TESTING MODELS OF STRATEGIC UNCERTAINTY: EQUILIBRIUM SELECTION IN REPEATED GAMES

MARTA BOCZOŃ, EMANUEL VESPA, TAYLOR WEIDMAN, AND ALISTAIR J. WILSON

Abstract. In repeated-game applications, where both the collusive and noncollusive outcomes can be supported as equilibria, it is crucial for researchers to understand when each equilibria is more likely to result. Controlled experiments have provided a selection criterion for the two-player repeated prisoner's dilemma. The theoretical selection criterion uses game primitives to measure the set of beliefs such that an agent prefers to defect in all rounds relative to cooperating conditionally. The set of such beliefs is large when uncertainty with respect to what others will do–strategic uncertainty–is large, and the prediction is that a noncooperative outcome would emerge. In this paper, we experimentally test this model of selection and its underlying mechanism of strategic uncertainty by manipulating the total number of players. Our results affirm the model as a tool for predicting when tacit collusion is likely/unlikely to be successful. Extending the analysis, we corroborate the findings and mechanism of the model, in alternative settings, and for the decisions of non-human AI agents.

1. Introduction

Answering questions on which of many possible equilibria best capture economic behavior is of central importance for applications with repeated interaction. For example, in models of oligopoly in which firms interact repeatedly, both collusive and noncollusive equilibria can arise. To better guide assumptions over equilibrium selection, and hence policy conclusions, experimental work has uncovered basic theoretical criteria that predict the likelihood of collusion based on game's payoffs and discount rates. Such work has focused on the two-player indefinitely repeated prisoner's dilemma (RPD). However, it is unknown to what extent the uncovered predictive criteria can be used to predict collusion in applications with features beyond this stark setting.

Testing whether equilibrium selection criteria hold beyond the two-player RPD presents an important challenge: first and foremost, the selection criteria have to be adapted to a new setting. For concreteness, let us introduce the *basin of attraction for always defect*, which has been shown to organize experimental data (Dal Bó and Fréchette, 2018). The measure takes stage-game payoffs and the discount factor as the input, and outputs the minimum belief that the other player will cooperate such that it is worthwhile to collude. The higher the minimum belief needed for cooperation, the wider the set of beliefs that lead to defection, hence the name.[1] A large set of beliefs leading to defection captures

Date: March, 2023.

We would like to thank: David Cooper, Guillaume Fréchette, Daniella Puzzello, and Lise Vesterlund. This research was funded with support from the National Science Foundation (SES:1629193).

[1]The basin measure, which we describe in detail in Section 2, makes the stark assumption that agents use either the grim-trigger (cooperative) strategy or the always-defect (noncooperative) strategy. It essentially

high risk of cooperation, which is why this measure is sometimes thought of as a proxy of strategic uncertainty (i.e. uncertainty with respect to whether the other player will cooperate). Experimental data starting with Dal Bó and Fréchette (2011) show that when the theoretical minimum belief in cooperation is high (low), observed cooperation rates tend to be low (high).

Now, consider a simple but practically relevant extension to an environment with more than two players, the focus of this paper. To extend the basin-of-attraction measure to an environment with N players the agent needs to assess the chances that *more than one* other player will cooperate. A natural benchmark is to compute the minimum belief that others cooperate by treating other agents as symmetric but independent of each other. We formally describe such an extension in the next section, referring to it as the *independent extension*. However, an experimental test of this extension in isolation may not be informative. Specifically, if the measure were not to explain the data well, the failure could be attributed either to an intrinsic problem with the notion (for example, that strategic uncertainty through the number of players plays no role) or to participants not treating others' behavior as independent. Hence, as an alternative hypothesis anchored at the other possible extreme, we develop the *correlated extension* in which the beliefs over others are perfectly correlated. If the beliefs over others are perfectly correlated, a change in N has a null effect on behavior. In contrast, if beliefs are independent, increasing N makes it more difficult to coordinate on a collusive outcome. Finally, it is possible that both extensions have low predictive power, which could suggest that the issue is entirely unrelated to strategic uncertainty and beliefs over others, or that the underlying correlation in beliefs lies somewhere in between the two extremes.

An experimental design that manipulates N in isolation would not provide strong evidence to support either extension. The comparative-static prediction of N is asymmetric—with no change predicted under perfect correlation and a large directional change expected under full independence. Given experimental evidence on the effects of group size in other settings (which we review below), a purely directional test over N is likely to favor the independent extension. To provide a more stringent evaluation, in addition to N, we introduce a second treatment variable. This second parameter, x, affects the stage-game payoffs, increasing the temptation to defect, while also creating a directional prediction under the correlated extension. In particular, our treatment parameterizations are designed so that any increase in strategic uncertainty in the independent extension when N increases (from N_0 to N_1), can be directly compensated for with a shift in x (from x_0 to x_1). This allows us to develop a 2×2 design directly over the two basin extensions that forms a stronger test: holding constant the correlated (independent) extension we can shift the independent (correlated) extension.

Our main result is that the independent basin extension best organizes longer-run behavior across treatments. Not only do we observe large shifts in the predicted direction when varying N in isolation, but we also find substantial similarity in the longer-run behavior when keeping constant the independent-extension prediction by varying x and N

consists of an extension of the notion of equilibrium selection described in Harsanyi and Selten (1988) and was first proposed by Blonski and Spagnolo (2015).

in opposing directions (here from $\{N_0, x_0\}$ to $\{N_1, x_1\}$). However, the independent extension does not predict all measures of cooperation equally well. Although it succeeds in measuring ongoing cooperation, which may be the most relevant for many applications, there are distortions in the quantitative predictions of initial cooperation that captures intentions to collude.

Strategic uncertainty regarding the other players' actions is put front and center in our main treatments as the driver of equilibrium selection. Hence, if strategic uncertainty is the causal mechanism, removing or reducing doubts about others' play should make the model's predictions moot. We pursue this idea in an additional set of treatments in which we allow for pre-play communication but where the chosen parameterization makes it difficult to collude. Participants are given the opportunity to exchange free-form messages before the game, a feature that can be used to reduce the uncertainty about the intentions of others (see Kartal and Müller, 2018). For the same experimental parameterization without pre-play communication, we observe ongoing cooperation rates below one percent. On the contrary, the effect of adding communication shifts behavior to the other extreme: with initial (ongoing) cooperation rates of 95 (80) percent.[2] Given these results, we conclude that strategic uncertainty is the causal channel. In addition to mechanism validation, our results here clearly outline the limitation of the basin of attraction as a model for understanding *tacit* collusion: it fails to provide useful guidance when collusion is more explicit.

While we find that the basin-of-attraction model makes poor predictions once strategic uncertainty is removed via pre-play communication, it predicts well across a number of extensions in which strategic uncertainty is not dissipated. For a second extension, we test whether an equilibrium is sticky by examining a situation in which the primitives of the game change within a session. Specifically, we introduce a change in N (from $N_0 = 4$ to $N_1 = 2$, and vice versa) halfway through an experimental session. If selected equilibria are sticky—because so are the beliefs over others—then a change in N will not affect behavior, and the selection model becomes moot. However, if the strategic uncertainty is reset with a change in the environment, an increase (decrease) in N will decrease (increase) the uncertainty over others resulting in a change in behavior. Our findings indicate that there is no stickiness in the long run. Specifically, behavior adjusts after a change in N, moving with experience toward the levels of cooperation observed in sessions with fixed N. In summary, this suggests that changes in strategic uncertainty lead to changes in behavior in a predictable way, with little evidence of stickiness.

The third extension weakens the requirement for successful cooperation. Whereas our main treatments require joint cooperation by all N players for a stable cooperative outcome, here, we weaken this requirement so that even if only half of the players in our

[2] As an additional check that the provided communication channel is not driving our results separate from equilibrium coordination (for example, increasing other regarding concerns) we implement a second treatment with communication in which the collusive outcome is not an equilibrium. Here, we find that communication does not lead to successful collusion. As such, communication has an effect only when there is a clear motive for selection of the equilibrium.

$N = 4$ treatment cooperate, a success will ensue. Despite the weakened condition for success (only two cooperators, versus all four in our original treatments) the strategic uncertainty predicts lower cooperation, as coordination over which two players cooperate becomes harder. Even though the structure for successful cooperation is different, the evidence is again broadly consistent with the theoretical prediction.

Taking a step back, the main goal of our paper is to help construct an empirical criteria for equilibrium selection in repeated games. In games with a large set of equilibria, such a tool can be very useful to evaluate policy recommendations and to use such models for predicting behavior. However, a clear shortcoming of any experimental paper such as ours is that conclusions are specific to the chosen environment and parameterizations. Ideally, one would want to evaluate the criterion for equilibrium selection in a large set of repeated games, and in each set for several possible parameterizations. While this goal is outside the scope of the paper, we make two final contributions that may help further evaluate empirical criteria of equilibrium selection.

First, we introduce a methodology that allows for a more expansive exploration without running an infeasible number of experimental conditions. We show that the experimental results for both the previous RPD literature and our main environments with $N > 2$ can be replicated with artificial intelligence algorithms (AIAs) that companies use for pricing decisions (Calvano, Calzolari, Denicolo, and Pastorello, 2020; Asker, Fershtman, and Pakes, 2021). Though still beyond the scope of this paper for a full exploration, we outline how such AIA driven experimental conditions can be used in future research both to explore the limitations of theoretical selection devices like the basin, but also as an aide in designing treatments for future studies. Given that we find a qualitative and a quantitative match between the long-run behavior of AIAs and our lab participants, the former can be used to predict behavior of human subjects in counterfactual environments that are not directly studied in the laboratory. Although not as analytically tractable as our basin calculation—which provides closed-form solutions for the direction of an effect and any interactions—such AIAs can be used to expand the scope of experimental studies if partially validated on the narrower domains studied within the laboratory. Finally, while we believe that studying the extension of analytic selection criteria such as the basin to $N > 2$ is of natural importance, we also offer a proof-of-concept method for validating and proving extensions to other domains.

1.1. **Literature.** This paper is connected to several strands of the literature. Our design is based on the recent consolidation of the experimental RPD literature presented in Dal Bó and Fréchette (2018). While one of our baseline treatments replicates a standard finding in the literature,[3] our core contribution is to generalize the equilibrium selection model outlined in the Dal Bó and Fréchette (2018) meta-study, adding an additional source of strategic uncertainty: the number of players, N.[4] Whereas the literature has developed

[3]As pointed out in Berry, Coffman, Hanley, Gihleb, and Wilson (2017) experimental replications are rare as the papers often leave out details of their experimental design.

[4]The extension of the notion of equilibrium selection described in Harsanyi and Selten (1988) to the RPD was first proposed by Blonski and Spagnolo (2015) (with further details in Blonski, Ockenfels, and Spagnolo (2011)) and was first shown to organize data by Dal Bó and Fréchette (2011). See also Fudenberg, Rand,

this model for explanatory purposes, our approach is both to expand the model to a new setting, but also to test it as the core experimental object.

Our generalization of the strategic uncertainty model is carried out in two ways. The first extension (and most standard, given its use of independent beliefs) formalizes a distinct source of strategic uncertainty from the payoff-based source in the meta-study. An alternative extension (based on fully correlated beliefs) reflects a null effect, that the newly introduced source has no effect. As such, our generalization offers a potentially profitable design approach for future research examining other channels for strategic uncertainty effects—asymmetries in the action space or payoffs, the effects of sequentiality, etc.[5]

Our environment also allows us to better distinguish between the empirical measures linked to the selection model. That is, using literature-level data assembled by Dal Bó and Fréchette (2018), we show that the two-player RPD strategic uncertainty model is suitable to predict both initial and ongoing cooperation.[6] However, with more than two players, this is no longer the case. Here, we demonstrate that the strategic uncertainty model is better suited to predict successful ongoing collusion rather than initial intentions to collude.[7]

This paper is part of a broader literature that seeks to understand and document regularities in equilibrium selection, in particular, regularities that are amenable to theoretical modeling. To this end, the strategic-uncertainty measure that we examine is particularly promising, as the equilibrium objects required for calculation are computationally simple: the stationary noncollusive equilibrium and the history-dependent collusive equilibrium. In environments beyond the RPD in which the equilibrium outcomes are held constant, the model can be similarly extended per our illustration with a move to N players. However, in more complex environments with changing sets of equilibria, the constraint to two focal equilibria in the strategic uncertainty model may lose validity and/or raise questions as to which two strategies are focal. Examples of more-complex settings include dynamic games in which the stage environment changes across the supergame, and the space of strategies becomes substantially larger. Vespa and Wilson (2020) focus on a horse-race examination of which two equilibria are focal (from a wider set of possible alternatives) to rationalize behavior in dynamic games. In this paper, we identify a similar strategic uncertainty measure constructed around the most-efficient Markov perfect equilibrium and the best *symmetric* collusive equilibrium. A strategic-uncertainty model

and Dreber (2010) for an examination of the effects with imperfect monitoring and Kartal and Müller (2018) for a test of a selection theory based on individual heterogeneity in preferences over dynamic strategies.

[5]See Ghidoni and Suetens (2022) and Kartal and Müller (2018) for experimental examinations of the effect of sequentiality in RPD settings through a reduction in strategic uncertainty.

[6]With two players, but when sequential moves are allowed, there is additional variation for identification. Ghidoni and Suetens (2022) also find that ongoing measures are better predicted than initial rates.

[7]Ongoing cooperation is a measure that is likely to be more relevant for empirical applications where collusion may be a worry. For instance, from Harrington, Gonzalez, and Kujal (2016), page 256: "(...) collusion is more than high prices, it is a mutual understanding among firms to coordinate their behavior. (...) Firms may periodically raise price in order to attempt to coordinate a move to a collusive equilibrium, but never succeed in doing so; high average prices are then the product of failed attempts to collude."

based on these strategies predicts behavior, where these strategies dovetail with repeated game strategies in the simpler environment studied here.[8]

An experimental literature on behavior in oligopolies documents that collusion clearly responds to the number of players. Both Cournot (Huck, Normann, and Oechssler, 2004; Horstmann, Krämer, and Schnurr, 2018) and Bertrand settings (Dufwenberg and Gneezy, 2000) indicate that as the number of players increases collusion becomes less likely, often as soon as N exceeds two.[9] We contribute to this literature on two margins. First, the mentioned papers focus on out-of-equilibrium behavior in settings with a finite-time horizon, and a subsequently unique theoretical prediction. In contrast, we examine how changes to N affect outcomes in an infinite-horizon with both collusive and noncollusive equilibria. Second, and crucially, our focus is not only on the qualitative directional effects of N, but foremost on validating the model suitability for studying strategic uncertainty. Specifically, the model, if validated, will allow us to understand the extent of substitutability between game primitives, which can help to predict the directional effects of more-nuanced, multi-dimensional counterfactuals. Clearly, any equilibrium-selection notion suggested for such a task requires a great deal of scrutiny. However, our findings suggest some optimism for future research.

