damages – form the base of the fine (United States Sentencing Commission, 2021, p.526).

overcharge reduces the optimal cartel price compared to the monopoly price, while a profit base leaves the optimal cartel price unaffected. Intuitively, in an overcharge regime, the fine is strictly increasing in price, which gives the cartel an incentive to mitigate the price; 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; fines based on revenue serve as a tax pushing prices up.

In contrast to earlier findings, our model suggests that cartel stability is lowest when cartel fines are based on the price overcharge. The reason is that defection in the revenue and profit regimes increases the expected fine, while defection decreases the price overcharge, and hence the expected fine in the overcharge regime. 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 in line with antitrust practice (Buccirossi and Spagnolo, 2007).

To confront 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. 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.[4] To empirically shed light on the generality of our theoretical results, therefore, we use a laboratory experiment to test our predictions.

Of course, any theoretical model or laboratory experiment only partially resembles real cartels, but our approach allows us to overcome significant challenges posed by field data. First, it is difficult for the researcher to observe cartels in the field because of their illegal nature. 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 regarding the various fining regimes, which is difficult to obtain in the field because it is hard to establish exogenous variation and to measure variables of interest like marginal costs and demand. Finally, experiments allow researchers to optimize internal validity as laboratory control ensures the theory's assumptions are met as closely as possible. For these reasons, laboratory experiments are widely employed as wind-tunnel tests of theory-based policy recommendations (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

[4]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."

that individual's revenue, profit, or price overcharge, respectively. We equalize excepted 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 formation, cartel incidence, and cartel recidivism.

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 formation, incidence, and recidivism across treatment. Therefore, market price differences across treatments are entirely determined by cartel prices. Our findings suggest benefits from antitrust authorities moving away from revenue and profit bases towards overcharge bases. We conclude by arguing that such a change is realistically implementable given current legal and institutional constraints.

We contribute to a strand of literature studying how cartel fining regimes influence market outcomes. in particular, revenue regimes are 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 Jr. (2004), and Harrington Jr. (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.[5] 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. Our main contribution to this theoretical literature is to test its predictions empirically by investigating equilibrium selection in the lab.[6]

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)), or based on revenue (e.g., Apesteguia et al. (2007); Hinloopen and Soetevent (2008)). Across different treatments, the size of the fine varies but the base of the fine is fixed. In contrast, we hold the size of the

[5] In contrast to our model, Katsoulacos et al. (2015) assume that defectors cannot be fined 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).

[6] As we compare commonly-studied fining bases while holding fixed the level of the fine, we remain silent on the optimal level and design of fines (e.g., Buccirossi and Spagnolo (2007); Katsoulacos and Ulph (2013); Houba et al. (2018)).

expected fine fixed and vary the fining base, to study the effect of fining structure on cartel behavior. As we are the first to investigate this question experimentally, and to limit the demands we place on experimental subjects, we abstract from leniency programs. We view the inclusion of leniency as an important avenue for future work.[7]

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. (2023b) for overviews), spillovers from legal cooperation in some markets to tacit collusion in others (e.g., Duso et al. (2014); Sovinsky (2022)), the effect of market transparency programs on collusion (e.g., Vega-Redondo (1997); Byrne and De Roos (2019)), and the impact of auction design on bid rigging (e.g., Robinson (1985); Marshall and Marx (2007)). Theoretical predictions in these domains are routinely tested in laboratory experiments.[8]

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 rate exceeding the critical discount rate 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 rate and the critical discount rate. While laboratory experiments on collusion typically equalize critical discount rates across treatments, this is impossible in our theoretical model as overcharge based fines always have higher critical discount rates than the other two regimes. However, we do not find differences across fining regimes in any of our measures of collusion, and do report 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 theoretical results on which we base our hypotheses. Section 3 contains our experimental design, experimental procedures, and hypotheses. Section 4 gives our experimental findings. Concluding remarks on implications of our findings and implementability are in Section 5.

[7] In a laboratory experiment on collusion, Fonseca et al. (2022) let subjects' payoff have a fixed component and a revenue-dependent component, and vary whether cartel fines are based on the revenue or subjects' total remuneration. Basing fines on total remuneration is found to reduce cartel formation rates.

[8] The experimental literature provides several lessons regarding the effects of the various competition policy instruments discussed above. For example, Hinloopen and Soetevent (2008) and Bigoni et al. (2012, 2015) observe leniency programs having the desired effects on cartel formation, cartel discovery, and the price; Normann et al. (2015) and Hinloopen et al. (2023a) report spillovers from legal cooperation in one experimental market to tacit collusion in another; Huck et al. (1999, 2000) and Offerman et al. (2002) find that transparency about competitors' actions increases competition; Hinloopen et al. (2020) observe more stable bidding rings and lower revenue in the English auction than in the first-price sealed-bid auction.

2 Theoretical framework

Out 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. In order to align our experiment with prior theory, we borrow heavily from existing work on antitrust penalties that relies on similar assumptions (e.g., Motta and Polo (2003); Aubert et al. (2006); Katsoulacos et al. (2015)).

Consider an infinitely repeated homogenous-goods oligopoly game with $n \geq 2$ firms that maximize expected profit and have a common discount rate $\delta \in (0, 1)$. Each firm i sets a price each period t, $p_{it} \in [0, \bar{p}]$. Market demand in period t, $q_t = q(p_t)$, depends on the market price, which is the lowest price set that period, $p_t = \min_i p_{it}$, and satisfies $q(\bar{p}) = 0$. 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 n firms that set the lowest price in a given period share the resulting market demand equally: $q_{it} = \frac{q_t}{n}$ if $p_{it} = p_t \; \forall \; i$. Firms that do not set the lowest price face no demand: $q_{it} = 0$ if $p_{it} > p_t$.

Absent explicit communication, we assume that the static Bertrand Nash-equilibrium is played each period: $p_{it} = c$, such that profit $\pi_{it} = 0 \; \forall \; (i, t)$. This assumption is in line with our experimental design where $n = 3$, as the literature finds that, absent communication, prices typically converge close to the static Nash-equilibrium for three or more players (e.g., Dufwenberg and Gneezy (2000); Fonseca and Normann (2012); Chowdhury and Crede (2020)). We denote the outcomes of this competitive benchmark by p^N and π^N.

Firms can choose to form a cartel and explicitly communicate, which allows market prices above p^N to emerge.[9] Such collusive prices are supported by a grim trigger strategy profile 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 unilaterally maximizing profits forever, leading to market price p^N. We assume that cartels coordinate on the joint-profit-maximizing price. In the absence of antitrust, unrestricted joint profit maximization requires setting the monopoly price $p^M = p^M(c)$ $(c < p^M < \bar{p})$. Note that our assumptions imply that $p^M(c)$ increases with c.