Our work is also related to the experimental literature on mergers that manipulates the number of players. As surveyed by Goette and Schmutzler (2009), some experiments deal with "pseudo-mergers," where a subset of the original firms remains in the market (see, for example, Huck, Konrad, Müller, and Normann, 2007). Other experiments implement "real mergers," where mergers introduce other changes in the market beyond N (Davis, 2002). Our contribution is that the strategic-uncertainty measure can predict counterfactual behavior in both settings. Another discussion in this literature is whether merger effects are evaluated within the same group of participants (within-subject designs) or across different groups (between-subject designs). In this paper, while our main treatments rely on between-subject identification, we also conduct within-subject sessions at the same parameterization, demonstrating that although there can be meaningful short-run differences, with enough experience the results align.[10]

The effects of communication devices as a bolster for collusion are well established in the experimental literature. As surveyed in Cason (2008) and Harrington, Gonzalez, and Kujal (2013), more-structured, limited forms of communication usually result in small, temporary collusive gains, where free-form communication generates large, long-lasting

[8]Work on equilibrium selection in dynamic games builds on recent contributions in this area. For example: Battaglini, Nunnari, and Palfrey (2012, 2016); Agranov, Frechette, Palfrey, and Vespa (2016); Kloosterman (2019); Vespa and Wilson (2019); Rosokha and Wei (2020); Salz and Vespa (2020); Vespa (2020).

[9]See also references in Potters and Suetens (2013) for similar findings.

[10]Differences in behavior can be stickier if changes are small or introduced gradually. Weber (2006) shows that gradually increasing the number of players in a coordination game leads to different results relative to a situation in which the game starts with a large group. The gradual introduction of changes to the payoff primitives has also been found to have effects in repeated games; see Kartal, Müller, and Tremewan (2017). These results suggest that the selection notions we are examining are relevant for "large" counterfactual changes, and where future research can help clarify how to integrate "large" into a predictive model of selection.

effects.[11] For these reasons, in one of our extensions we examine unrestricted chat messages as a strong coordination device. Our collusive results indicate that the domain for our strategic-uncertainty measure based on tacit collusion does not include environments where explicit collusion is allowed. However, we show that there are clear limits on the effects of explicit collusion, and these limits are directly predicted by theory. Using a change to the payoff primitives (here the discount rate) we make collusion a knife-edge, nonrobust equilibrium, and show that the effects of communication dissipate entirely.

While much of the literature on repeated games studies the standard two-player RPD, there is a large literature studying a canonical N-player social dilemma: the voluntary contribution public-goods game (see Vesterlund, 2016, for a survey). Although much of this literature focuses on finite implementations, one notable exception is Lugovskyy, Puzzello, Sorensen, Walker, and Williams (2017). Similar to our paper, the authors use experimental variation over both N and the payoff primitives (in their case, the return to the group contribution). However, this is done with a different end goal: to identify the isolated effect of the stage game's MPCR. Instead, our objective is to isolate strategic uncertainty and test a predictive theory of selection.

Beyond social dilemmas, our paper is also related to the literature on coordination games (see Devetag and Ortmann, 2007, for a survey). The strategic-uncertainty measure examined in our paper works because the RPD has a stag-hunt normal-form representation (Blonski and Spagnolo, 2015), adapting the risk-dominance notion for one-shot coordination games as in Harsanyi and Selten (1988).[12] Risk dominance (and the cardinal implementation through the measure of strategic uncertainty) has been shown to have substantial predictive content in simple coordination games with trade-offs over payoff-dominance and risk-dominance (see Battalio, Samuelson, and Van Huyck, 2001; Brandts and Cooper, 2006, and references therein). Therefore, strategic uncertainty has demonstrated its usefulness as a theoretical selection device in both static and repeated games. Our contribution to this literature is to design an experiment that will explicitly test and show how the predictive effects extend further, to multiplayer infinite-horizon settings.

2. Generalizing the Basin of Attraction

Developing empirical criteria for equilibrium selection in games in which collusion is possible requires two measures: one theoretical, one empirical. On the theory side, we need a prediction, a model that maps the primitives of the game into a scalar where upward/downward movements are clearly interpretable as increasing/decreasing the likelihood of collusion. On the empirical side, we need a precise target against which the theoretical notion can be contrasted and validated. This empirical measure should examine a behavior that differs starkly under the collusive and noncollusive equilibria.

[11] For further details on the effect of communication in repeated games with an unknown time horizon, see Fonseca and Normann (2012), Cooper and Kühn (2014), Harrington, Gonzalez, and Kujal (2016), and Wilson and Vespa (2020).

[12] The difference in our setting is that neither total payoffs nor strategic choices are directly provided to the participants, as these are extensive-form objects. Instead, they are provided with the stage-game payoffs/actions, where strategies (such as grim trigger or tit for tat) and gross payoffs are endogenously formulated.

We begin this section by summarizing the progress made towards validating a theoretical prediction in the two-player RPD literature. The theoretical notion here is the size of the basin of attraction for always defect. The focal outcome measures are the initial and ongoing cooperation rates of individual players. Then, we extend this framework by introducing a new source of strategic uncertainty, the number of players N—both for the theoretical and empirical measures.

2.1. **Two-players.** Consider an RPD with a discount rate $\delta \in (0,1)$. In each period $t = 1,2,\ldots$ players $i \in \{1,2\}$ simultaneously select actions $a_i \in \mathcal{A}$:={(C)ooperate, (D)efect}. The period-payoff for player i is a function of both players' choices, $\pi_i(a_i,a_j)$, where all symmetric PD stage-games can be expressed in a compact form by normalizing all payoffs relative to the joint-defection payoff $\pi_0 := \pi(D,D)$, and rescaling with the relative gain from joint cooperation: $\Delta\pi := \pi(C,C) - \pi_0$.[13] Defining scale and normalization in this way, the PD stage-game can be expressed with two parameters g and s for the different-action payoffs $\pi_i(D,C) = \pi_0 + (1+g)\Delta\pi$ and $\pi_i(C,D) = \pi_0 - s \cdot \Delta\pi$. The parameters $g > 0$ and $s > 0$ thereby capture the relative temptation- and sucker-payoffs, respectively.

The PD stage-game payoffs can be used as primitive inputs into a risk/reward model of collusion based upon strategic uncertainty. Here, strategic uncertainty is distilled into a decision between two focal extensive-form RPD strategies.[14]

(i) The *always defect*, $\alpha_{\text{All-D}}$, which plays the stage-game Nash in all rounds (the unique MPE of the game).[15]

(ii) The *grim trigger*, α_{Grim}, a strategy that begins by cooperating but switches to the always defect after observing any defections in past play (the best-case collusive SPE).[16]

As functions of the observable history h_t, these two strategies are given by:

$$\alpha_{\text{Grim}}(h_t) = \begin{cases} C & \text{if } t = 1 \text{ or } h_t = ((C,C),(C,C),\ldots,(C,C)), \\ D & \text{otherwise;} \end{cases}$$

$$\alpha_{\text{All-D}}(h_t) = D.$$

Strategic uncertainty in the two-player RPD is measured through the size of the basin of attraction for always defect. The model considers the expected reward for player i when uncertainty on the other player j is represented by a believed strategy mixture

[13] More exactly, game payoffs π can be transformed as $\tilde{\pi} = (\pi_i - \pi_0)/\Delta\pi$ to measure all payoffs relative to joint defection in units of the optimization premium.

[14] In the Online Appendix E, we describe why it is useful and not very restrictive to focus on these two strategies.

[15] A Markov strategy is history independent, removing any conditioning on past play. A Markov-perfect equilibrium (MPE) is a subgame-perfect equilibrium (SPE) in which agents use Markov strategies. In an RPD, if choices are forced to be history independent then there is a unique equilibrium: playing the stage-game Nash equilibrium in all periods.

[16] The strategy here is 'best case' in three senses: (i) It can support the best-case outcome. (ii) It uses the harshest possible punishment, and so can support collusion at smaller values of δ than any other strategy. (iii) Any realized miscoordination is minimal and resolves in a single round.

$p \cdot \alpha_{\text{Grim}} \oplus (1-p) \cdot \alpha_{\text{All-D}}$. The basin for always defect is defined as the set of beliefs p for which the player i receives a higher expected payment from $\alpha_{\text{All-D}}$ than α_{Grim}. The always-defect belief basin is therefore the interval $[0, p^\star(g,s,\delta)]$, where the critical-point/interval-width is given by:[17]

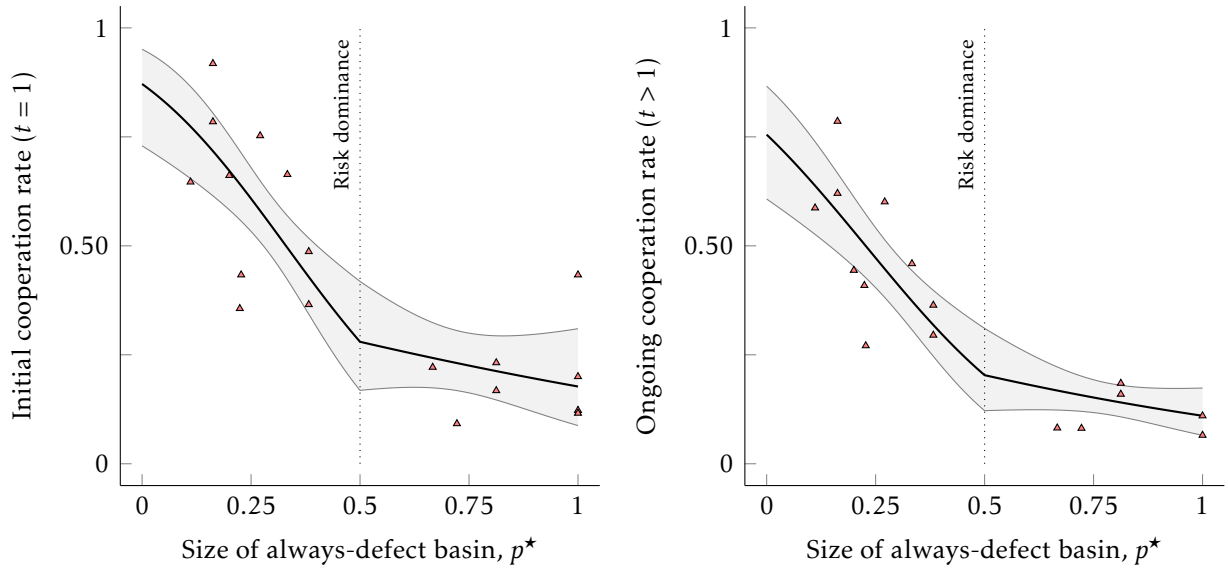
$$p^\star(g,s,\delta) \equiv \frac{(1-\delta)\cdot s}{\delta - (1-\delta)\cdot(g-s)}. \tag{1}$$

The size of the basin $p^\star(g,s,\delta)$ has a clear interpretation for strategic uncertainty: for any belief $p > p^\star(g,s,\delta)$ that the other player uses the collusive strategy α_{Grim}, the player does strictly better choosing α_{Grim}; for any belief $p < p^\star(g,s,\delta)$, the player does strictly better by selecting the noncollusive strategy $\alpha_{\text{All-D}}$. As such, the smaller $p^\star$, the lower the strategic uncertainty surrounding collusion. Moreover, the cardinal basin-size measure directly implies the ordinal risk-dominance relationship between the two strategies. If $p^\star(g,s,\delta) < 1/2$ the collusive strategy α_{Grim} risk-dominates $\alpha_{\text{All-D}}$, and vice versa. Henceforth, by 'basin of $p^\star$' we mean the maximal belief in the other players' cooperating for which the strategy $\alpha_{\text{All-D}}$ is optimal.

Equation (1) represents an easy-to-derive theoretical relationship between the payoff primitives of the game (here g, s, and δ) and a critical strategic belief over the other player's likelihood of collusion. The hypothesized relationship is monotone, where the higher $p^\star$, the lower the probability of cooperation, allowing unambiguous directional predictions for any counterfactual change in the primitives. The posited mechanism within this model is also clear cut: strategic uncertainty introduces a risk/reward trade-off for collusion attempts, which can be solved using standard economic analysis. This has two benefits. First, we can test the underlying strategic-uncertainty mechanism through other channels outside of the model, where we will do exactly that in an extension examining coordination devices. Second, the necessary assumptions for extending this model to analyze alternative sources of strategic uncertainty are straightforward.

We now turn to the empirical criteria used to validate this theoretical measure. As a summary of the RPD literature, we focus on the recent meta-study on the two-player RPD (Dal Bó and Fréchette, 2018). One of our main results shows that the scalar basin-size measure of strategic uncertainty is highly predictive of behavior, though with a nonlinear relationship. We illustrate this relationship in Figure 1 for two empirical outcome measures. In both panels of Figure 1 the theoretical measure of strategic uncertainty (the size of the basin $p^\star$) is presented on the horizontal axis. The empirical outcome measures are presented on the vertical axes. In Panel (A) we present the results for *initial cooperation* in the supergame, a measure that tracks collusive *intentions* at the individual level; in Panel (B) we present results for *ongoing cooperation*, choices in rounds two and beyond, a measure of the extent to which collusion attempts are successful. The solid line in both panels indicates the predicted cooperation rate at each $p^\star$ from the piecewise-linear probit

[17] In the case that the strategy $(\alpha_{\text{Grim}}, \alpha_{\text{Grim}})$ is not an SPE of the repeated game, the basin size for always defect is defined as $p^\star(x_T, x_S, \delta) = 1$.



(A) Initial cooperation

(B) Ongoing cooperation

FIGURE 1. Meta-study relationship: strategic uncertainty and RPD cooperation

Note: Figures show estimated effects and 95-percent confidence intervals for initial/ongoing cooperation in RPD meta-study (Dal Bó and Fréchette, 2018). Each point indicates a separate treatment.

estimates; the shaded region represents the 95 percent confidence interval for the prediction (clustering by treatment).[18]

For both initial and ongoing cooperation, the same pattern is found: a consistently low cooperation level when always-defect is risk dominant ($p^\star > \frac{1}{2}$); and a significantly decreasing relationship with $p^\star$ when collusion is risk dominant ($p^\star < \frac{1}{2}$). Specifically, Figure 1(A) illustrates Results 3 and 4 from Dal Bó and Fréchette (2018) that pertain to initial cooperation and Figure 1(B) portrays a relationship between strategic uncertainty and ongoing cooperation.