Cartels leave a paper trail that the antitrust authority can detect. In particular, once a cartel has been formed, it is detectable and remains so in later periods until the antitrust authority has discovered 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 and fined. This is in line with reality, and our

[9] 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 often assume that communication is required for collusion to emerge (e.g., McCutcheon (1997); Motta and Polo (2003)).

experiment, as defectors do not face reduced fines in either the US or the EU (Buccirossi and Spagnolo, 2007). In addition, the paper trail used to convict cartels typically originates years prior to detection and conviction.[10] In line with profit maximization, we assume that cartels immediately reform after the antitrust authority has discovered and convicted them, as long as no firm has previously defected from the agreement.[11]

Each period, after prices are set and the market clears, the antitrust authority detects, prosecutes, 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. The main focus of this article is how different choices for B_{it} affect cartel pricing and stability. In theory, cartels can be deterred entirely by ensuring that F_{it} is sufficiently large. For instance, in the spirit of Becker (1968), by imposing a penalty which ensures that the expected profit of forming a cartel is non-positive. 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, and bankruptcy concerns put downward pressure on fines in particular instances (Buccirossi and Spagnolo, 2007; Houba et al., 2018).[12]

In the absence of side payments, and due to the symmetric nature of firms, we focus on collusive agreements where all firms set the same price and share output. Denote the single-period before-fine profit of an individual firm in a cartel whose members all set price p^C by π^C, and the concomitant fine upon detection by F^C. A firm's expected present value of participating in the cartel and its expected present value of the competitive benchmark are, respectively, given by

$$V^C = \frac{\pi^C - \alpha F^C}{1 - \delta} \qquad \text{and} \qquad V^N = \frac{\pi^N}{1 - \delta}. \tag{1}$$

Denote the optimal defection of a cartel member who assumes that all other firms will set p^C by p^D, and the resulting profit and fine by π^D and F^D, respectively. If the cartel is convicted immediately after defection, all firms revert to playing the static-Nash price forever. However, if the cartel is not immediately convicted after defection, firms select the price that maximizes their unilateral profit in the static game, denoted by p^{PD}, until the cartel is detected. Denote the concomitant profit and fine by π^{PD} and F^{PD}. Prices p^{PD} and p^N could differ because a cartel can still be convicted and fined post-defection, which can incentivize firms to set a price different from p^N. After a firm has defected from the

[10] Kwoka and White (2018) provides a description of prominent cartel cases.

[11] A similar assumption is made in Motta and Polo (2003) and Chen and Rey (2013), among others.

[12] To not further burden subjects in what is already a complicated experiment, we have opted to follow the literature and let fines depend on current conduct only. Note that we do introduce dynamic detection. In reality, fines are often based on the estimated duration of the cartel. Introducing a dynamic component to fines substantially complicates the analysis and is, therefore, typically ignored in the literature. Notable exceptions are in Harrington (2004; 2005).

agreement *and* the cartel has been detected, all firms set p^N again. Therefore, the expected present value of defection is given by

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

where the last equality follows from the assumption that $\pi^N = 0$, and the fact that expected profit in the Bertrand game is 0 when firms set prices unilaterally.

For stable cartels to be part of a subgame-perfect Nash equilibrium, two conditions must be met. 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, set the price that solves

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

We next analyze how different choices for F^C influence cartel pricing and stability. We consider three fining regimes. First, fines based on a firm's revenue. Second, fines based on incremental profit enjoyed by or consumer damages caused by a firm. Finally, fines based on a firm's price overcharges relative to the competitive price. The first two fining regimes represent fining practice worldwide, particularly in Europe and the US. To our knowledge, fines based on the overcharge have not been implemented in practice but have been theoretically shown to have advantages over the other two fining regimes (e.g., Katsoulacos et al. (2015)). Indeed, below we show that overcharge-based fines not only reduce prices of stable cartels compared to the other fining regimes – as is well known in the literature – but also destabilize cartels by making defecting more attractive – a novel finding.

2.1 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. Revenue bases

are widely used 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.[13] While US 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 (US Sentencing Commission, 2021, p.311).

Let p_R^C denote the price set by a stable cartel, and let δ_R^* denote the critical discount rate above which cartels are stable if fines are based on revenue.[14]

Proposition 1. *If fines are based on revenue:*
i) *The price set by a stable cartel exceeds the monopoly price:* $p_R^C > p^M$.
ii) *Cartel stability does not depend on antitrust:* $\delta_R^* = \frac{n-1}{n}$.

Proposition 1 shows that revenue-based fines have a perverse price effect. The fine acts as a tax on revenue, reducing marginal revenue but leaving marginal cost unaffected. 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. It is an established fact in the theoretical literature that revenue bases can increase cartel prices (e.g., Bageri et al. (2013); Katsoulacos and Ulph (2013)).

A firm's best response to all other firms setting p_R^C is not the monopoly price p^M, as defectors can be detected and fined. Instead, 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 profit by n, this is the only relevant parameter for cartel stability. As n increases, collusion becomes more difficult in the sense that the critical discount rate increases.

An additional effect of revenue-based fines is the fact that even following a defection, cartel prices will remain above the competitive benchmark. Since a cartel can still be detected and fined post-defection, and since setting p^N results in positive revenue but no profit, cartel members will set the price $p_R^{PD} = \frac{c}{1-\alpha r^R} > p^N$, until the cartel is detected.

2.2 Fines based on incremental profit

We implement incremental 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. The incremental profit base closely resembles the US guidelines for non-antitrust offences, where the base fine is the maximum of the incremental profit due to the offense – pecuniary gain

[13] 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).

[14] All proofs are in Appendix A.

– and the caused damages – pecuniary loss (US Sentencing Commission, 2021, p.526).[15] Note that, in the US, the standard formula for consumer damages in cartel cases is $(p^C - p^N)q^C$ (Harrington, 2014). That is, only damages for goods that were sold are taken into consideration. As in our Bertrand setting $p^N = c$ and $\pi^N = 0$, profit bases, incremental profit bases, and consumer damage bases are all identical.[16]

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

Proposition 2. *If fines are based on incremental profit or consumer damages:*
i) *The price set by a stable cartel equals the monopoly price: $p_\pi^C = p^M$.*
ii) *Cartel stability does not depend on antitrust $\delta_\pi^* = \frac{n-1}{n}$.*

As the incremental profit-based fine acts as a tax on profit, it does not affect the profit-maximizing price, and the cartel sets the monopoly price. Similar results on cartel pricing 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 rate only depends on n and is identical to the revenue-based critical discount rate. To investigate the generality of this result, consider what happens if consumer damages and incremental profit are bench-marked against a but-for price p^{BF} above marginal costs instead of against p^N. If fines are based on a firm's profit, Proposition 2 applies. It is straightforward to show that a cartel in an incremental profit-based regime still sets p_π^C, but that the critical discount rate is *below* δ_π^* as defecting scales up the fine more than the before-fine profit. In a consumer-damages-based regime, the critical discount rate is still given by δ_π^*, but the cartel price lies below p_π^C. This echoes Harrington (2005), who shows – in a model where detection depends on price changes and penalties accumulate over time – that the steady-state cartel price is below the monopoly price when fines are based on damages, unless those damages are proportional to profit.

2.3 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: $F_{it} = r^O(p_{it} - p^N)\frac{q^N}{n}$, where r^O is the exogenous penalty rate. Note that we multiply the overcharge by the competitive output of an individual firm, following Katsoulacos et al. (2015), who introduced

[15] Penalty rates are based on a culpability score, which depends on several aggravating and mitigating factors. For price-fixing cases, the minimum penalty rate is at least 0.75, and the maximum penalty rate is at most 4. Penalty base and the two penalty rates jointly determine a range of possible fines from which courts select a fine based on the perceived seriousness of the infringement. As in the EU, bankruptcy concerns and leniency applications could substantially lower the amount paid by the defendant (US Sentencing Commission, 2021).

[16] In addition, the results in this section carry over to fixed fines, as is the case in the setting studied by Katsoulacos et al. (2015).

this overcharge-based fine. In principle, we could multiply the overcharge by any constant the cartel cannot control. What sets overcharge-based fines apart from fines based on (incremental) profit is that the latter multiply the overcharge by the firm's output rather than a constant. This distinction is crucial for generating the attractive properties of an overcharge-based fine, as a price increase will put upward pressure on the fine in both regimes but, in addition, put downward pressure on profit-based fines by reducing the cartel's output. Although several jurisdictions mention the overcharge as relevant for determining the fine, to our knowledge, an overcharge base has not been implemented in practice even though it directly targets the distortion created by the cartel.[17]

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 rate above which cartels are stable if fines are based on the overcharge.

Proposition 3. *If fines are based on the overcharge:*
i) *The price set by a stable cartel lies below the monopoly price:* $c \leq p_O^C < p^M$.
ii) *The critical discount rate increases in antitrust:* $\delta_O^* \equiv \frac{(n-1)\pi_O^C}{n\pi_O^C - \alpha F_O^C}$

An overcharge-based fine reduces the cartel price compared to p^M 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, a price reduction decreases the 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 defectors can be fined and cartels formed in previous periods can be detected.

In contrast to the revenue and profit regimes, defecting from the cartel agreement does not increase the fine in an overcharge-based regime, but does increase before-fine profit n-fold. This has two effects. First, antitrust affects the critical discount rate, which increases in both the the penalty rate and the detection probability. Second, defection is incentivized compared to the other two fining regimes, where defection increases both the before-fine profit and the fine by a factor n.

For a range of discount rates – $\bar{\delta} \equiv \frac{n-1}{n-\alpha r^O} < \delta < \delta_O^*$ – stable cartels that set the unconstrained cartel price do not exist, but stable cartels that set a lower price can be part of a subgame-perfect Nash equilibrium. This price – given in the proof of Proposition 3 – is always below the unconstrained price, and the $\bar{\delta}$ is always higher than the critical discount rates of the revenue and profit regimes. Therefore, our focus on the comparison of unconstrained cartel prices and stability conditions across fining regimes in the remainder of the paper is without loss of generality.

[17] US 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 (US Sentencing Commission, 2021, p.312).

2.4 Comparison of fining regimes

Propositions 1 to 3 imply the following ranking across fining regimes of the critical discount rate and cartel prices.

Corollary 1.
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 rate 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. Incremental profit-based fines leave prices unaffected as they are essentially a proportional tax on firm profit.

A novel result is that overcharge-based fines always increase the critical discount rate 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 incremental profit-based regimes. Hence, the only stability-relevant parameter is n: antitrust is unrelated to cartel stability. If deterrence is an antitrust objective, revenue and profit regimes fail to deliver. Overcharge-based fines have the desirable property that defectors do not see their fine increase proportionally with the before-fine benefits of defecting so that more stringent antitrust measures deter more cartels.

Consider a distribution of discount rates δ over firms segmented by market. Corollary 1 implies that the average price – averaged over stable cartels and competitive markets – follows the same ranking as Corollary 1(i). 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 incremental profit-bases induce identical deterrence, prices of undeterred cartels are higher when fines are based on revenue.

Corollary 1 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 might not hold in practice, such as the grim trigger strategy and expected profit maximization.[18] Therefore, we conduct a laboratory experiment to test the validity of the predictions in Corollary 1. An experiment allows us to randomize the fining regime and accurately track cartel formation, demise, and pricing. In contrast,

[18]We discuss another potential reason – implementability – in Section 5 and argue that it is insufficient to explain the discrepancy between theory and practice.

data on discovered cartels suffer sample selection bias, and identifying the cartel's duration and marginal costs is challenging.

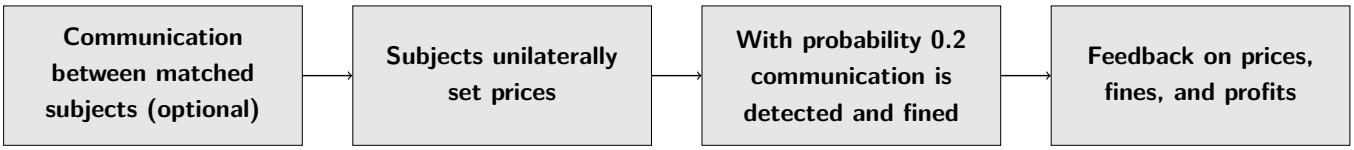


Figure 1: Timeline of a single period

3 Experimental design, procedures, and hypotheses

3.1 Experimental design and procedures

Our experiment tests the pricing effects of the different cartel fining regimes studied in Section 2. Subjects play an infinitely repeated Bertrand triopoly game.[19] Each period in each treatment follows the timeline displayed in Figure 1. Subjects first engage in optional communication 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 playing the next period. We now explain all phases of a period in more detail.

After being matched with two subjects, each subject unilaterally votes for or against cartel formation. 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 competition authority or subjects are re-matched.[20] 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 (e.g., Cooper and Kühn (2014); Harrington et al. (2016)).[21] In addition, unrestricted

[19]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.

[20]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.

[21]According to Cooper and Kühn (2014, p.250), "Limited message treatments may miss the types of

communication using natural language is a central feature of discovered hard-core cartels (e.g., 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 no more than two players (Huck et al., 2004). 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.[22] With more than two players, the market price in Bertrand experiments closely resembles the static Nash equilibrium in the absence of explicit communication (e.g., Dufwenberg and Gneezy (2000); Fonseca and Normann (2012)). Triopolies are, therefore, the simplest setting where tacit collusion is unlikely to occur.

We believe a Bertrand game stimulates subjects' understanding of the game. In addition, a Bertrand setting is the norm in existing theory on cartel fines and is used in many oligopoly experiments (e.g., Dufwenberg and Gneezy (2000); Fonseca and Normann (2012); Hinloopen et al. (2023a)). Fines can vary substantially over periods in this setting as defectors capture the entire market. While unrealistic, 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 firms are challenging to implement and place strong demands on experimental subjects.[23] To further aid subjects' understanding, an on-screen profit calculator was made available.

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.[24] 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`, $F_{it} = 1.85(p_{it} - 47)\left(\frac{53}{3}\right)$.[25] Penalty rates are selected to equalize fines across treatments. 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

messages that actually matter and the available messages are used differently than they would be in a natural conversation."

[22]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 are lowest for Bertrand duopolies (Fonseca and Normann, 2012).

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

[24]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.

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

including a treatment without antitrust. We are interested in studying how the cartel fining regime influences cartel pricing and stability 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. This includes Europe, North America, most South American countries, and many others in Africa, Asia, and Oceania (DLA Piper, 2020).

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 (2018)).[26] 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.[27] 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 rate $\delta = 0.9$.[28]

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. Students were recruited by public announcement. In total, 279 students,

[26]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)).

[27]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.

[28]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.

mainly from the university's undergraduate population, participated across 21 sessions covering the three treatments. Each session had either 9 or 18 participants.[29] 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 C). Table 1 lists the number of subjects, supergames, and observations across treatments.

Sessions took 70-90 minutes to complete. 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.[30] 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 REVENUE are in Appendix B.

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 price set by an unconstrained stable cartel, and δ^* the critical discount rate above which this price can be part of a subgame-perfect Nash equilibrium, derived in Section 2.

3.2 Hypotheses

Table 2 displays the theoretical predictions on prices and critical discount rates which are based on the model in Section 2 and the parameters introduced in Section 3.1. Parameters were selected based on simplifying the presentation towards subjects while ensuring that no focal prices emerged that could guide subject behavior. We test the following hypotheses against the null of no differences between treatments.