The theoretical model used in the basin construction posits a connection between initial and ongoing cooperation. If collusion functions through conditional cooperation with grim-trigger punishments, the expected ongoing cooperation rate is the probability that the players jointly cooperate in the first round: the initial cooperation rate squared. Thus, if cooperation were effectively governed by the grim trigger, both measures of empirical cooperation would carry the same information. Since, in fact, grim-trigger punishments have been documented to be used by subjects (for example, Dal Bó and Fréchette, 2011),

[18] We estimate the probit regression using meta-study data from 996 participants across 18 experimental treatments, where we focus on late-session cooperation (supergames 16–20, the data we focus on in our experiments). Individual-level cooperation decisions are the left-hand side variable, where the basin size enters the right-hand side in a piecewise-linear fashion around the risk-dominance dividing point. The econometric specification is motivated by Dal Bó and Fréchette (2018, Table 4); however, to enforce level-continuity in the estimated relationship, we remove a degree of freedom from their specification that allowed a discontinuity at $p^\star = 1/2$.

data from the RPD do not provide enough variation to identify whether theoretical notions track one of these measures more closely. Consequently, with only two players it is challenging to identify the extent to which the strategic-uncertainty measure predicts initial intentions versus successful coordination.[19] However, as we will show below, adding more players provides additional variation that will allow us to differentiate between the two measures of cooperation.[20]

2.2. **Extending to $N > 2$.** We now extend the strategic-uncertainty model to an N-player environment. The core extension is intuitive: instead of considering one other player choosing a mixture $p \cdot \alpha_{\text{Grim}} \oplus (1-p) \cdot \alpha_{\text{All-D}}$, we consider $N-1$ other players choosing a strategic mixture.

Our goal in this section is to introduce two simple ways to construct the extension to $N > 2$. Perhaps the most standard approach is to assume that each player treats the decisions of the other players to play α_{Grim} as independent. An alternative approach is to assume that the choices of all other $N-1$ players are perfectly correlated.[21] In the first approach, increasing N increases strategic uncertainty because the players are independent, and adding one more player adds one more person whose behavior is unknown. In the second approach, strategic uncertainty is not affected by N, but continues to depend on other model primitives, as discussed for the two-player RPD. We focus on these two extreme cases of full independence and full correlation for the following reasons. First, they allow us to produce an experimental design (introduced in the next section) that has stark behavioral predictions. Second, both stances are simple to compute in settings beyond our environment.[22]

To outline the extension and set up our design, we consider a family of symmetric social dilemmas that nest the standard two-player RPD. However, to maintain a constant 2×2 stage-game representation for all N, our family of dilemmas makes use of an aggregate (and deterministic) signal of the other agents' actions. All players $i = 1, \ldots, N$ continue to make a binary action choice $a_i \in \mathcal{A} \equiv \{C, D\}$, but their payoffs do not vary with (and they do not receive feedback on) the separate actions of the other $N-1$ players. Instead, payoffs are determined by the own-action a_i and a deterministic binary signal $\sigma(a_{-i}) \in \{S(\text{uccess}), F(\text{ailure})\}$ of the actions of the others, a_{-i}. In particular, the generic player i's

[19] For a setting that achieves this with sequentiality of moves, see Ghidoni and Suetens (2022).

[20] For any setting in which collusion requires N agents to initially cooperate to produce ongoing cooperation, the relationship is simply given by initial cooperation rate to the N-th power. Separate identification between the two measures occurs upon comparing treatments with different values of N.

[21] Notice that both extensions of the measure capture beliefs over supergame strategies (i.e a full specification of what to do in every possible round). In the strategies underlying the measure, actions will be perfectly correlated in all rounds but the first. For instance, consider α_{Grim}. Either all N players successfully coordinate on cooperation, or after a failure in round one, the punishment path is triggered with all N players choosing defect in subsequent rounds. As such, the independent and correlated models only differ in the potential for correlation in play in the very first round.

[22] Clearly, one can define an intermediate hypothesis with an extra parameter that captures the extent to which beliefs are independent (with complementary probability on the extent to which beliefs are correlated). In Section 4 we will discuss this alternative in further detail and report an estimate of such a parameter.

stage-game payoff and signal function are given, respectively, by:

$$\pi_i\big(a_i,\sigma\big)=\begin{cases}\pi_0+\Delta\pi & \text{if } a_i=C,\sigma=S,\\ \pi_0+\Delta\pi\cdot(1+x) & \text{if } a_i=D,\sigma=S,\\ \pi_0-\Delta\pi\cdot x & \text{if } a_i=C,\sigma=F,\\ \pi_0 & \text{if } a_i=D,\sigma=F;\end{cases}$$

$$\sigma(a_{-i})=\begin{cases}S & \text{if } a_j=C \text{ for all } j\neq i,\\ F & \text{otherwise.}\end{cases}$$

These choices lead to a symmetric game, where payoffs can be summarized with a 2×2 table over: (i) the own action C or D; and (ii) the signal outcome, an S signal if the other $N-1$ players jointly cooperate, or an F signal if at least one other player defects. The exact payoffs (with the same implicit scale $\Delta\pi$ and normalization π_0 as before) implement a PD-like environment where payoff-based strategic uncertainty is collapsed to a single parameter x.[23]

The success/failure signal is a deterministic function of the actions of the $N-1$ other players, corresponding to the standard RPD two-player game. The choice for the signal function $\sigma(\cdot)$ maximizes the coordinative pressure, duplicating a Bertrand-like tension: collusion is successful only when all other $N-1$ players cooperate.[24]

Ignoring the scale and normalization of the game (held constant in our experiments with $\Delta\pi=\$9$ and $\pi_0=\$11$) the repeated games we examine are summarized by three primitives: (i) The relative cost of cooperating, x. (ii) The number of players, N. (iii) The continuation probability, δ. Our experiments fix $\delta=3/4$ in all but one diagnostic treatment in Section 5. This leaves us with two key experimental parameters: the relative cost x (actual cost $X=x\Delta\pi$) and the number of participants N.

In building a model of strategic uncertainty for arbitrary N, we use a symmetric belief over the others' choices. That is, we assume each player chooses a mixture $p\cdot\alpha_{\text{Grim}}\oplus(1-p)\cdot\alpha_{\text{All-D}}$ over the two strategies.[25] Our family of social dilemmas require cooperation from all N players for everyone to get an S signal. Thus, the strategic uncertainty reduces to the probability that the other $N-1$ players *jointly* coordinate on the collusive strategy,

$$Q(N)=\Pr\{N-1 \text{ others all choose } \alpha_{\text{Grim}}\}.$$

[23] In the meta-study notation this is implemented with $s=g=x$. This single-parameter formulation is equivalent to the Fudenberg, Rand, and Dreber (2010) benefit/cost formulation, where their benefit/cost ratio parameter (b/c) is given by $(1+x)/x$ here.

[24] In Section 5, we introduce a manipulation in which only two of four players are needed for the cooperative outcome to occur. If cooperative outcomes can be achieved with some players not cooperating, a free-riding problem emerges. In our main treatments, we abstract from this issue by having efficient outcomes only if all players cooperate.

[25] For the N-player dilemma we define the grim-trigger strategy with imperfect-signals as:

$$\alpha_{\text{Grim}}(h_t)=\begin{cases}C & \text{if } t=1 \text{ or } h_t=((C,S),(C,S),\ldots,(C,S)),\\ D & \text{otherwise.}\end{cases}$$

In every other case, at least $N-1$ players will receive an F signal and the punishment path will be triggered.

Identically to the case of two players, the critical belief $Q^\star(N)$ is given by the point of indifference between the amount given up with certainty from a single round of cooperation, $x \cdot \Delta\pi$, and the continuation gain from collusion, $\frac{\delta}{1-\delta} \cdot \Delta\pi$, obtained with probability $Q(N)$. The critical belief is therefore given by:

$$Q^\star(N) = \frac{(1-\delta)}{\delta} x,$$

where the RHS is identical to the two-player construction in Equation (1) for $x = g = s$.

Next, we need to relate the joint cooperation of the other $N-1$ players to the probability p that every other player individually attempts to collude. However, even though we have specified the marginal belief distribution and assumed symmetry, we must still resolve the relationship between the joint and marginal distributions: the extent to which beliefs are correlated. In particular, our design focuses on two extremes. The 'standard' extension in which beliefs are fully independent; and an alternative/null-effect model in which beliefs are perfectly correlated.[26] Assuming perfect correlation for the other $N-1$ agents, joint and individual probabilities are identical, $Q(N) = p$, and so the extended critical belief (and correlated model outcome) is given by:

$$p^\star_{\text{Corr.}}(x) = \frac{1-\delta}{\delta} \cdot x. \tag{2}$$

In contrast, when beliefs are independent, we have $Q(N) = p^{N-1}$, and the critical belief (and independent-model extension) is given by:

$$p^\star_{\text{Ind.}}(x, N) = \left(\frac{1-\delta}{\delta} \cdot x\right)^{\frac{1}{N-1}} \equiv \left(p^\star_{\text{Corr.}}(x)\right)^{\frac{1}{N-1}}. \tag{3}$$

Obviously, when $N = 2$ the basin measures in Equations (2) and (3) are identical, matching the standard construction. However, for $N > 2$ the two measures of strategic uncertainty are distinct, where the standard model extension under independence also depends on the group-size N.

3. Experimental Design

The basin of attraction for always defect serves as our measure of strategic uncertainty. Ceteris paribus, the greater the uncertainty on successful strategic coordination, the more likely the subject is to take refuge in a safer strategy—in the case of an RPD game, the stage-game Nash outcome of defecting.

Extending the environment to $N > 2$ raises a question of how strategic uncertainty is affected by N. If others' behavior is highly correlated, adding players but holding the payoffs constant will do little to affect the behavior. In contrast, in a more-standard extension in which beliefs over the other players are independent, $p^\star_{\text{Ind.}}$ will model the strategic uncertainty. Under this extension, we will be able to use shifts in $p^\star_{\text{Ind.}}$ to understand

[26] See Cason, Sharma, and Vadovič (2020) for an example of correlated beliefs that arise where independence would be the standard prediction.

changes in the selected behavior. While one implication is that additional players ceteris paribus reduce collusion, a deeper implication of the model is to help us understand substitution effects across the two sources of strategic uncertainty, x and N.

Our experimental design attempts to untangle the effects of strategic uncertainty. The aim of the design is to embed comparative-static tests on the effect of N (and thus rule out the $p^{\star}_{\text{Corr.}}$ null-effect model), but also to examine the possible substitution effects by constructing perfect substitution treatments with the $p^{\star}_{\text{Ind.}}$ model. We achieve this with a series of experimental implementations of the N-player two-action–two-signal repeated game outlined above. In particular, the first part of our design aims to distinguish and separate between the two extremes of independence and perfect correlation, leveraging the theoretical relationships derived above.

While we cannot directly manipulate strategic uncertainty—as the basin-size measures are indirect, theoretical relationships derived from the primitives—Equations (2) and (3) allow us to implicitly manipulate each measure through shifts in x and/or N. Increases in x increase the strategic uncertainty in both models: higher costs of cooperation require a greater belief that the other(s) are cooperating. In contrast, increases in the number of players N only increase strategic uncertainty for the independent-basin measure, interacting with x in a nonlinear manner.

Using Equations (2) and (3), the two notions can be varied in isolation. As such, it is possible to construct a 2×2 design that orthogonally varies each strategic-uncertainty measure. Next, we outline our design, which we also summarize in Table 1.

Panel (A) of Table 1 illustrates our first treatment dimension, which manipulates the payoff cost of cooperating $X = x \cdot \Delta\pi$, where $\Delta\pi = \$9$. The two values of X—a high temptation of \$9 (illustrated on the left, $x = 1$), and a low temptation of \$1 (on the right, $x = 1/9$)—lead to two payoff environments over own-actions and the signals.[27]

Our design also manipulates the number of players N, captured in the column headings of Panel (B) in Table 1. In total, we create four treatment environments, each defined by an (N, X)-pair. The two rows of Panel (B) indicate how the choices over X and N affect the basin-size measures of strategic uncertainty under the correlated and independent extensions.

To manipulate each basin-size measure separately, our design takes $\binom{N=2}{X=\$9}$ as its starting point. For this treatment, the values for both the independent and correlated basin-size measures are the same: $p^{\star}_0 = 0.33$. Holding the relative cooperation cost fixed and increasing the number of players to $N = 4$ do not affect the correlated measure in Equation (2). However, a shift to $N = 4$ increases the independent-basin measure to $p^{\star}_0 + \Delta p^{\star}_{\text{Ind.}} = 0.69$.

Now, consider the manipulation of X. Comparing $\binom{N=4}{X=\$1}$ with $\binom{N=2}{X=\$9}$, we hold constant the independent-basin measure at $p^{\star}_0 = 0.33$. The shifts in both X and N have perfectly substituting effects in Equation (3). However, the same change in both variables under the correlated basin has a substantial effect, as the change in N does not offset the change in X. As such, the correlated-basin measure is lowered to $p^{\star}_0 - \Delta p^{\star}_{\text{Corr.}} = 0.04$. Finally,

[27] See Figure C.1 in the Online Appendix C for representative lab screenshots.

TABLE 1. Experimental design

Panel A. **Stage-game payoffs**	$X = \$9$		$X = \$1$	
	$\sigma(a_{-i}) = S$	$\sigma(a_{-i}) = F$	$\sigma(a_{-i}) = S$	$\sigma(a_{-i}) = F$
Coop., $\pi_i(C, \sigma)$	\$20	\$2	\$20	\$10
Defect, $\pi_i(D, \sigma)$	\$29	\$11	\$21	\$11
Panel B. **All-D Basin Size**	$X = \$9$ $(x = 1)$		$X = \$1$ $(x = 1/9)$	
	$N = 2$	$N = 4$	$N = 4$	$N = 10$
Cor. basin, $p^\star_{\text{Cor.}}(x)$	$p^\star_0$ [0.33]	$p^\star_0$ [0.33]	$p^\star_0 - \Delta p^\star_{\text{Cor.}}$ [0.04]	$p^\star_0 - \Delta p^\star_{\text{Cor.}}$ [0.04]
Ind. basin, $p^\star_{\text{Ind.}}(x, N)$	$p^\star_0$ [0.33]	$p^\star_0 + \Delta p^\star_{\text{Ind.}}$ [0.69]	$p^\star_0$ [0.33]	$p^\star_0 + \Delta p^\star_{\text{Ind.}}$ [0.69]
Sessions	3	3	3	2
Subjects	60	60	72	60
Panel C. **Meta-study predictions**	$p^\star_0$	Marginal effect from:		
	[0.33]	Basin increase to [0.69]	Basin decrease to [0.04]	
Initial coop. $(t = 1)$	0.50	−0.26	+0.35	
Ongoing coop. $(t > 1)$	0.37	−0.21	+0.50	

Note: Meta-study predictions in Panel (C) correspond to estimates from the treatment-clustered probits illustrated in Figure 1.

in the $\binom{N=10}{X=\$1}$ treatment we complete the 2×2 design over the two basin-size measures. Comparing $\binom{N=4}{X=\$1}$ with $\binom{N=10}{X=\$1}$, we hold constant the correlated basin at $p^\star_0 - \Delta p^\star_{\text{Corr.}} = 0.04$, as it does not depend on N. However, more players increase strategic uncertainty in the independent basin. In particular, our parameterization matches the independent-basin sizes for $\binom{N=10}{X=\$1}$ and $\binom{N=4}{X=\$9}$ at $p^\star_0 + \Delta p^\star_{\text{Ind.}} = 0.69$.