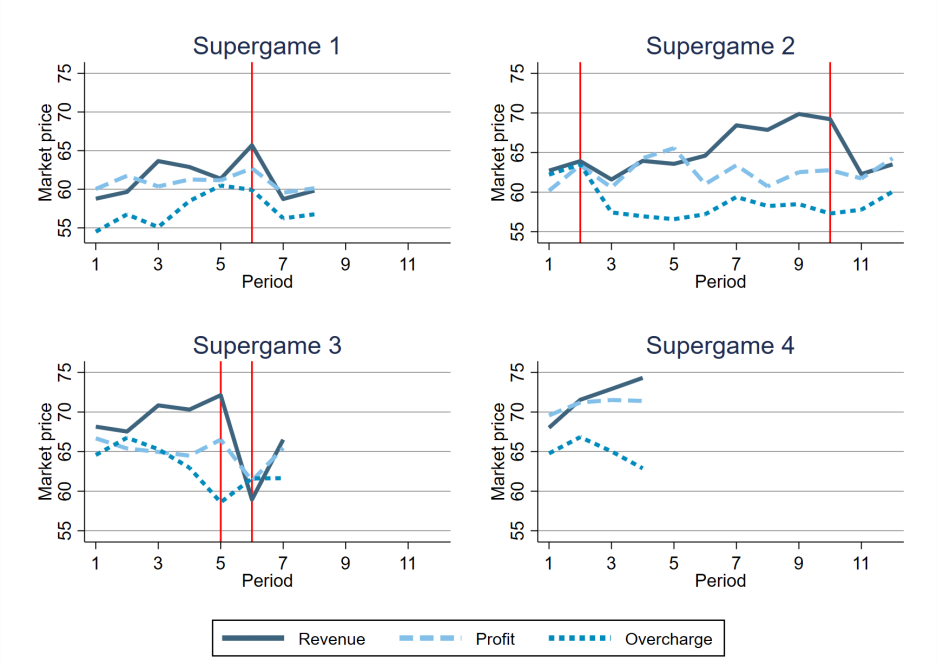
H1: Market prices are highest in REVENUE and lowest in OVERCHARGE
H2: Stable cartels are least likely in OVERCHARGE

[29] The share of sessions with only 9 subjects was equal across treatments.
[30] Out of 279 participants, this happened 11 times.

H1 originates in Corollary 1(i). 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 rate, which implies that stable cartels can, in theory, be the norm regardless of fining base.

Given the continuation probability of 0.9, our theoretical model provides no reason to reject the null hypothesis of no differences in cartel stability across treatments. However, we posit as an alternative hypothesis regarding cartelization that stable cartels are less likely to emerge in OVERCHARGE than in PROFIT and REVENUE. The experimental literature on infinitely repeated games generally finds that the discount rate exceeding the critical discount is a necessary, but not sufficient, condition for coordination. Indeed, the literature suggests that subjects are more likely to cooperate the further is the discount rate above the critical discount rate (Dal Bó and Fréchette, 2018). **H2** follows as this difference is smallest in OVERCHARGE.

Figure 2: 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.

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 cartelization.[31] We show that differences in prices across treatments are driven by differences in prices of cartels rather than differences in the prevalence of cartels or prices in uncartelized markets. In Sections 4.3 and 4.4, 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.[32]

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)

Notes: Table 3 compares prices across treatments; 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 deviations in brackets; ***, **, and * indicate statistical significance at the 1%, 5%, and 10% level, respectively.

[31] 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).

[32] We use Wilcoxon rank-sum tests for comparisons across treatments and the Wilcoxon signed-rank-sum test for within-treatment comparisons. All tests are two-sided, with the average of a variable within a re-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-period or subject-period level data – depending on the outcome – while clustering the standard errors at the re-matching-group level (although regression-based p-values are typically lower than those reported in the paper).

4.1 Prices

Figure 2 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 of 31 periods). With experience, subjects learn to set higher prices. In the first supergame, market prices are initially between 55 and 60; in the last supergame, this has increased to 65 to 70. While market prices tend upward over time, they typically decrease to similar levels in all treatments in the period immediately following cartel detection (the vertical red lines in Figure 2 indicate periods at the end of which detection occurs).

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 ($p = 0.000$ 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 cartelization, 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 $p = 0.000$, respectively). In line with the theoretical predictions, market prices exceeding the monopoly price are most common when fines are based on revenue.[33] Market prices in uncartelized markets lie between 51.21 in REVENUE and 48.82 in OVERCHARGE. This is only somewhat above the one-shot Nash equilibrium price of 47, suggesting that subjects do not manage to collude tacitly, and consistent with previous work on repeated Bertrand experiments with more than two players (e.g., Dufwenberg and Gneezy (2000); Fonseca and Normann (2012)). Results are even more in line with the prediction when subjects gain more experience – by the last two supergames – as then cartel prices in all treatments are significantly different while none of the differences in market prices of uncartelized markets are significant.

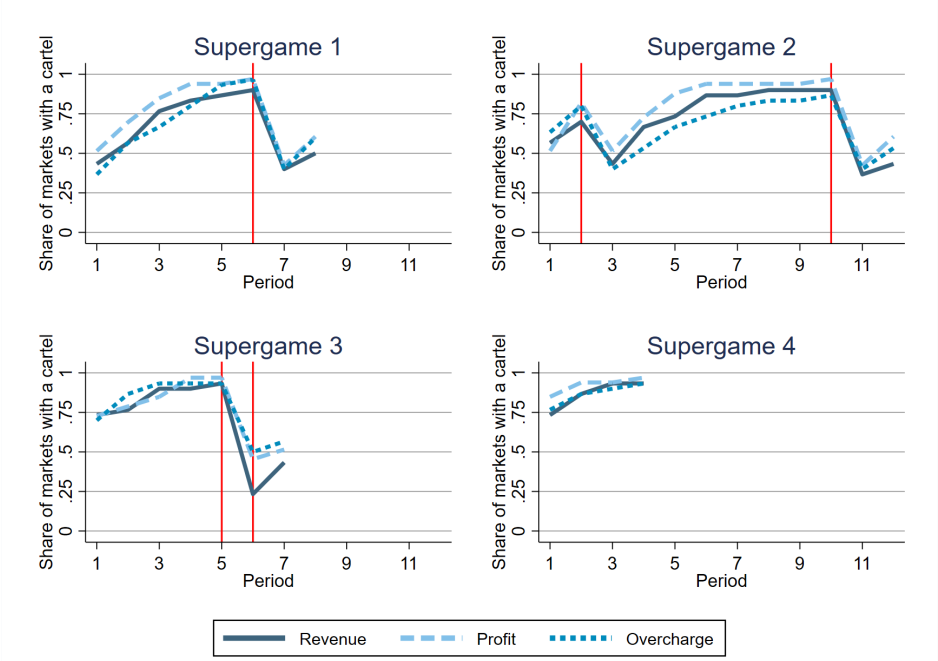
Summing up, we conclude that the data are in line with alternative hypothesis **H1**: market prices are highest in REVENUE and lowest in OVERCHARGE. While uncartelized markets yield prices close to the one-shot Nash equilibrium price in all three fining regimes, cartel prices are highest when fines are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

[33] 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 when fines are based on the overcharge ($p = 0.049$ and $p = 0.008$, respectively; $p = 0.564$ when comparing profit to overcharge bases).

4.2 Cartel formation, incidence, and recidivism

Figure 3 displays cartel incidence over time by treatment and supergame, and Table 4 presents the aggregate results on cartel formation, incidence, and 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.[34] 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. 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.

Figure 3: 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.

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

[34]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.

treatments. Indeed, cartel formation rates when fines are based on revenue (0.40), profit (0.50), or the price overcharge (0.43) do not differ significantly (p-values lie between 0.557 and 0.7394). Given these results, 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 ranges 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 C1 in Appendix C 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 appetite for risk form a cartel. We conclude that risk-preference-based selection into cartels does not differ across treatments.

Overall, none of our measures of cartelization significantly differ across treatments. We interpret this as aligning with the null hypothesis of no differences rather than the alternative hypothesis **H2**: stable cartels are equally likely in all treatments. Together with the fact that market prices in uncartelized markets do not differ across fining regimes either, this implies that all observed variation in market prices across treatments originates 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 agreements or stability cause differences in cartel prices across treatments.

4.3 Classifying cartel agreements

That subjects form cartels and send chat messages to each other does not necessarily imply that cartel members form agreements on which prices to set. We, therefore, classify the chat data according to whether an agreement is in place in a given period. We use two definitions of cartel 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 before any subject leaves the chat. As this classification misses many clear cases of

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)

Notes: Table 4 compares measures of cartelization across treatments; 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 deviations in brackets; ***, **, and * indicate statistical significance at the 1%, 5%, and 10% level, respectively.

price coordination, we also employ a broader definition of cartel agreements – titled 'price agreements' – that adds implicit agreements where the context makes it clear that all subjects agree on a particular price. An example of an implicit agreement is one subject commenting "great! let's keep it going" after several periods of successful coordination, followed by "ok" from the other two subjects. A detailed explanation of how chat data were classified, as well as several illustrative examples of chat contents, are given in Appendix D.[35]

Explicit price agreements are present in 1089 of all 2122 market-periods with a cartel (51.32 percent). When an explicit agreement is in place, subjects manage to successfully coordinate on a market price above the one-shot Nash equilibrium market price of 47 in 901 market-periods (82.74 percent of all explicit price agreements), and the average market price is 70.75, suggesting that our measure of explicit price agreements successfully captures cartel agreements. However, cartels without explicit price agreements still manage to coordinate on market prices above 47 in 49.95 percent of all market-periods (516 of 1033), compared to only 10.78 percent in uncartelized market-periods (82 out of 761). Unsurprisingly, therefore, the average market price in cartelized markets without explicit price agreements (64.68) substantially exceeds that of uncartelized markets (50.29). Moreover, Figure E1 in Appendix

[35]Recall that the content of the chat in our experiment does not determine the illegality of the cartel. Once subjects have voted to form a cartel, the chat window opens, and from that period onward, the subjects can be detected and fined until the cartel is detected or the supergame ends. In line with reality, discussing prices itself is punishable (and detectable in later periods).

D shows that the fraction of explicit price agreements tends downward within supergames, while the incidence of successful price coordination does not. These findings suggest that our measure of explicit price agreements substantially underestimates the frequency of cartel agreements, prompting us to construct a broader measure: 'price agreements.'

Price agreements are present in 80.21 percent of all market-periods with a cartel (1702 of 2122 instances). The average market price of cartelized markets with price agreements is 70.92, while market prices of cartelized markets without price agreements (55.13) are now much closer to those of uncartelized markets (50.29). Figure E1 in Appendix D shows that the incidence of price agreements tracks the movement of successful coordination over time, while the level is higher, which suggests that our broader measure of cartel agreements is substantially more accurate than explicit price agreements alone. Hence, in the remainder of the analysis, we present results using our broad measure of price agreements. Table F1 in Appendix F shows that price agreements, rather than cartel stability, still explain our aggregate results if we use explicit price agreements instead.

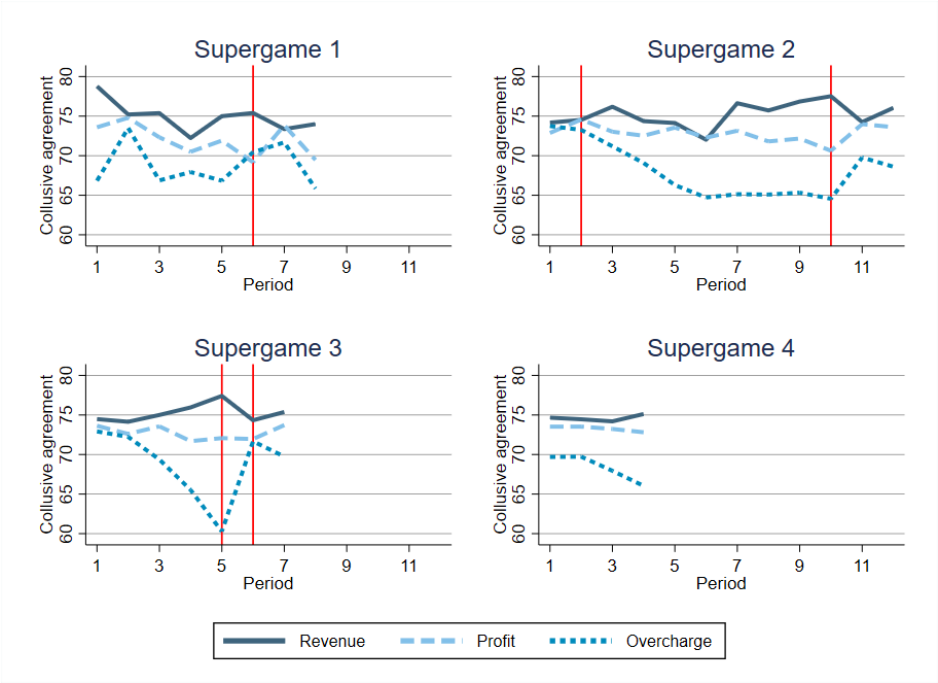
Figure E2 in Appendix D reveals no apparent differences in price agreement incidence between treatments throughout the four supergames. Moreover, the shares of cartelized market-periods with price agreements or in REVENUE (0.81), PROFIT (0.77), and OVERCHARGE (0.83) do not differ significantly, as is the case for the average risk aversion in market-periods with price agreements.[36] 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 when fines are based on revenue (61.13) are higher than when fines are based on profit (52.34) or the overcharge (53.06) ($p = 0.049$ and $p = 0.001$, respectively). Hence, one explanation underpinning our aggregate results, albeit minor, given the prominence of price agreements, is that basing fines on revenue leads to higher market prices even when cartels fail to reach an agreement on which price to set. Recall that, according to theory, when fines are based on revenue, firms that are still detectable but no longer coordinate prices with other firms set price $p_R^{PD} = \frac{c}{1-\alpha r^R} = 58.75$. In line with theory, therefore, even unsuccessful cartels increase market prices when fines are based on revenue.[37]

[36] Comparisons of price agreement incidence result in p-values between 0.512 and 0.912, 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.

[37] This is unlikely to be the result of our cartel agreement categorization being too conservative in REVENUE, as in this treatment cartels only coordinate on market prices above 47 in 10 out of 124 market-periods (8.06 percent), which is less than the incidence of such coordination in the absence of cartels (10.78 percent). This result is also unlikely to be driven by learning effects as it is robust to only using data from the final two supergames.

Figure 4: 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.

4.4 Price agreements and cartel stability

Figure 4 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 (75.03) are significantly higher than those in PROFIT (72.67) and OVERCHARGE (68.92) ($p = 0.010$ and $p = 0.000$, respectively). 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.

Price agreements above the monopoly price of 73.5 are common in REVENUE. In 326 of all 535 market-periods with a price agreement subjects agree to set such a price (60.93 percent). While price agreements and the fraction of above-monopoly-price agreements are stable over time, the standard deviation of price agreements in REVENUE decreases from 6.85 in the first two supergames to 3.60 in the final two supergames as agreements of different cartels converge. Moreover, 9.91 percent of all price agreements fall between 79 and 80, a percentage that is stable over the supergames. Subjects appear to converge on a price between the monopoly and predicted cartel prices. Therefore, the perverse incentives inherent in revenue-based fines push price agreements above the monopoly price, but to a lesser extent than predicted by theory.