Through variation in the primitives X and N, our design thereby generates four correlated/independent basin measure pairs with a 2×2 structure:[28]

$$\left(p^\star_{\text{Corr.}}, p^\star_{\text{Ind.}}\right) \in \left\{p^\star_0, p^\star_0 - \Delta p^\star_{\text{Corr.}}\right\} \times \left\{p^\star_0, p^\star_0 + \Delta p^\star_{\text{Ind.}}\right\} := \left\{0.33, 0.04\right\} \times \left\{0.33, 0.69\right\}.$$

This design achieves the goal of orthogonal variation over the two basin measures. However, the above parameterization was also chosen so that the shifts in each dimension are expected to have quantitatively large effects. Through the Dal Bó and Fréchette (2018) meta-study we generate level predictions for the behavioral effects of each directional

[28] We note that our choices of $\Delta\pi = \$9$ and $\delta = 3/4$ were motivated by simplicity of the presentation: our results are integer-valued for both N and X. Our design over the basin measures is more-exactly given by:

$$\left(p^\star_{\text{Corr.}}, p^\star_{\text{Ind.}}\right) \in \left\{3^{-1}, 3^{-3}\right\} \times \left\{3^{-1}, 3^{-1/3}\right\}.$$

change. Our design generates three basin-size measures: $p_0^\star = 0.33$, $p_0^\star - \Delta p^\star_{\text{Corr.}} = 0.04$, and $p_0^\star + \Delta p^\star_{\text{Ind.}} = 0.69$. Using the probit-model estimates illustrated in Figure 1 we make *quantitative* predictions for the cooperation rates under each basin-size measure. These predictions are indicated in Panel (C) of Table 1. The first column reports the initial and ongoing cooperation rates expected at $p^\star = 0.33$.[29] The next two columns indicate the expected treatment effect from a shift in the strategic uncertainty from $p^\star = 0.33$ to either $p_0^\star - \Delta p^\star_{\text{Corr.}} = 0.04$, or $p_0^\star + \Delta p^\star_{\text{Ind.}} = 0.69$.

We formalize the two competing hypotheses as:

Correlated-Basin/Null-effect Hypothesis. *Cooperation increases as we decrease X, but there is no effect as we vary the number of players N.*

Independent-Basin Hypothesis. *Cooperation decreases as we increase X and/or N. Moreover, the substitution effects between X and N indicate no effect on cooperation if we decrease X and increase N to hold constant the $p^\star_{Ind.}$ measure of strategic uncertainty.*

That is, consider now the predictions under the standard independence-based extension of strategic uncertainty. In treatments $\binom{N=2}{X=\$9}$ and $\binom{N=4}{X=\$1}$ the independent basin size is 0.33, and it increases to 0.69 in treatments $\binom{N=4}{X=\$9}$ and $\binom{N=10}{X=\$1}$. If the strategic uncertainty relationship estimated from the two-player RPD meta-data is perfectly extrapolatable to our setting, we should expect: (i) A reduction of 26 (21) percentage points in initial (ongoing) cooperation across the treatment pairs, caused by an increase in strategic uncertainty. (ii) A null effect on cooperation within each treatment pair, reflecting the designed perfect substitution across X and N in the independence-based measure.[30]

Notice that our hypotheses are silent with respect to which of the two outcome measures, initial and/or ongoing cooperation, we are supposed to match. The two measures have different interpretations—whereas initial cooperation captures intentions, ongoing cooperation reflects successful coordination. In the case of the two-player RPD, Figure 1 shows that the basin size tracks both cooperation measures relatively well and that the effects are hard to disentangle. Through N, we are able to generate additional variation in the theoretical relationship between initial and ongoing cooperation variables that separates the two relationships on another dimension. An advantage of this design is that it will allow us to identify the measure better predicted by either of the two basin-size models.

Experimental Specifics. In our experiment, we used a between-subject design over the four distinct treatments described in Table 1. Participants for each treatment were recruited from the undergraduate population at the University of Pittsburgh, and each participated in only one session. We recruited a total of 584 participants, 252 for the first four treatments and 332 for the extensions that we outline in Section 5. Three sessions

[29] All predictions are based on late-session meta-study data (supergames 16–20).

[30] Alternatively, under a null-effect from N, given by the correlated-basin measure, the basin size is reduced from 0.33 to 0.04 as we move between the $\binom{N=2}{X=\$9}$ and $\binom{N=4}{X=\$9}$ treatment pair and the $\binom{N=4}{X=\$1}$ and $\binom{N=10}{X=\$1}$ pair. The RPD prediction from the meta-study then is for an increase in the initial (ongoing) cooperation rate of 35 (50) percentage points (and again, a null effect within each pair).

were held for each treatment, with the goal of recruiting at least 20 participants per session, with one exception of the $\binom{N=10}{X=\$1}$ treatment for which we ran two sessions of 30.[31] Sessions lasted between 55 and 90 minutes with participants receiving an average payment of approximately $19.

Each session comprised 20 supergames, with a common random termination chance of $1-\delta = 1/4$ after each completed round.[32] The participants were randomly and anonymously matched in the 20 supergames in a stranger design.[33] The 20 supergames were divided into two parts of ten supergames.[34] For final payment, one supergame from each part was randomly selected, where only the actions/signals from the last round in the selected supergame counted for payment.[35]

4. Results

We begin this section by describing the aggregate cooperation rates at the treatment level. Then, we proceed to discussing inferential tests of our two basin-extension hypotheses. Our main finding is that while neither extension contains all the relevant information for predicting initial cooperation, the basin-size measure based on a standard independence assumption delivers more definitive results for ongoing cooperation within the experimental supergames.

4.1. Main Treatment Differences.

Table 2 reports average cooperation rates broken out by the four treatments, where we separately report initial (the first round) and ongoing cooperation (all subsequent rounds). The averages are computed for the last five supergames—which capture late-session behavior, after subjects have amassed experience in the environment—though including all rounds generates similar results (see Table A.1 in the Online Appendix A). Overall, the results indicate large shifts in cooperation as we vary the cost of cooperation X and/or the size of the group N.

The first row of Table 2 summarizes initial cooperation rates. The initial cooperation rate in the $\binom{N=2}{X=\$9}$ treatment is 50.3 percent, essentially identical to the 50 percent cooperation rate predicted by the RPD meta-study. However, holding constant the cooperation cost at $X = \$9$ and doubling the group size to four virtually eliminates cooperative behavior,

[31] In more detail, our design called for sessions to have at least 20 participants but allowed us to recruit an additional group of size N depending on realized show ups. For $\binom{N=10}{X=\$1}$ we instead opted to recruit 30 participants for each session so that we had at least three groups in each supergame.

[32] However, we used common draws to keep supergame lengths matched at the session-level by treatment.

[33] All subjects received written and verbal instructions on the task and payoffs, where instructions are provided for interested readers in the Online Appendix D.

[34] Subjects received full instructions for the first part and were told they would be given instructions for the second part after completing supergame ten. For the four between-subject treatments outlined in Section 3, part two was then identical to part one. Later in the paper, we will outline a further set of treatments with a within-subject change across the parts. The design choice for two identical parts here allows for direct comparisons in first-half play.

[35] This method is developed in Sherstyuk, Tarui, and Saijo (2013) to induce risk neutrality over supergame lengths. Another benefit from this design choice is that there are no wealth effects within a supergame; moreover, history only matters as an instrument for the future play of others.

TABLE 2. Cooperation rates and basin-effect decomposition

Action and signal rates	$X = \$9$		$X = \$1$	
	$N = 2$	$N = 4$	$N = 4$	$N = 10$
Initial coop.	0.503 (0.058)	0.035 (0.017)	0.792 (0.042)	0.357 (0.055)
Ongoing coop.	0.450 (0.055)	0.006 (0.003)	0.409 (0.050)	0.184 (0.048)
Initial success	0.503	0.000	0.578	0.000
Ongoing success	0.450	0.000	0.293	0.000

Note: Results are calculated using data from the last-five supergames. Cooperation rates present raw proportions (with subject-clustered standard errors).

with just 3.5 percent initial cooperation in $\binom{N=4}{X=\$9}$. In low-temptation settings ($X = \$1$), groups of $N = 4$ show highly cooperative initial behavior (79.2 percent), while groups of $N = 10$ generate moderate first-round cooperation rates (35.7 percent).

The second row of Table 2 indicates ongoing ($t > 1$) cooperation rates. Here, the data indicate a decrease in cooperation when compared to the initial behavior in all treatments, although the quantitative effects are largest in the $\$X = 1$ treatments.[36]

The third and fourth rows of Table 2 present the fraction of rounds in which a success signal was observed.[37] Focusing on success signals, similar patters emerge to ongoing cooperation, though with starker quantitative effects. Although a success is the modal signal in the $\binom{N=2}{X=\$9}$ and $\binom{N=4}{X=\$1}$ treatments, in the $\binom{N=4}{X=\$9}$ and $\binom{N=10}{X=\$1}$ treatments we observe *no* successes at all.[38]

Using only the raw averages, the evidence clearly falsifies the correlated basin/null hypothesis on the effect of N, for both initial and ongoing cooperation. The experimental results clearly indicate large changes in behavior as we move N as a comparative static, fixing the value of X. For $N = 4$ we find the comparative-static effect predicted by the correlated-basin measure as we move X, though this directional effect is also predicted by the independent basin.

[36] In the Online Appendix A, Table A.2 further breaks out ongoing cooperation by the observed history in the previous round. The results indicate that individual cooperation is highly conditional on successful coordination. However, strategies are significantly more forgiving after failed cooperation at $X = \$1$ than $X = \$9$.

[37] A success requires that the other $N - 1$ participants jointly cooperate. Success is a direct function of group-level cooperation, where the *expected* success rate with an independent cooperation rate q is q^{N-1}. In two-player games, the success rate is identical to the cooperation rate. For the initial round the expected success rates (in the Table 2 column order) are: 0.503, 4.2×10^{-5}, 0.497 and 9.5×10^{-5}.

[38] As success is a direct aggregate of individual-level cooperation we do not report standard errors (where we also cannot calculate standard errors when there is no variation). However, the starkness of the effect with no successes when the independent-basin size is high clearly illustrates the underlying economic effects.

TABLE 3. Basin-effect decomposition

Cooperation decomposition	$p_0^\star$	Marginal effect from:	
	[0.33]	Ind. basin increase to $p_0^\star + \Delta p_{\text{Ind.}}^\star = [0.69]$	Corr. basin decrease to $p_0^\star - \Delta p_{\text{Corr.}}^\star = [0.04]$
Initial coop.	0.464 (0.058)	−0.395 (0.048)	+0.357 (0.053)
Ongoing coop.	0.366 (0.051)	−0.293 (0.051)	+0.115 (0.061)

Note: Results are calculated using data from the last-five supergames. The cooperation decomposition runs two subject-clustered probits, one for the initial, one for the ongoing cooperation. RHS variables are: a constant and two dummies, one for a low-correlated–basin treatment ($X = \$1$, both N values), one for a high-independent–basin treatment ($X = \$9/N = 4$ and $X = \$1/N = 10$).

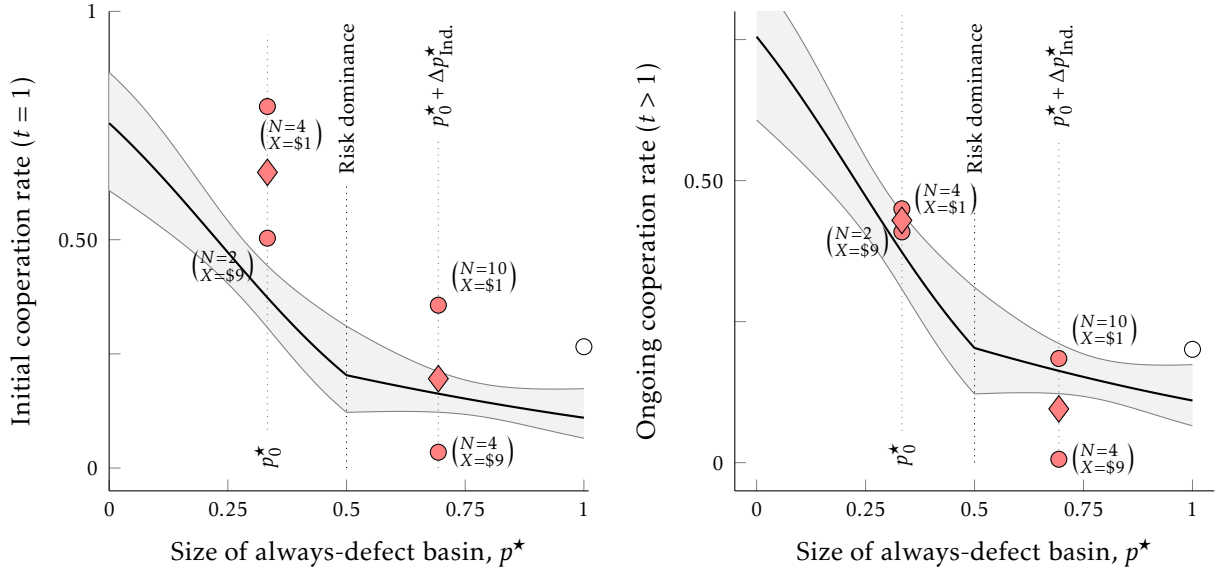
The evidence suggests that the independent-basin hypothesis is better at organizing the data. Both initial and ongoing cooperation respond in the predicted direction as we shift either X or N in isolation. However, for initial cooperation, we do not find perfect substitution as we move both X and N. While the independent-basin hypothesis does not predict any change in cooperation rates between $\binom{N=2}{X=\$9}$ and $\binom{N=4}{X=\$1}$ or $\binom{N=4}{X=\$9}$ and $\binom{N=10}{X=\$1}$, we find substantive differences in the laboratory.[39] Also, the average initial cooperation rate when pooling $\binom{N=2}{X=\$9}$ and $\binom{N=4}{X=\$9}$ differ from the average initial cooperation when pooling $\binom{N=4}{X=\$1}$ and $\binom{N=10}{X=\$1}$. Initial cooperations in the pooled treatments for $X = \$9$ and $X = \$1$ are 28.6 and 59.4 percent, respectively, and the difference is more than 30 percentage points.

Whereas initial cooperation rates are inconsistent with the independent-basin hypothesis, the results for ongoing cooperation under independence are considerably better, see treatments $\binom{N=2}{X=\$9}$ and $\binom{N=4}{X=\$1}$. Yet, we still note a difference between $\binom{N=4}{X=\$9}$ and $\binom{N=10}{X=\$1}$, which is primarily driven by a very stark finding of near-zero cooperation in $\binom{N=4}{X=\$9}$. We explore this further below. However, in the aggregate, when pooling the ongoing cooperation rates across X, we find similar rates at 22.8 percent for $X = \$9$ and 29.8 percent for $X = \$1$.