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	0.81 (0.39)	75.03 (5.70)	0.78 (0.41)	73.91 (6.58)
	$\vee$	$\vee^{**}$	$\wedge$	$\vee^{**}$
PROFIT	0.77 (0.42)	72.67 (4.70)	0.81 (0.39)	71.81 (5.71)
	$\wedge$	$\vee^{***}$	$\wedge$	$\vee^{***}$
OVERCHARGE	0.83 (0.38)	68.29 (9.43)	0.85 (0.36)	67.03 (9.77)
	$\vee$	$\wedge^{***}$	$\vee$	$\wedge^{***}$
REVENUE	0.81 (0.39)	75.03 (5.70)	0.78 (0.41)	73.91 (6.58)

Notes: Table 5 compares measures based on price agreements across treatments; 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 deviations in brackets; ***, **, and * indicate statistical significance at the 1%, 5%, and 10% level, respectively.

Price agreements converge to the jointly optimal monopoly price in PROFIT. From Figure 4, 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, price agreements between 73 and 74 characterize 233 of 345 market-periods with an agreement (67.54 percent). This percentage increases to 94.46 in the final two supergames (256 of 271 instances), and the standard deviation of price agreements decreases from 5.52 to 3.35, indicating near-complete convergence to the monopoly price.

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 that are adhered to is high in all treatments, ranging from 0.78 in REVENUE to 0.85 in OVERCHARGE, and does not differ significantly across treatments (p-values lie between 0.315 and 0.796). Over time, cartel stability increases, up to 0.88 across all treatments in the last two supergames. As a result, market prices in market-periods with a price agreement in place in REVENUE (73.91) are significantly higher than those in PROFIT (71.81) and OVERCHARGE (67.03) ($p = 0.010$ and $p = 0.001$, respectively), and market prices in PROFIT exceed those in OVERCHARGE ($p = 0.001$). Therefore, our aggregate results on market prices are primarily driven by the prices that stable cartels agree to set.

5 Concluding remarks

In this paper, we have investigated the relative performance of three bases for cartel fines: overcharge, profit, and revenue. We have done so using a theoretical model and a laboratory experiment. While we observe no significant differences across treatments regarding cartel formation, incidence, or recidivism, we find that average prices are lowest when fines are based on the overcharge and highest when they are based on revenue.

Policymakers may wonder whether our experimental results are generalizable to practice. In general, economic laboratory experiments frequently, albeit not always, replicate in the field (Camerer, 2015) or with professional participants rather than students (Fréchette, 2015). Of course, policymakers should keep in mind that lab experiments have their limitations, like every method (Falk and Heckman, 2009). Following the current consensus in the experimental-economics literature (Schram (2005); List (2020)), our experimental results do not provide reasons why policymakers should be hesitant to implement an overcharge regime in anti-cartel enforcement. Of course, there may be practical hurdles, including the data requirements that are arguably more demanding than for at least the revenue base. Policymakers should remember that such data need to be made available for private damage cases in any case, giving them an incentive to team up with customers that the cartel harmed. Moreover, the competition authorities should make firms aware of a regime switch.[38]

[38]Early empirical evidence by Block et al. (1981) indeed suggests that the visibility of anti-cartel enforcement has a downward pressure on markups.

References

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.

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.

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, 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. (2018). On the determinants of cooperation in infinitely repeated games: A survey. *Journal of Economic Literature*, 56(1):60–114.

DLA Piper (2020). Cartel enforcement global review 2020.

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.

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.

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.

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., and Treuren, L. (2023a). Spillovers from legal cooperation to tacit collusion. *KU Leuven Discussion Paper Series DPS 23.12*.

Hinloopen, J., Onderstal, S., and Soetevent, A. (2023b). 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 annual conference.

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. 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. (no. w27535) National Bureau of Economic Research.

Marshall, R. C. and Marx, L. M. (2007). Bidder collusion. *Journal of Economic Theory*, 133(1):374–402.

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.

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.

Robinson, M. S. (1985). Collusion and the choice of auction. *RAND Journal of Economics*, 16(1):141–145.

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.

Schram, A. (2005). Artificiality: The tension between internal and external validity in economic experiments. *Journal of Economic Methodology*, 12(2):225–237.

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.

United States Sentencing Commission (2021). Guidelines manual 2021.

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

Appendices

A Proofs of propositions

Proof of Proposition 1

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

$$p_R^C = \arg\max_p (1-\alpha r^R) q(p) \left(p - \frac{c}{1-\alpha r^R}\right) = p^M\left(\frac{c}{1-\alpha r^R}\right). \tag{4}$$

That is, the cartel acts like a monopolist in the absence of antitrust, facing marginal cost $\frac{c}{1-\alpha r^R}$. As the monopoly price increases in c, $p_R^C > p^M$. Denote the corresponding firm-level profit and fine by π_R^C and F_R^C.

The optimal defection is not the monopoly price p^M, as defectors can be detected and fined. 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 the defector's before-fine profit and fine 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, firms do not revert to the Nash equilibrium of the static Bertrand game as the possibility of being fined implies that expected profit is negative if all firms set price equal to marginal cost. Instead, the unique pure-strategy Nash equilibrium is $p_R^{PD} = \frac{c}{1-\alpha r^R}$. 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}: $\pi_R^{PD} - \alpha F_R^{PD} = 0$. Together with $\pi^N = 0$ and the preceding paragraph, this implies that $V^D = n\left(\pi_R^C - \alpha F_R^C\right)$. The stability condition can, therefore, be written as

$$\frac{\pi_R^C - \alpha F_R^C}{1-\delta} \geq n\left(\pi_R^C - \alpha F_R^C\right) \iff \delta \geq \delta_R^* \equiv \frac{n-1}{n}. \tag{5}$$

■

Proof of Proposition 2

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

$$p_\pi^C = \arg\max_p (1-\alpha r^\pi)\,(p-c)\,q(p) = p^M(c). \tag{6}$$

As the profit-based fine acts as a tax on profit, it does not affect the profit-maximizing price, and the cartel sets the monopoly price. Denote the corresponding firm-level profit and fine by π_π^C and F_π^C.

As an incremental 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 $V^D = n\left(\pi_\pi^C - \alpha F_\pi^C\right)$. The stability condition can, therefore, be written as

$$\frac{\pi_\pi^C - \alpha F_\pi^C}{1-\delta} \geq n\left(\pi_\pi^C - \alpha F_\pi^C\right) \iff \delta \geq \delta_\pi^* \equiv \frac{n-1}{n}. \tag{7}$$

■

Proof of Proposition 3

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. The unconstrained cartel price is then given by

$$p_O^U = \arg\max_p (p-c)\,q(p) - \alpha r^O\left(p - p^N\right)q(p^N) < p^M(c). \tag{8}$$

The inequality follows from the first-order condition of the maximization problem underlying equation (8): $(p-c)\frac{\partial q(p)}{\partial p} + q(p) - \alpha r^O q(p^N) = 0$. 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 to reduce the expected fine and increase expected profit. The difference between p^M and p_O^U increases with the detection probability, the penalty rate, and the Nash-equilibrium quantity of the static Bertrand game. All of these factors are unrelated to the behavior of the cartel.

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

$$\max_q (p(q)-c)(q - \alpha r^O q(p^N)). \tag{9}$$

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 $p_O^U > c$, and also that $p(\alpha r^O q(p^N)) > p_O^U$, which is used below.