4.2. **Evaluation of Independent- and Correlated-Basin Hypotheses.** Table 3 provides a direct statistical evaluation of our two competing hypotheses. it reports results of probit regressions that assess subjects' cooperation decisions using dummy variables for the 2×2 design in Table 1 Panel (B). The dummy covariates are an indicator for the $\Delta p_{\text{Cor.}}^\star$ decrease in the correlated-basin size (as we decrease X), and an indicator for the $\Delta p_{\text{Ind.}}^\star$ increase in the independent-basin size (as we increase N within each X).

Each row of Table 3 provides the results of a separate estimation, one over initial cooperation and one over ongoing. The first column reports the estimated cooperation rate when both dummy variables are zero: essentially the cooperation rate for a game with a basin size of $p_0^\star = 0.33$. The other two columns report the estimated marginal effects on the cooperation rate for each basin shock, holding the other constant. If either of the two

[39] 29 percentage points in the first comparison and 35 percentage points in the second.



(A) Initial cooperation

(B) Ongoing cooperation

FIGURE 2. Cooperation and the independent basin-size model

Note: Filled circles indicate separate treatments and filled diamonds treatments pooled over each value of the independent-basin measure. Empty circles show the unilateral cooperation rates in the extension treatment discussed in Section 5.

basin hypotheses fully explains behavior, we would expect a significant estimate for the corresponding dummy and an insignificant effect on the other.

The estimation procedure is designed to directly parallel the probit model used to generate predictions from the meta-study. The estimated cooperation rates at $p_0^\star = 0.33$ are in fact quantitatively very close to the meta-study prediction in Panel (C) of Table 1. Whereas the meta-study predicts an initial (ongoing) cooperation of 49.5 (37.3) percent, our data at $p_0^\star = 0.33$ indicate similar (and statistically inseparable) rates of 46.4 (36.6) percent. To illustrate this, in Figure 2 we plot the fitted relationships from the meta-study overlayed with our results from the four treatments using the independent-basin size on the horizontal axis. Filled circles indicate separate treatments and filled diamonds treatments pooled over each value for the independent-basin measure. While there is substantial divergence for initial cooperation, Figure 2 indicates quantitatively similar results for ongoing cooperation.[40]

Our tests of the two competing hypotheses focus on the second and third columns of Table 3. If the independent-basin (correlated-basin) measure captures all relevant facets of behavior, we would expect a significantly negative (insignificant) estimate for the independent-basin increase (correlated-basin decrease) and an insignificant (significantly negative) effect on the correlated-basin (independent-basin) increase.

[40]In the Online Appendix A, Figure A.3 presents analogous results organized under the correlated-basin model. The figure illustrates much poorer organization of the data, both relatively across the treatment comparisons, and quantitatively.

For initial cooperation, we find that changes to both basin measures generate significant effects ($p < 0.001$). The magnitudes of each estimated marginal effect are very similar, but moving in opposite directions, as predicted. Since neither effect dominates, we conclude both X and N contain information for predicting initial cooperation that is not fully absorbed by either basin measure.

However, consistent with our descriptive presentation of the raw treatment rates in Table 2, the probit estimates for ongoing cooperation in Table 3 Panel (B) break toward the independent-basin construction. The coefficient on the increase in the independent basin is negative and significant ($p < 0.001$), while the estimate on the decrease for the correlated basin is much smaller in magnitude and insignificant at the 5 percent level ($p = 0.061$). Beyond qualitative directional effects, the quantitative change in ongoing cooperation under the independent-basin measure is close to the predicted effects of the standard basin shifts we would expect from the meta-study data.[41] That is, the meta-study predicts a drop of 21 percentage points (see Panel (C) of Table 1) in ongoing cooperation when the size of the basin increases from $p_0^\star = 0.33$ to 0.69, whereas our estimates indicate a decrease by 29 percentage points.[42]

As mentioned above, the differences in the ongoing cooperation rates between our data and the out-of-sample prediction from the meta-study are driven by the stark (essentially boundary) behavior in the $\binom{N=4}{X=\$9}$ treatment. As illustrated in Figure 1(B) a two-player repeated game with a basin of size $p^\star = 0.69$ has a predicted ongoing cooperation rate of 16.3 percent, where we should be able to reject 11 percent cooperation at 95 percent confidence. While the $\binom{N=10}{X=\$1}$ treatment is close to the predicted rate (as are the two other treatments at $p^\star = 0.33$), Figure 1(B) clearly indicates that the $\binom{N=4}{X=\$9}$ treatment is significantly below the predicted level. However, when the noncollusive strategy is risk dominant (the independent-basin size is greater than 1/2), the related literature suggests that we should not expect substantial cooperation. While the model predicts low cooperation at this basin size (~20 percent), the evidence from the $\binom{N=4}{X=\$9}$ treatment pushes towards the same conclusion, just in a starker way. Aside from the more-extreme coordination effects at $\binom{N=4}{X=\$9}$, for the other three treatments the theoretically standard extension of

[41] Our measures of equilibrium selection aim to capture strategic uncertainty in a setting that differs from the two-player RPD. Ideally, a theoretical measure would incorporate all relevant features of the environment and provide guidance on equilibrium selection for any setting. The proposed measures (independent and correlated) aim to capture aspects of strategic uncertainty. Finding that results in our setting are comparable to those in the two-player RPD is useful because it suggests that a measure of strategic uncertainty may be a good predictor of collusion regardless of the specific details of the environment.

[42] In contrast, for a decrease to 0.04 we should expect an increase in ongoing cooperation of 50 percentage points; instead, we observe an increase of 11.5 percentage points.

the strategic-uncertainty measure comes very close to *quantitatively* predicting the ongoing cooperation level using the out-of-sample relationship estimated from the two-player RPD meta-data.[43]

Finally, we attempt to measure *how much* correlation is necessary to rationalize the data. To do this, we allow for the beliefs to be a convex combination of the two belief models, all $N-1$ agents choosing grim with probability p with perfect correlation σ of the time, and each choosing grim independently with probability p the remaining $(1-\sigma)$ of the time. As such, for an N-player game in which a player considers conditional cooperation, the probability of a coordination is given by

$$\sigma \cdot p + (1-\sigma) \cdot p^{N-1}.$$

This generalized basin of attraction solves for the indifferent belief $p^{\star}(\sigma, x, N)$ for which the agent is indifferent between choosing a grim or always-defect strategy, here with the additional parameter σ that nests the models at the two extremes: $\sigma = 0$ for full independence, $\sigma = 1$ for perfect correlation.[44] Looking at the best fitting parameter, for initial cooperation, we estimate $\hat{\sigma}_I = 0.091$, while the comparable estimate for ongoing cooperation is $\hat{\sigma}_O = 0.031$. As such, the estimated degree of correlation in the relaxed model is quantitatively small.

We summarize our main findings as:

Result 1 (Independent-Basin Measure)**.** *The independent-basin measure qualitatively organizes the results for ongoing cooperation, and in all but one treatment matches the quantitative level predictions. However, it does not contain all relevant information for predicting initial intentions to cooperate.*

Result 2 (Correlated-Basin Measure)**.** *Our data are inconsistent with the predictions from the correlated-basin hypothesis, both for initial and ongoing cooperation. In particular, whereas the correlated basin predicts that behavior should ceteris paribus be unaffected by N we find decreases in cooperation as N increases. Quantitatively, the estimated degree of belief correlation is small.*

---

[43] Some of the differences in cooperation rates at fixed values of the basin are driven by differences in the number of unconditional cooperators at $X = \$9$ in comparison to $X = \$1$. In Tables E.1 and E.2 in the Online Appendix E we present strategy frequency estimates from the first and last seven supergames. Our results here suggest much greater rates of unconditional cooperation in the $X = \$1$ treatments than $X = \$9$, though this falls across the session. Because variation in X is associated with shifts in the correlated-basin value (invariant to N), this presents a confound in interpretations for the small positive effect for the correlated basin. While it could represent belief correlation, it could also be driven by other-regarding preferences.

[44] Because the model now needs to make a *quantitative* prediction on the effect at different basin values—where the extreme models are designed for perfect nulls effects—we use the estimated meta-study cooperation model $Q(p)$, illustrated in Figure 1(A) and (B) for initial and ongoing rates. As such, the log-likelihood equation across our four treatments is given by:

$$l(\sigma; Q) = l\Big(Q\big(p^{\star}(\sigma, \$9, 2)\big)\Big) + l\Big(Q\big(p^{\star}(\sigma, \$9, 4)\big)\Big) + l\Big(Q\big(p^{\star}(\sigma, \$1, 4)\big)\Big) + l\Big(Q\big(p^{\star}(\sigma, \$1, 10)\big)\Big).$$

This is a single equation in σ, which can be estimated via maximum likelihood, see Figure A.2 in the Online Appendix A for illustration.

5. Extensions

Our analysis so far has abstracted away other features of the coordination problem to focus on the pure effects of the primitives of the strategic game. In this section, we consider four extensions—with relevance both inside and outside the laboratory—that allow us to study possible limitations of the strategic-uncertainty model to predict changes in equilibrium selection.

First, we consider the extent to which beliefs about collusive behavior of others may be distorted by prior experience. While a policy change can alter market primitives and the strategic-uncertainty measure, the underlying variable in this model is belief in others' cooperating. It seems plausible that beliefs may be driven by prior experience before any change in the primitives takes place, and so the model may perform poorly at predicting changes within a population. For example, if a player has engaged extensively with the *same* population of market participants under a status quo and adapted noncollusive behavior, this behavior may be sticky, and therefore unresponsive to shifts in the primitives. Our treatments in the previous section used a *between*-subjects design: identification was based on comparisons of late-session behavior between different populations, each with experience in a fixed environment. Here, in a modified treatment, we examine the effects of varying the number of players N *within* the same population. In this extension, we show that outcomes do not exhibit long-run stickiness, where the between- and within-subject results are largely in line.

In a second set of extensions, we examine the strategic-uncertainty mechanism underlying the basin-size selection device.[45] In particular, we examine the extent to which our results are affected by the possibility of explicit coordination, holding constant X and N. Here, we seek to mirror an empirical finding that when collusion in industries is detected, it is often accompanied by evidence of explicit collusion—despite the illegality of such meetings.[46] We show that once explicit collusion is allowed, neither the independent- nor the correlated-basin measures are good at predicting collusive behavior. Once parties can explicitly collude, we find very high levels of sustained cooperation. This suggests that, indeed, uncertainty over the other strategic choices is a main driver of behavior in our treatments. However, the extremeness of the effect once communication is allowed for raises a question over the extent to which explicit collusion might lead to high cooperation rates even when collusive outcomes are not an equilibrium. To examine this, we show that there are clear limits to what explicit collusion can achieve. In fact, our findings suggest pre-play communication no longer helps at sustaining collusive behavior once it can no longer be theoretically supported in equilibrium.

[45] Strategic uncertainty is a term that captures challenges when a player has to think about the behavior of a human opponent. Free-form communication treatments can reduce strategic uncertainty because players can share information that reveals their strategic intentions. However, our design is not equipped to identify the mechanism through which strategic uncertainty is reduced. It could be that messages convey the opponent is reasonable and understands the tensions of the game. It could be that messages do not directly convey information on rationality but simply reduce social distance. Perhaps, once social distance is reduced, it is easier to trust the other player and, as a result, strategic uncertainty is reduced.

[46] See Marshall and Marx (2012) for a more comprehensive treatment.

Our third extension tests a possible limitation of the basin-of-attraction measure along a different dimension. In our previous treatments we required all N players to cooperate to achieve the efficient outcome. Here, we soften that requirement by allowing for the number of cooperating players to be less than N. At first sight, this less demanding requirement for a cooperative outcome suggests that more cooperation would emerge. However, the basin of attraction makes a prediction in the opposite direction. The reason is that reducing the number of players needed for a first-best outcome introduces a new coordination problem: which two players will be the ones who cooperate. In fact, the new coordination problem introduces so much strategic uncertainty that both measures predict no cooperation. As such, the treatment provides a stark test of the basin-of-attraction notions. Although the cooperation rates we document are not zero, they are significantly lower than in the comparison treatment where $N = 2$ and both players have to cooperate.

The final extension we discuss proposes a procedure to study the validity of the basin-of-attraction measure beyond the laboratory. The section introduces AIAs that are currently used for pricing in several markets. While there are multiple AIAs currently studied in the literature for their potential capacity to collude, we show that a specific AIA algorithm that uses only past experience to learn matches quite well the collusion predictions under the basin-of-attraction measure. In other words, in this extension we show that expected collusion for a given measure of the basin of attraction is comparable between a specific AIA algorithm and what has been documented for humans in the laboratory. We leverage this finding to propose several alternatives to test the equilibrium selection theory beyond our setting and parameterizations.

5.1. **Between vs. Within Identification.** The motivating idea for our first extension is that in many settings of interest the policy-relevant comparative static varies within a population. However, if agents have strong beliefs about others due to previous experiences, it may be that our theoretical construction lacks the ability to predict outcomes as policy changes. If selected equilibria are very sticky within a population, then more-standard assumptions maintaining the equilibria across the counterfactual may have greater validity. For example, if a participant's experience with others is that they play the stage-game defection every period, this belief can persist despite a policy shift that makes collusion easier.

Ideally, we would introduce a primitive change within a supergame; for example, a move from $N = 4$ to $N = 2$, where the matched player after the modification is one of the three matched participants from before. However, in exploring potential designs, we were not satisfied that they would produce clear results. First, it is well documented that repeated-game environments require several supergames of experience for participants to internalize the environment (Dal Bó, 2005). While implementing a surprise change in N—as a mid-supergame manipulation—would mirror an outside-the-laboratory consolidation, it would provide a single supergame observation. An alternative design choice could implement a change in N with some probability within each supergame. However, any observed effects would then be confounded with the expectations over the primitive change (and greater complication in the instructions) and would no longer be comparable to our between-subject treatments.