In an overcharge regime, the downward pressure on the price is smaller by a factor n for a defector than for the entire cartel as the defector's fine is scaled by $\frac{q(p^N)}{n}$ instead of $q(p^N)$. This is true regardless of whether we consider a constrained or unconstrained cartel price. Consider the first-order condition of a cartel that faces the defector's fine: $(p-c)\frac{\partial q(p)}{\partial p} + q(p) - \alpha r^O \frac{q(p^N)}{n} = 0$. 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 intact, 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 not been convicted yet, 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 $V^D = n\pi_O^C - \alpha F_O^C$. The stability condition can, therefore, be written as

$$\frac{\pi_O^C - \alpha F_O^C}{1-\delta} \geq n\pi_O^C - \alpha F_O^C \iff \delta \geq \delta_O^* \equiv \frac{(n-1)\pi_O^C}{n\pi_O^C - \alpha F_O^C}. \tag{10}$$

The stability condition is associated with a maximum price that potentially constrains the cartel price to lie below p_O^U. We can see this by rewriting the stability condition.

$$\frac{(p-c)(q(p) - \alpha r^O q(p^N))}{1-\delta} \geq (p-c)(nq(p) - \alpha r^O q(p^N)). \tag{11}$$

If $p = c$, the stability condition is always satisfied, so we restrict attention to $p - c > 0$, divide both sides by $p - c$, and solve for $q(p)$.

$$q(p) \geq \frac{\delta \alpha r^O q(p^N)}{1-(1-\delta)n} \iff p \leq p_O^{max} \equiv p\left(\frac{\delta \alpha r^O q(p^N)}{1-(1-\delta)n}\right). \tag{12}$$

As inverse demand strictly decreases with quantity, p_O^{max} increases in δ. At $\delta = \frac{n-1}{n-\alpha r^O}$, $p_O^{max} = c$, and p_O^{max} increases until is equals $p(\alpha r^O q(p^N))$ at $\delta = 1$. At $\delta = \delta_O^*$, $p_O^C = p_O^{max}$. Stable cartels that set a price above marginal cost are, therefore, possible for $\bar{\delta} \equiv \frac{n-1}{n-\alpha r^O} < \delta < \delta_O^*$, but the stability condition restricts the cartel price. We can now characterize the price of stable cartels.

$$p_O^C = \begin{cases} p_O^U & \text{if } \delta_O^* \leq \delta < 1, \\ p_O^{max} & \text{if } \bar{\delta} \leq \delta \leq \delta_O^*. \end{cases} \tag{13}$$

■

B 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. For brevity, we include only the instructions for REVENUE and for the risk preference test – discussed in Appendix C. The instructions for the other treatments are available from the authors upon request.

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 perioded to two decimal points.

The market

The price you set must be between 0.01 and 99.99 (all inputs are perioded 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. Recall that only if all three vote in favor, a cartel is formed. 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 about prices, you will still need to set a price independently.

Chatting about anything that can be used to 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, the cartel will automatically be active in the next period. If your cartel is detected, you will pay a fine and the cartel is no longer active. If your cartel is detected, you can vote to form a new cartel in the next period.

If your cartel is detected and you set price p in that period, your fine will be:

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

$Fine = p\frac{100-p}{n}$ if n firms set the lowest price.

Note that 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 and get:

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

The cartel is not detected this period, so total profit is 132 this period for all three firms. 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. 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 firm 3 $= (50 - 47)\frac{100-50}{1} = 150.$

The cartel is detected this period. Firms 1 and 2 pay no fine as they did not set the lowest price. Their profit for this period is 0.

Firm 3 pays the following fine:

$Fine = 50\frac{100-50}{1} = 2500.$

Firm 3's profit for this period is 150 - 2500 = -2350.

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

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.

C 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 C1 displays the lotteries as seen by the participants, and Appendix B 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.

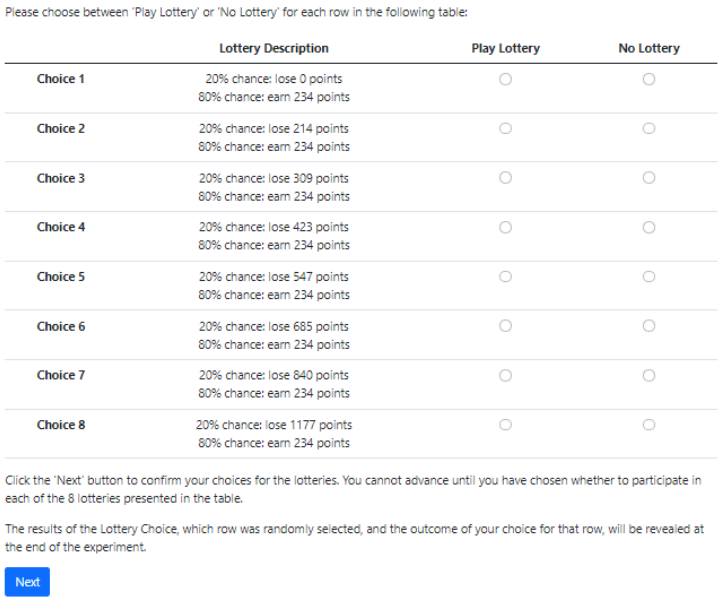
We construct as a measure of risk preferences the first row that a subject opts to not play the lottery. This measure ranges from 1 (subject does not play the lottery in row 1) to 9 (subject plays all eight lotteries), with higher values indicating a higher appetite for risk. Table C1 describes this 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 C1: Risk preferences, by treatment

	REVENUE	PROFIT	OVERCHARGE
25th percentile	4	4	4
50th percentile	5	5	5
75th percentile	6	6	6
Mean	5.02	4.99	5.19
Standard deviation	1.85	1.96	1.80
Observations	90	99	90

Notes: Descriptive statistics on our measure of risk preferences: the first lottery that a subject opted not to play.

Figure C1: Risk preference test



D 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 utilize two definitions of cartel agreements in our analysis. Explicit price agreements are classified purely based on the content of the chat. As discussed in Section 4.3, this definition seems too conservative as it misclassifies both the level and the trend of cartel agreements. Therefore, we construct a broader measure of price agreements that are 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. We next provide a description of how we construct both measures, followed by chat excerpts that provide examples of implicit agreements or are referenced in the main text.

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.

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 none-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 on 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 can not 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 D1 and Table D2 give examples where lack of communication was classified as a price agreement and not as a price agreement, respectively. Our results are robust to tightening the stability condition by only classifying periods without communication as price agreements if the cartel had been stable in all prior periods, but omit these results for the sake of brevity.