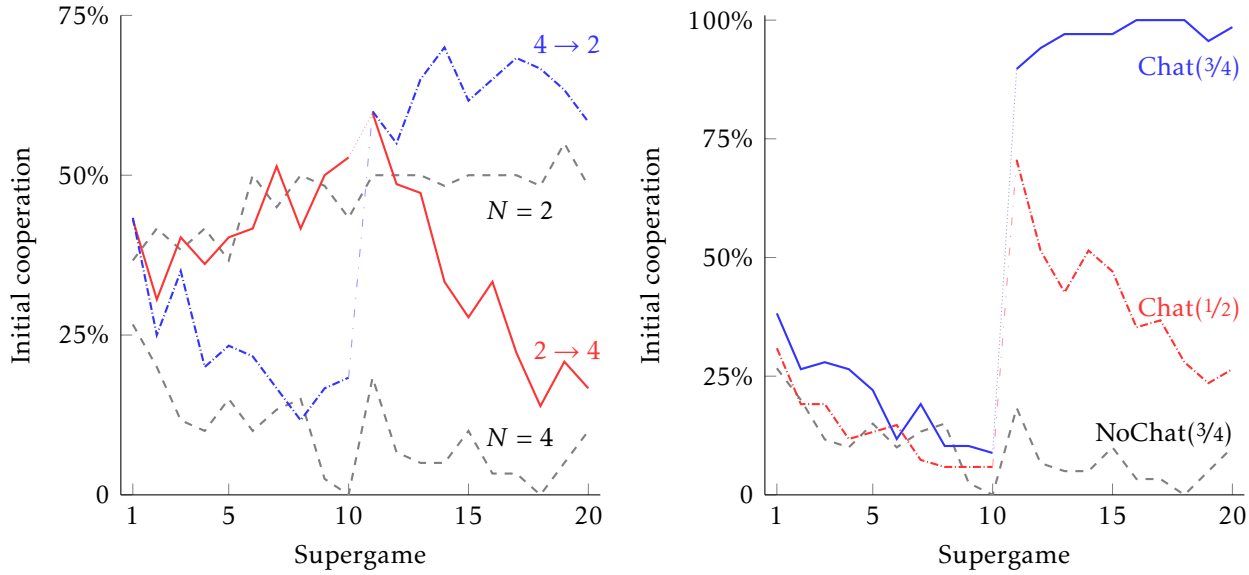
Given these potential confounds, we instead opted for a design with a surprise one-time change in the number of players occurring in a fixed session-level population. Holding constant the cooperation cost at $X = \$9$, we initially set a value of N (either two or four) for the first ten supergames. Then, we changed the value of N for the last ten supergames (to four or two, respectively).

This led to two additional experimental treatments, one with $\binom{N=2}{X=\$9}$ in the first half, and $\binom{N=4}{X=\$9}$ in the second; and the converse treatment from $\binom{N=4}{X=\$9}$ in the first half, to $\binom{N=2}{X=\$9}$ in the second. Given that we hold constant $X = \$9$, for simplicity we label the treatments as $2 \rightarrow 4$ and $4 \rightarrow 2$, for the changes from the first half to the second half. In both treatments, the change in N comes as a surprise: subjects know that there is a second part, but do not receive instructions until supergame ten ends.[47] In terms of the standard strategic-uncertainty model this creates a shift across the session from a low basin size of 0.33 when $N = 2$, and a high basin size of 0.69 when $N = 4$. In particular, this is a change in N that for our between-subject design generated a substantial treatment effect.

In Figure 3(A) we present the initial cooperation rates in each supergame from 1 to 20 averaged across all sessions. The cooperation rates refer to between- and within-subject treatments with $X = \$9$. The between-subject treatments with $N = 2$ and $N = 4$ are indicated by two gray dashed lines (separately labeled), while the within-subject treatments are represented by two colored lines: a solid red line for the $2 \rightarrow 4$ treatment and a dash-dotted blue line for the $4 \rightarrow 2$ treatment. The figure illustrates a substantial between-subject effect, with more cooperation in $N = 2$ over $N = 4$ for the twenty supergames. Further, the figure indicates that our within- and between-session results are identical for the first ten supergames. Pooling the between and within treatments with $N = 2$ in supergames 6–10 the initial cooperation rate is 47.4 percent. In contrast, the pooled cooperation rate for $N = 4$ is just 13.9 percent.[48] As we move into supergames 11–20, the number of players matched in each supergame changes for our within treatments. The vertical lines in Figure 3(A) indicate the immediate shift in behavior as the primitive changes. For the $2 \rightarrow 4$ treatment (the solid red line), initial cooperation levels remain fairly high after changing from $N = 2$ to $N = 4$. In fact, cooperation in the first four-player interaction (supergame 11) actually exhibits an increase to 59.7 percent from the 53.0 percent from the last two-player interaction (supergame 10). However, while there is no immediate drop in cooperation, and thus stickiness, as subjects gain experience the cooperation rate falls rapidly, reaching 16.7 percent by supergame 20. In contrast, moving in the other direction from $N = 4$ to $N = 2$ (the blue dash-dotted line), we find an immediate jump as the primitive shifts. While the initial-round cooperation in supergame 10 with four-players is 18.3 percent, the reduction to $N = 2$ pushes the cooperation rate up to 60.0 percent in supergame 11. The immediate jump in behavior is then sustained across the remaining supergames, with 58.3 percent cooperation in supergame 20.

[47] For our between-subject treatments sessions were also divided into two parts, except that once the subjects reached the second half of the session, they were told that part two is identical to part one.

[48] Testing the initial cooperation rate differences in supergames 6–10 over N (so across the between and within sessions with identical treatment at this point) we find $p = 0.150$ for $N = 2$ and $p = 0.981$ for $N = 4$ from t-tests for a level difference, and $p = 0.353$ for a joint test.



(A) Between vs. within ($X = \$9$)

(B) Explicit vs. implicit $\binom{N=4}{X=\$9}$

FIGURE 3. Initial cooperation rates in extensions (by supergame)

Inspecting the results illustrated in Figure 3(A) it is clear that there is little evidence for the hypothesis that equilibrium selection is sticky in the long run under a within-population shift in N. Despite exposure to the alternative environment in the first half, longer-run behavior in the second half is not dissimilar from that observed in the between-subject design. This is indicated by the close proximity of the two colored/gray line pairs in supergame 20, and the relative distance from the other pair.

In the Online Appendix B, we provide a more detailed like-with-like comparison of the between-subject and within-subject results. These more detailed findings do not indicate differences with the between-subject results as we move from $2 \to 4$. However, in opposition to the hypothesis that the selected equilibrium is sticky, we find a significant *increase* in the response to N over the between analysis as we move from $4 \to 2$.

Overall, we find that:

Result 3 (Between vs. Within)**.** *Switching the identification to within does not substantially change our qualitative results. We find no evidence that the selected equilibrium is sticky in the long run as we shift a primitive within the population. If anything, our within-subject identification shows a larger shift than in the between-subject design as we decrease N.*

5.2. **Explicit Correlation.** The focus so far and for the RPD literature in general has been on an examination of *implicit* coordination. Specifically, our results suggest that a model of strategic uncertainty extended through independence is relatively successful in organizing the data. This is particularly so for ongoing cooperation. On the contrary, a measure based on perfectly correlated play fails to predict the substantial effect of N for both initial and ongoing behavior.

TABLE 4. Cooperation: implicit vs. explicit

	Implicit	Explicit	
	NoChat(3/4)	Chat(3/4)	Chat(1/2)
Initial coop.	0.035 (0.017)	0.988 (0.007)	0.300 (0.037)
Ongoing coop.	0.006 (0.003)	0.806 (0.030)	0.044 (0.018)
Initial success	0.000	0.971	0.094
Ongoing success	0.000	0.756	0.002

Note: All treatments have $X = \$9, N = 4$, where NoChat(3/4) refers to the core 2×2 between-subject design discussed in Section 4. Results are calculated using data from the last-five supergames. Cooperation rates present raw proportions (with subject-clustered standard errors).

These results validate a standard extension of the basin-of-attraction notion. However, correlation across participants' beliefs becomes much more plausible when agents have access to a channel that enables explicit coordination. In an extension to our previous experiments, we now examine the extent to which multiple parties successfully collude when given a coordination device. To this end, we designed two additional 'chat' treatments by modifying an already-studied environment with the greatest coordination on the noncollusive outcome, the $\binom{N=4}{X=\$9}$ treatment.

In our first chat treatment, the first 10 supergames are identical to the $\binom{N=4}{X=\$9}$ treatment, but in supergames 11–20 we introduce pre-supergame chat between the four matched players. The second chat treatment is identical to the first in terms of timing for when the chat coordination device is introduced; the difference between the two treatments is that in the latter we reduce the continuation probability to $\delta' = 1/2$ for the entire session. The designed effect of this change is that while the stage-game payoffs and the number of players remain constant, lowering the value of δ makes the grim-trigger strategy a knife-edge-only SPE, with the critical belief $p^\star(\delta') = 1$ required on the other cooperating under both Equations (2) and (3). The δ' treatment therefore serves as a test for whether explicit coordination can implement outcomes that are not supportable as a robust equilibrium (that is, with arbitrarily small trembles in behavior).

In Figure 3(B) we plot the initial cooperation rates in each supergame averaged across sessions $\left(\text{with the } \binom{N=4}{X=\$9} \text{ treatment provided as a baseline, labeled here as NoChat(3/4)}\right)$. Late-session cooperation and success rates (in supergames 16–20 with subject-clustered standard errors) are provided in Table 4.

Our first chat treatment delivers unequivocal results: providing pre-play communication at $\delta = 3/4$ takes the near-zero cooperation rate in the absence of chat to almost complete cooperation. While ongoing cooperation drops slightly from the very high initial-cooperation levels, the large majority of supergames exhibit coordination by all four participants on the efficient/collusive outcome. Such high levels of cooperation with communication are inconsistent with the predictions of either model of equilibrium selection for these primitives. This suggests that once explicit coordination devices are allowed

for and strategic uncertainty is reduced, the independent model—that captures behavior when collusion is tacit—becomes redundant.

However, at the δ' boundary, even with pre-play communication, participants find it hard to coordinate. While initial cooperation continues to be substantially higher than in the treatment without chat (30.0 percent), ongoing cooperation falls to just 4.4 percent (with an ongoing success rate of only 0.2 percent). The findings indicate that for explicit communication to play a role, there need to be clear incentives for collusion.

Overall, the results of this extension can be summarized as follows:

Result 4 (Implicit vs. Explicit)**.** *Explicit coordination leads to very high cooperation levels with multiple players, in a setting where implicit cooperation achieves near-zero cooperation. However, in the limiting case, where cooperation is a knife-edge SPE outcome, even pre-play chat fails to support cooperation.*

5.3. **Easing Requirements for a Success .** Our previous result illustrates a limitation of both basin measures: if communication is allowed, neither extension predicts behavior well. Here, we explore another extension that, similarly to the one detailed above, may presumably show limitations of either of the two measures in predicting behavior. In our prior games, we generalized the two-player RPD to the N-player setting by requiring that all N players jointly coordinate to obtain the first-best outcome. In this way, we maintained the stark coordination problem in the two-player RPD over two Pareto-ranked outcomes: efficiency with joint-cooperation, and the stage-game Nash outcome. However, once we allow for N players—even holding constant our 2×2 stage-game representation—it is possible to allow for weaker coordination requirements. In this extension, we weaken this requirement by studying an environment with $N = 4$ players and $X = \$9$, but where only two cooperative players are required for everyone to receive a success signal. A single cooperator is enough to provide a success for the other three, but with two cooperators all four participants will get a success. In this way, the cooperative requirements are identical to the standard two-player RPD (two cooperators yield a stable efficient outcome) but the number of participants is larger, $N > 2$.

In this new treatment, we can again follow the same logic for both extensions of the basin-of-attraction measure. While in this new environment efficient cooperation seems *easier* to sustain than in the two-player RPD—requiring just two of four to cooperate, rather than two of two—*both* extensions indicate that this environment is strictly harder. Both the independent and correlated basin extensions suggest that not only will the rate of cooperation be lower here, but so too will the rate of successful cooperation.

Fixing the same grim-trigger versus always-defect thought experiment, we can focus on the event that *exactly one* other player is a conditional cooperator: with more than one, the best-response is always defect (cooperation is costly, and defection does not affect the outcome); with less than one, the best response is again always defect (as any cooperation attempt will fail). An automatic inference then is that if others are perfectly correlated in their strategy choice, choosing grim can never be a best-response, as the number of cooperators will be equal to either zero or three. Consequently, the correlated extension predicts very low cooperation (with $p^\star = 1$). However, even when other participants

choose grim independently with probability p, it is easy to show that there is no $p \in [0,1]$ for which grim is a best response.[49] As such, despite weakening the requirement for cooperation relative to the standard two-player RPD, the prediction of the basin-based extensions is that cooperation attempts will fail.

The reason for this non-intuitive prediction is that in this alternative setting, strategic uncertainty is substantially increased. Not only is there the same uncertainty from before over whether the group will succeed, there is also uncertainty over which group-members, if any, will free-ride on the cooperation of others. The basin calculations indicate that the effect of increased strategic uncertainty dominates the weaker requirement for success. To examine this prediction, we conducted three sessions of this treatment (with 64 unique participants recruited), using $X = \$9$ but where only two of the $N = 4$ group members needed to cooperate for every player in the group to receive a success signal.

Before describing the results of the new treatment, we review the comparable behaviors for the baseline two-player RPD. Taking averages at the partnership level, across 480 late-session rounds in $\binom{N=2}{X=\$9}$, we find an overall cooperation rate of 44.6 percent, with a 36.0 percent rate of group-wide success (here both players cooperating in the round). Having fixed the temptation parameter at $X = \$9$, our new treatment has the weakened requirement of two or more cooperators from the four matched players for a group-wide success. In the data from the new treatment, we find a cooperation rate of 22.7 percent (192 rounds) with a group-wide success rate of 25.5 percent.[50,51] Per the prediction from the basin, the results indicate significantly reduced levels of coordination in the game ($p = 0.006$).

Despite making it mechanically easier to generate a success than in the two-player RPD, the results mirror the directional prediction from the basin calculation. Furthermore, at the full basin size of one, the unilateral cooperation rates in the treatment are in line with the quantitative predictions of the basin. To see this, in Figure 2 we illustrate the treatment's initial and ongoing cooperation rates as empty circles.

In summary, we find that:

Result 5 (Easing Requirements for a Success)**.** *In a treatment where the set of players needed for a successful outcome is lower than the group size, the basin-of-attraction extension predicts a reduction in coordination due to an increase in strategic uncertainty (here distributive). The*

[49] It is sufficient to show that the difference in payoff between grim and always defect is negative at $p = \frac{1}{(N-2)(1+x/9)+1}$, which is the best-case for grim (the strategy payoff difference is single peaked and uniquely maximized here).

[50] To make fair comparisons, our analysis here focuses on the group-level analysis, in particular the group-wide success outcome. However, we note that cooperation at the individual level is significantly lower in the new treatment ($p<0.001$ all comparisons).

[51] An alternative comparison here is to the $\binom{N=4}{X=\$9}$ treatment. However, the treatment results are already at a hard boundary (a late-session cooperation rate of 1.4 percent, with a zero percent rate of group-wide success). Given this, the fact that we find greater success with the weakened threshold is not surprising.

treatment results indicate low cooperation rates in line with empirical rates observed for extreme basin-value in other RPD experiments. In terms of successful coordination, the effect from weakening the coordination requirements matches the basin prediction, with a significant decrease in successful coordination.

5.4. **Moving Beyond the Laboratory .** Data to evaluate selection criteria such as the basin-of-attraction have thus far come from the behavior of human subjects in the laboratory, with treatment parameters often selected to study other hypotheses. While meta-studies have brought some of this together, the ability to study a wider set of parameters across many other environments is necessary to pin down which measures are most predictive and the domains they can cover. However, experimental methods are often best-placed to examine relatively coarse hypotheses, across a sparse set of parameters. As such, it is particularly useful to find empirically driven methods that might supplement and target experiments for maximum inference.

To that end, in this last extension we propose one option as a potential guide for future exploration. Our approach here is based on an emerging literature in industrial organizations that examines the pricing behavior of AIAs. Calvano, Calzolari, Denicolo, and Pastorello (2020) outline how a commonly used AI-learning algorithm (Q-learning, Watkins, 1989) is capable of maintaining supracompetitive prices in a standard dynamic Bertrand environment with implicit coordination, via standard dynamic strategies. In a response to this, Asker, Fershtman and Pakes (2021, 2022) outline how this result is sensitive to the form of the algorithm, showing that supracompetitive prices rely on the extent to which AIAs can learn counterfactually from what would have happend with alternative choices (termed *synchronous* learning), as opposed to learning solely from the on-the-path experiences (termed *asynchronous* learning).

Given the growing interest in AIAs as pricing agents, and the potential for collusion to emerge, there is a natural connection to the question we are asking in this paper. Specifically, in this section we examine how the steady-state behavior of the Q-learning AIAs is related to the behavior of lab subjects in our in our repeated N-player dilemma, and how both are predicted by the basin of attraction. In particular, we will demonstrate a strong parallelism between our laboratory results and the results from experimental simulations using AIAs, with extensive variation across x, N, and δ.

In total we simulate over 1.8 million games with an identical structure to our experimental N-player RPD environment, but here using two-state AIAs as the decision makers.[52] In our simulations we vary: (i) the number of players, $N = \{2, 3, \ldots, 10\}$; (ii) the discount factor, $\delta = \{0.75, 0.90,\ 0.95,\ 0.99\}$; the always-defect basin size ($p^\star(x, N, \delta) = \{0, 0.01, 0.03,$ $0.05, \ldots, 0.99\}$, chosen by varying x; and (iv) the algorithm learning mode. In the asynchronous learning mode the AIAs learn solely from the payoffs observed from their chosen decisions, whereas in the synchronous mode we allow the AIA to learn both from

[52]With two internal states the AIA decision makers have access to a conditioning variable that could be used to construct a history-dependent strategy such as the grim trigger. However, the way the algorithm makes use of this state variable is entirely endogenous, determined by the particular learning path.

the path and the counterfactual.[53] For each treatment environment/algorithm we simulate 1,000 distinct repeated games, where each simulated game runs for 10,000 rounds (this was a sufficient length to obtain convergent behavior for all treatments/algorithm modes). Our final measures from each simulation are the ongoing cooperation rate among the AIAs, where initial behavior is entirely random, driven by an initially diffuse uniform distribution over the action choice weights for each state.[54]

We present the results of our AIA simulations in Figure 4 as the triangular points, with the asynchronous results in panel (A) and the synchronous results in panel (B). Each triangle represents the average long-run cooperation rate across our AIAs at that value of $p^\star$, pooling our treatments across N and δ.[55] As such, each point represents an average across 18,000 AIA supergames. Behind each set of results for the algorithms we illustrate the fitted relationship between the basin and ongoing cooperation from the RPD meta-study using human subjects, cf. Figure 1(B). Figure 4(A) makes clear a top-level observation that the results from the asynchronous algorithm display behavior that is highly consistent with the predictions of the independent extension: collusion decreases as $p^\star$ increases and essentially disappears once $p^\star > 0.5$. In the region with $p^\star < 0.5$, asynchronous AIAs broadly mirror the behavior of subjects in the laboratory.[56] For values of $p^\star$>0.5, the asynchronous AIAs cooperate less than humans, although the difference is not large. In contrast, for the more sophisticated synchronous algorithm shown in Panel (B), we observe much larger differences in behavior between the AIAs and humans. Mirroring the results from Asker, Fershtman, and Pakes (2022), the synchronous algorithm is much less successful at colluding, only doing so at very low values of the basin.

We conclude that:

Result 6 (Exploration of AIAs Behavior Relative to Humans)**.** *Asynchronous AIAs that learn only from past experiences on the path display collusion behavior that is consistent with the prediction of the independent extension of the size of always-defect basin and track the behavior of human subjects quite closely. On the contrary, there are large differences between the behavior generated by sophisticated synchronous algorithm that also learns counterfactually and the behavior of humans.*

Our results here suggest that Q-learning algorithms can be predictive of human behavior in these repeated settings. Future research can explore and leverage this link, where

[53] We thank John Asker for sharing MATLAB code, which we re-implemented in *Python*.

[54] In general, we follow Asker, Fershtman, and Pakes (2021) in this setup, with the only substantial change being the switch from a dynamic Bertrand environment they study with many price actions, to the two-action environment studied in our laboratory treatments.

[55] The AIAs we study require a substantial degree of training to converge. For this reason, we examine the long-run, convergent behavior of Q-learning AIAs within our simple N-player social dilemma environment. The ongoing cooperation rates that we report correspond to the convergent behavior that AIAs achieve for a given parameterization.

[56] For higher N, and lower $p^\star$ the data does exhibit a non-monotonicity for $\delta = 0.75$. which appears in the graph as the flat region close to a zero basin. The reason for this is that at very low values of x ($<< 10^-5$), the asynchronous AIAs have a hard time learning the relevant punishment strategies to support cooperation, where the state is used instead for serial alternation between cooperation and non-cooperation.

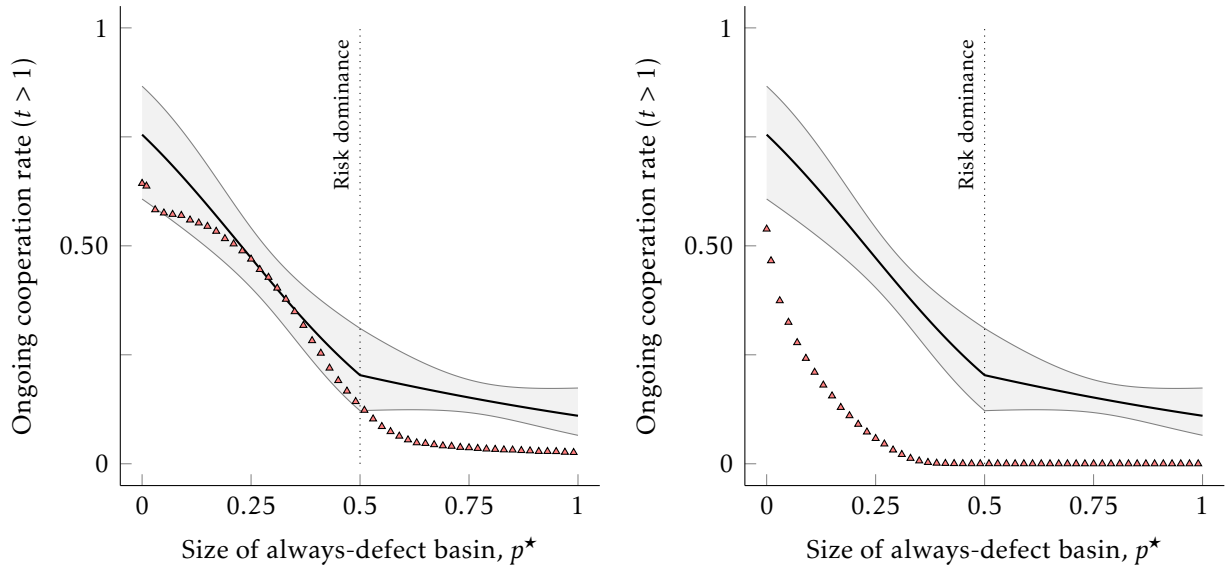
we now outline some of the possible ways this can be accomplished. First, the exercise suggests ways in which AIAs can complement the laboratory. For example, in a standard RPD environment, the asynchronous algorithm can be used to predict behavior for sets of parameters for which there is no/very little experimental data. So long as one can show some parallels between AIAs and human-subject behavior, testing the connections between behavior of AIAs and humans, then AIAs might be used for thought experiments or exploring the extremes of the parameter space.

Greater exploration of the parameter space may then help fine-tune empirical selection criteria, even in settings for which there exists substantial data. For example, some facets of the AIAs behavior may not be not fully captured by the summary basin $p^\star$. As an example, in two-player PD games with high temptations and low sucker payoffs, AIAs begin to exhibit serial alternation across the (C,D) and (D,C) actions well before this behavior becomes efficient ($1+g-s>=2$). This prediction from the AIAs can then be examined in the laboratory. The data from such experiments could clear up whether the predicted discrepancies were exclusive to AIAs, or whether they are shared by humans, suggesting a need for a correction to the selection criterion at these regions of the parameter space.

Finally, AIAs can be used to explore behavior and shape selection theory in extension environments that differ from the standard RPD or our N-player extension. For example, with AIAs it is relatively simple: (i) to expand the action set (as in the Bertrand/Cournot setting); (ii) to allow for state variables that evolve with the game (stochastic/dynamic games); (iii) to allow for imperfect monitoring (á la Green and Porter, 1984); or (iv) to study features that reduce strategic uncertainty such as sequential moves or explicit communication between the AIAs. Naturally, studying whether empirical selection criteria such as the basin of attraction for always defect work in these other settings are outside of the scope of this single paper. However, we suspect that AIAs will be a key aide for future explorations of these selections questions within experimental contexts. Moreover, the increasing interest in AIAs will mean that studying their behavior will have increasing external validity.

6. Conclusion

Our paper examines equilibrium selection in repeated games and the extent to which it can be predicted with a model of strategic uncertainty. We leverage a model of equilibrium selection that rationalizes behavior in the two-player RPD and design an experiment to stress test this specific theoretical model. The predictive model works by mediating the effects from multiple primitives into a single dimension, strategic uncertainty. As such, even for rich counterfactual policies with many changes to the setting, the model can still generate a directional prediction. We introduce a novel source of strategic uncertainty that has not yet been studied in the RPD setting (the number of players), while also manipulating a payoff parameter. Therefore, we can change both sources of strategic uncertainty simultaneously and study the extent to which the evidence is consistent



(A) Asynchronous

(B) Synchronous

FIGURE 4. Cooperation behavior of two different AIAs

with the predictions of the selection model. Our main finding is that the model of equilibrium selection can indeed be used as a device to understand successful ongoing coordination on the collusive outcome.[57] In particular, the model performs well in trading off the competing effects from the two distinct sources of strategic uncertainty.

After illustrating the theoretical power of the model for implicit coordination, we turn to several application-motivated extensions that probe the model's limitations. Our first extension is motivated by the extent to which prior history and experience could make the equilibria "sticky," even when the model suggests a change. To do this, we study the extent to which our findings still hold when treatment-variable manipulations take place within the same population, as opposed to between populations in our main study. The results indicate that even when participants experience a treatment-variable shift, the model continues to predict the longer-run outcomes. While we find some hysteresis in the short-run responses to a within-population policy change—initial stickiness in behavior in one direction and a large immediate response to the change in the other—behavior after accumulating experience under the new parameters is not distinct from that observed in the between-population treatments. As such, these results suggest that the model prediction fares better than the more-standard assumption of maintaining a selected equilibrium across a policy change.

In a second extension, we examine the potential effects from explicit collusion. Our main finding here is that once we allow for explicit coordination, by providing an explicit coordination device (pre-play chat), the selection-model prediction differs considerably from

[57] In applications outside of the laboratory that may want to leverage this finding (that theoretical prediction measures fare better when checked against ongoing measures of collusion), it would be useful to establish that the behavior in the game is either converging or at least not displaying large changes.

observed behavior. In fact, where the model suggests very low levels of cooperation, the observed behavior is highly cooperative, once coordination devices are present. The evidence from this extension indicates both that strategic uncertainty plays a clear role, but also that the selection model based upon it is entirely inappropriate for predicting behavior when explicit collusion is suspected.

Thus far, in our main treatments (and the previous extensions) we required all N players to cooperate to reach a cooperative outcome. Our third extension eases this requirement so that only a subset of players needs to cooperate. One may think that this relaxation would increase cooperation rates as reaching a first-best outcome is now less restrictive. However, the measures based on strategic uncertainty make a prediction in the opposite direction. The reason is that the softer requirements add strategic uncertainty with respect to which players will cooperate and which will free-ride on others cooperating. Our results are consistent with the strategic-uncertainty predictions and suggest lower rates of cooperation. This implies that in settings in which free-riding is possible (for example, an indefinitely repeated voluntary contribution game), strategic uncertainty increases and other things constant it is more difficult to cooperate. Our extension does not directly test such an environment but can be taken up in subsequent work.

Our final extension highlights a limitation of the experimental approach and lays out a possible path for future research. We show that some AIAs can be used to predict behavior in experimental settings beyond the standard RPD. We document that the collusive behavior of some AIAs is consistent with the strategic-uncertainty measure and that it closely tracks the behavior of humans. While these exercises are preliminary, we outline how future research can leverage AIAs to design better tests in commonly-used laboratory settings, and how to evaluate the adequacy of empirical selection indices beyond the simplest environments.

References

Agranov, Marina, Guillaume Frechette, Thomas Palfrey, and Emanuel Vespa (2016), "Static and dynamic underinvestment: An experimental investigation." *Journal of Public Economics*, 143, 125–141.

Aoyagi, Masaki, V Bhaskar, and Guillaume R Fréchette (2019), "The impact of monitoring in infinitely repeated games: Perfect, public, and private." *American Economic Journal: Microeconomics*, 11, 1–43.

Asker, John, Chaim Fershtman, and Ariel Pakes (2021), "Artificial intelligence and pricing: The impact of algorithm design." *National Bureau of Economic Research*.

Asker, John, Chaim Fershtman, and Ariel Pakes (2022), "Artificial intelligence, algorithm design and pricing." *AEA Papers and Proceedings*, 112, 452–56.

Battaglini, Marco, Salvatore Nunnari, and Thomas R Palfrey (2012), "Legislative bargaining and the dynamics of public investment." *American Political Science Review*, 106, 407–429.

Battaglini, Marco, Salvatore Nunnari, and Thomas R Palfrey (2016), "The dynamic free rider problem: A laboratory study." *American Economic Journal: Microeconomics*, 8, 268–308.

Battalio, Raymond, Larry Samuelson, and John Van Huyck (2001), "Optimization incentives and coordination failure in laboratory stag hunt games." *Econometrica*, 69, 749–764.

Berry, James, Lucas C Coffman, Douglas Hanley, Rania Gihleb, and Alistair J Wilson (2017), "Assessing the rate of replication in economics." *American Economic Review*, 107, 27–31.

Blonski, Matthias, Peter Ockenfels, and Giancarlo Spagnolo (2011), "Equilibrium selection in the repeated prisoner's dilemma: Axiomatic approach and experimental evidence." *American Economic Journal: Microeconomics*, 3, 164–92.