Table D1: Lack of communication classified as price agreement

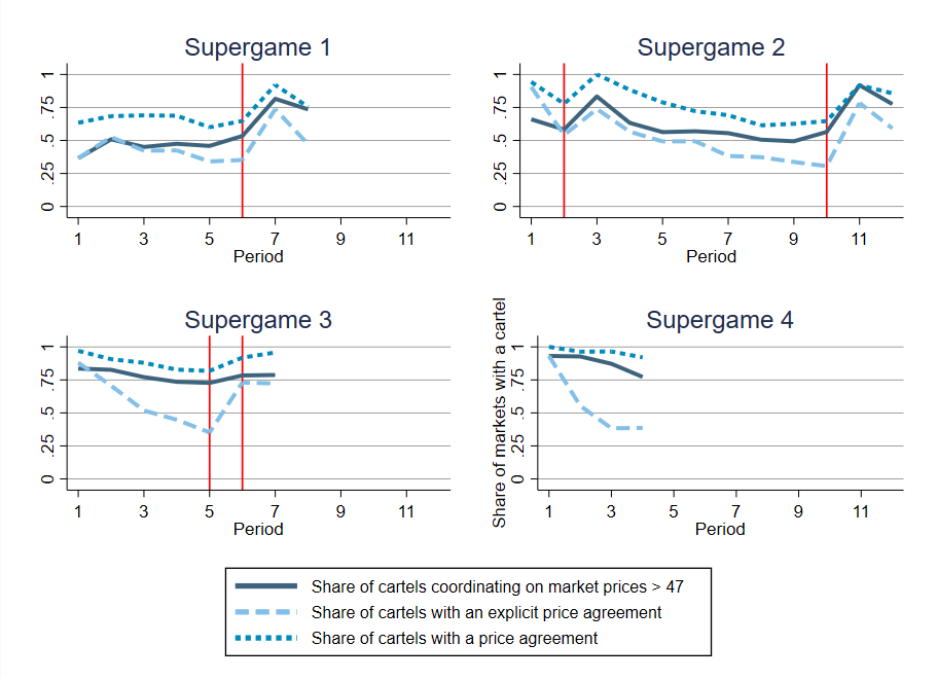
Supergame - Period	Firm 1	Firm 2	Firm 3	Category
2-1		75?		
2-1	85?			
2-1			The max profit is at 73.5	
2-1		73.5 it is then		
2-1			Shall we do that?	
2-1	okay			Explicit price agreement
2-1	*Firm 1 exits chat*			
2-1			Cool	
2-1		*Firm 2 exits chat*		
2-1			*Firm 3 exits chat*	
2-1	price: *73.5*	price: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

Table D2: Lack of communication not classified as price agreement

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

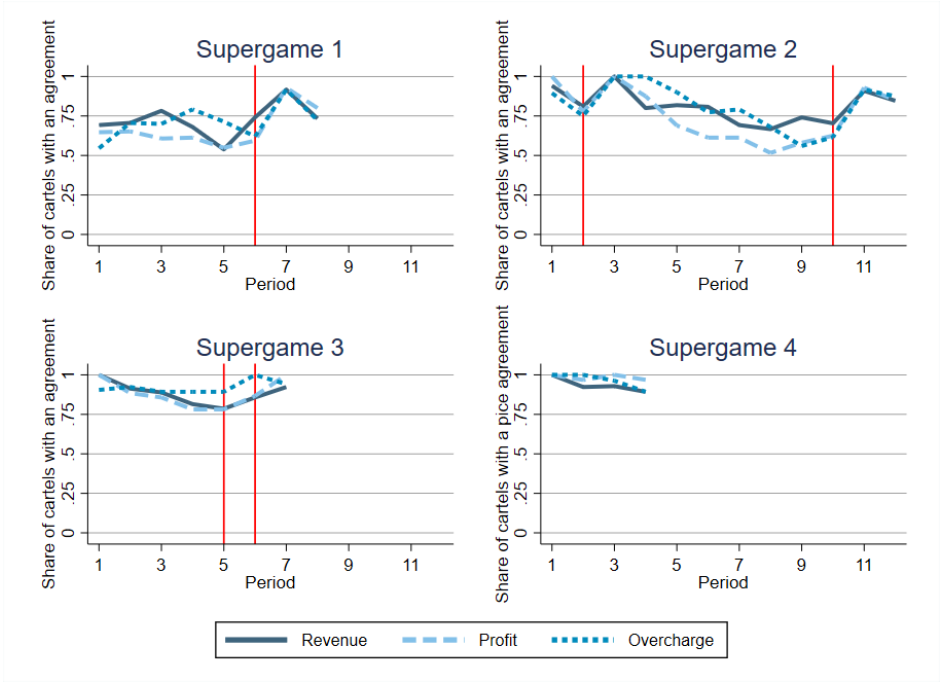
E Additional Figures

Figure E1: Coordination and agreement incidence over time, by supergame



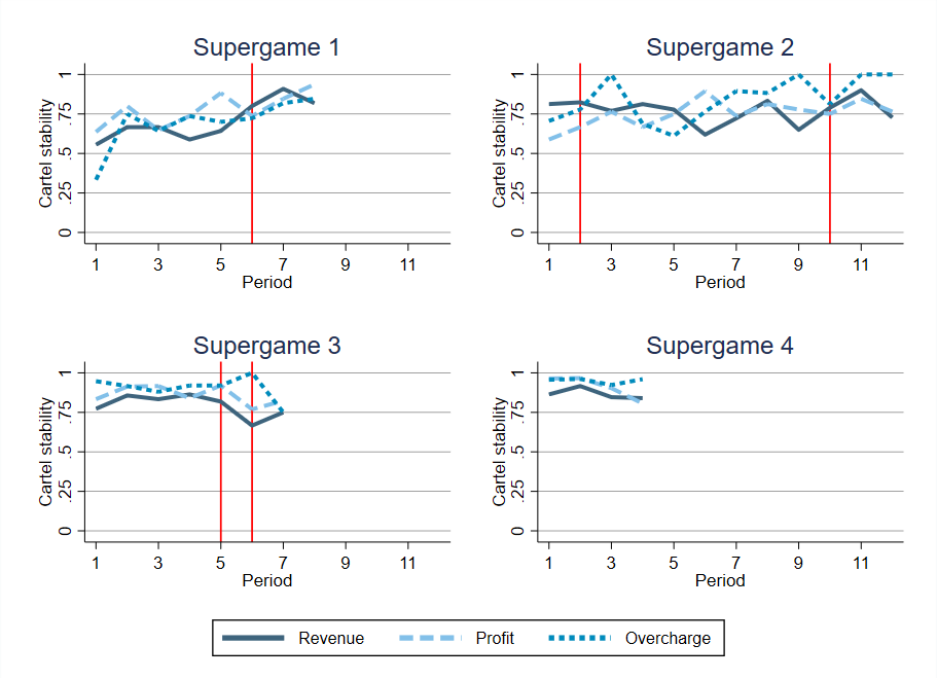
Notes: Share of cartels that coordinate on a market price above the one-shot Nash-equilibrium price of 47, and share of cartels with an (explicit) price agreement in place, over time and by supergame. Red vertical lines indicate a period at the end of which all cartels are detected.

Figure E2: 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 in a cartelized market-period. Red vertical lines indicate a period at the end of which all cartels are detected.

Figure E3: 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.

F Additional Tables

Table F1: Price agreements and cartel stability using explicit agreements, across treatments

	Agreement incidence	Price agreement	Cartel Stability	Market price
	(Cartels)	(Cartels with an explicit price agreement in place)		
REVENUE	0.46 (0.50)	74.86 (5.64)	0.79 (0.41)	73.77 (6.32)
	$\wedge$	$\vee^{***}$	$\wedge$	$\vee^{**}$
PROFIT	0.50 (0.50)	72.56 (5.10)	0.82 (0.38)	71.82 (6.02)
	$\wedge$	$\vee^{***}$	$\wedge$	$\vee^{***}$
OVERCHARGE	0.58 (0.49)	68.35 (9.48)	0.86 (0.34)	67.33 (9.85)
	$\vee^{**}$	$\wedge^{***}$	$\vee^{*}$	$\wedge^{**}$
REVENUE	0.46 (0.50)	74.86 (5.64)	0.79 (0.41)	73.77 (6.32)

Notes: Table F1 compares measures based on explicit price agreements across treatments; Agreement incidence = Indicator for a cartel with 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 deviations in brackets; ***, **, and * indicate statistical significance at the 1%, 5%, and 10% level, respectively.