Blonski, Matthias and Giancarlo Spagnolo (2015), "Prisoners' other dilemma." *International Journal of Game Theory*, 44, 61–81.

Brandts, Jordi and David J Cooper (2006), "A change would do you good.... an experimental study on how to overcome coordination failure in organizations." *American Economic Review*, 96, 669–693.

Calvano, Emilio, Giacomo Calzolari, Vincenzo Denicolo, and Sergio Pastorello (2020), "Artificial intelligence, algorithmic pricing, and collusion." *American Economic Review*, 110, 3267–97.

Cason, Timothy N. (2008), "Price signaling and 'cheap talk' in laboratory posted offer markets." *Handbook of Experimental Economics Results*, 1, 164–169.

Cason, Timothy N., Tridib Sharma, and Radovan Vadovič (2020), "Correlated beliefs: Predicting outcomes in 2×2 games." *Games & Economic Behavior*, 122, 256–276.

Cooper, David J. and Kai-Uwe Kühn (2014), "Communication, renegotiation, and the scope for collusion." *American Economic Journal: Microeconomics*, 6, 247–78.

Dal Bó, Pedro (2005), "Cooperation under the shadow of the future: Experimental evidence from infinitely repeated games." *American Economic Review*, 95, 1591–1604.

Dal Bó, Pedro and Guillaume Fréchette (2007), "The evolution of cooperation in infinitely repeated games: Experimental evidence." *American Economic Review*, 50, 54.

Dal Bó, Pedro and Guillaume R Fréchette (2011), "The evolution of cooperation in infinitely repeated games: Experimental evidence." *American Economic Review*, 101, 411–429.

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

Dal Bó, Pedro and Guillaume R Fréchette (2019), "Strategy choice in the infinitely repeated prisoner's dilemma." *American Economic Review*, 109, 3929–52.

Davis, Douglas D (2002), "Strategic interactions, market information and predicting the effects of mergers in differentiated product markets." *International Journal of Industrial Organization*, 20, 1277–1312.

Devetag, Giovanna and Andreas Ortmann (2007), "When and why? A critical survey on coordination failure in the laboratory." *Experimental economics*, 10, 331–344.

Dufwenberg, Martin and Uri Gneezy (2000), "Price competition and market concentration: an experimental study." *International Journal of Industrial Organization*, 18, 7–22.

Embrey, Matthew, Guillaume Fréchette, and Ennio Stacchetti (2013), "An experimental study of imperfect public monitoring: Efficiency versus renegotiation-proofness."

Fonseca, Miguel A and Hans-Theo Normann (2012), "Explicit vs. tacit collusion—the impact of communication in oligopoly experiments." *European Economic Review*, 56, 1759–1772.

Fudenberg, D., D.G. Rand, and A. Dreber (2010), "Slow to anger and fast to forgive: Cooperation in an uncertain world." *American Economic Review*.

Ghidoni, Riccardo and Sigrid Suetens (2022), "The effect of sequentiality on cooperation in repeated games." *American Economic Journal: Microeconomics*.

Goette, Lorenz and Armin Schmutzler (2009), "Merger policy: What can we learn from competition policy." *Experiments and Competition Policy; Hinloopen, J., Normann, HT, Eds*, 185–216.

Green, E.J. and R.H. Porter (1984), "Noncooperative collusion under imperfect price information." *Econometrica*, 87–100.

Harrington, Joseph E, Roberto Hernan Gonzalez, and Praveen Kujal (2013), "The relative efficacy of price announcements and express communication for collusion: Experimental findings." *Working paper. University of Pennsylvania, The Wharton School*.

Harrington, Joseph E, Roberto Hernan Gonzalez, and Praveen Kujal (2016), "The relative efficacy of price announcements and express communication for collusion: Experimental findings." *Journal of Economic Behavior & Organization*, 128, 251–264.

Harsanyi, John C and Reinhard Selten (1988), *A general theory of equilibrium selection in games*. MIT Press, Cambridge, MA.

Horstmann, Niklas, Jan Krämer, and Daniel Schnurr (2018), "Number effects and tacit collusion in experimental oligopolies." *Journal of Industrial Economics*, 66, 650–700.

Huck, Steffen, Kai A. Konrad, Wieland Müller, and Hans-Theo Normann (2007), "The merger paradox and why aspiration levels let it fail in the laboratory." *Economic Journal*, 117, 1073–1095.

Huck, Steffen, Hans-Theo Normann, and Jörg Oechssler (2004), "Two are few and four are many: Number effects in experimental oligopolies." *Journal of Economic Behavior & Organization*, 53, 435–446.

Kartal, Melis and Wieland Müller (2018), "A new approach to the analysis of cooperation under the shadow of the future: Theory and experimental evidence." University of Vienna working paper.

Kartal, Melis, Wieland Müller, and James Tremewan (2017), "Building trust: The costs and benefits of gradualism." University of Vienna working paper.

Kloosterman, Andrew (2019), "Cooperation in stochastic games: A prisoner's dilemma experiment." *Experimental Economics*, 1–21.

Lugovskyy, Volodymyr, Daniela Puzzello, Andrea Sorensen, James Walker, and Arlington Williams (2017), "An experimental study of finitely and infinitely repeated linear public goods games." *Games & Economic Behavior*, 102, 286–302.

Marshall, Robert C and Leslie M Marx (2012), *The economics of collusion: Cartels and bidding rings*. MIT Press.

Potters, Jan and Sigrid Suetens (2013), "Oligopoly experiments in the current millennium." *Journal of Economic Surveys*, 27, 439–460.

Rosokha, Yaroslav and Chen Wei (2020), "Cooperation in queueing systems." *Working Paper*.

Salz, Tobias and Emanuel Vespa (2020), "Estimating dynamic games of oligopolistic competition: An experimental investigation." *RAND Journal of Economics*, 51, 447–469.

Sherstyuk, Katerina, Nori Tarui, and Tatsuyoshi Saijo (2013), "Payment schemes in infinite-horizon experimental games." *Experimental Economics*, 16, 125–153.

Vespa, Emanuel (2020), "An experimental investigation of cooperation in the dynamic common pool game." *International Economic Review*, 61, 417–440.

Vespa, Emanuel and Alistair J Wilson (2019), "Experimenting with the transition rule in dynamic games." *Quantitative Economics*, 10, 1825–1849.

Vespa, Emanuel and Alistair J Wilson (2020), "Experimenting with equilibrium selection in dynamic games." *Working Paper*.

Vesterlund, Lise (2016), "Using experimental methods to understand why and how we give to charity." *Handbook of Experimental Economics*, 2, 91–151.

Watkins, Christopher (1989), *Learning from delayed rewards*. Ph.D. thesis, Cambridge, United Kingdom.

Weber, Roberto A (2006), "Managing growth to achieve efficient coordination in large groups." *American Economic Review*, 96, 114–126.

Wilson, Alistair J. and Emanuel Vespa (2020), "Information transmission under the shadow of the future: An experiment." *American Economic Journal: Microeconomics*, 12.

Appendix A. Additional Tables and Figures

Table A.1. Cooperation rates across all supergames

Cooperation rates	$X = \$9$		$X = \$1$	
	$N = 2$	$N = 4$	$N = 4$	$N = 10$
Initial coop.	0.466 (0.046)	0.100 (0.021)	0.719 (0.039)	0.457 (0.044)
Ongoing coop.	0.296 (0.029)	0.044 (0.012)	0.433 (0.034)	0.243 (0.039)
Initial success	0.466	0.003	0.408	0.010
Ongoing success	0.296	0.002	0.275	0.009

Note: Results are calculated using data from all supergames. Cooperation rates present raw proportions (with subject-clustered standard errors).

Table A.2. Cooperation in reaction to previous round's history

	$X = \$9$		$X = \$1$		Chat ($X = \$9, N = 4$)	
History	$N = 2$	$N = 4$	$N = 4$	$N = 10$	$\delta = 3/4$	$\delta = 1/2$
(C, S)	0.977 (0.011)	–	0.988 (0.013)	–	0.980 (0.006)	0.750 (0.217)
(C, F)	0.317 (0.063)	0.000	0.521 (0.085)	0.739 (0.077)	0.342 (0.0073)	0.255 (0.104)
(D, S)	0.150 (0.060)	–	0.263 (0.110)	–	0.143 (0.136)	0.750 (0.217)
(D, F)	0.033 (0.006)	0.006 (0.004)	0.023 (0.009)	0.025 (0.009)	0.019 (0.019)	0.006 (0.004)

Note: Data are taken from the last five supergames in each treatment (with subject-clustered standard errors). Cells marked "–" have no observations at the relevant history. History shows the own-action-signal pair from the previous round, (a_{t-1}, σ_{t-1}).

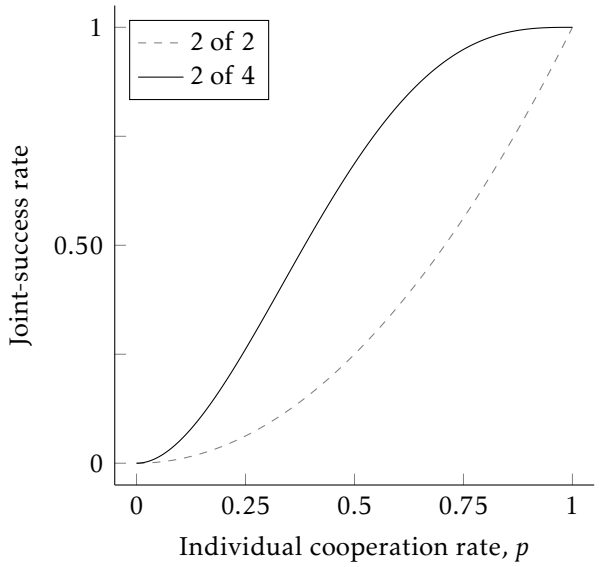
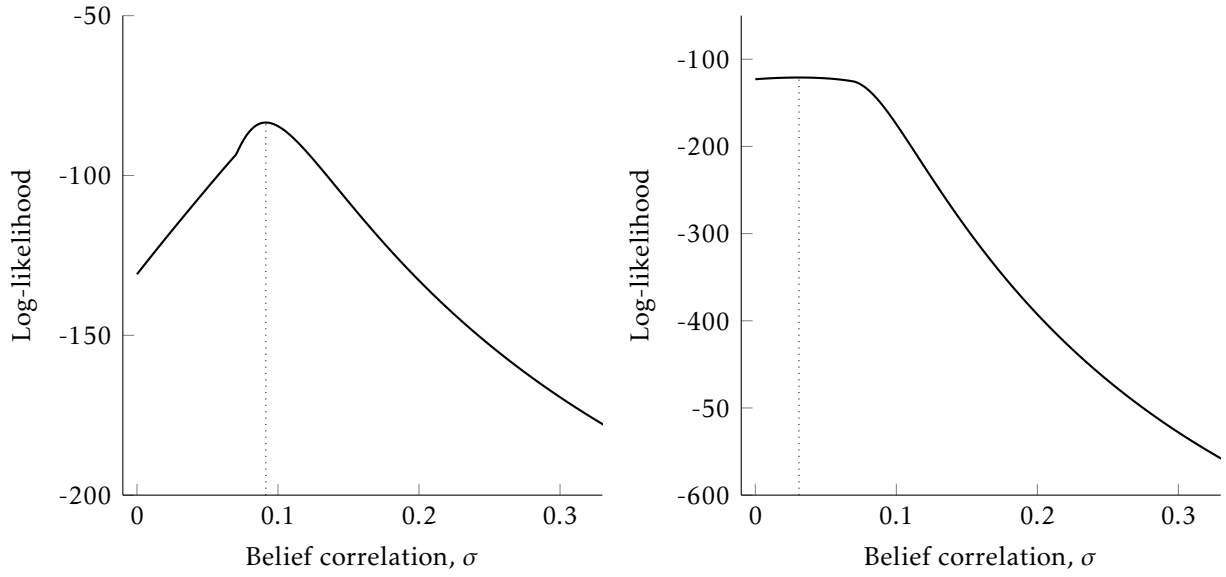


FIGURE A.1. Success rate as a function of individual cooperation rate

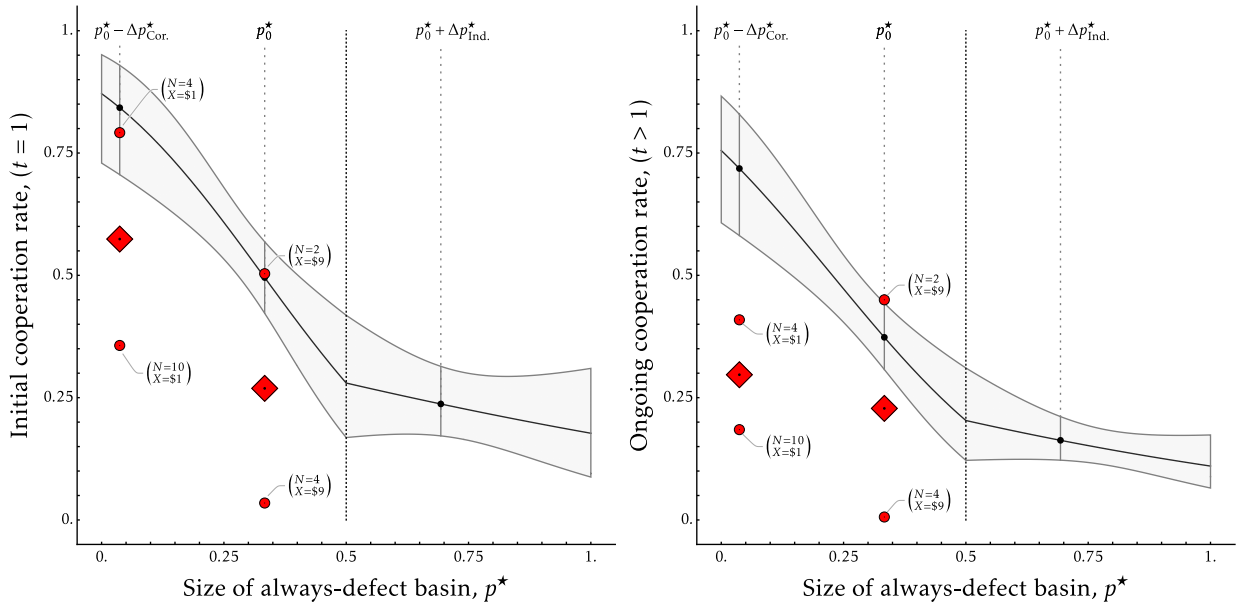


(A) Initial cooperation

(B) Ongoing cooperation

FIGURE A.2. Belief correlation

Note: Data are taken from the last five supergames (16–20) in our four main between-subject treatments. Log-likelihoods are calculated using the imputed cooperation rate from the Dal Bó and Fréchette (2018) meta-study with belief correlation σ calculated as a σ proportion of independent beliefs and a $(1-\sigma)$ proportion of perfectly correlated beliefs.

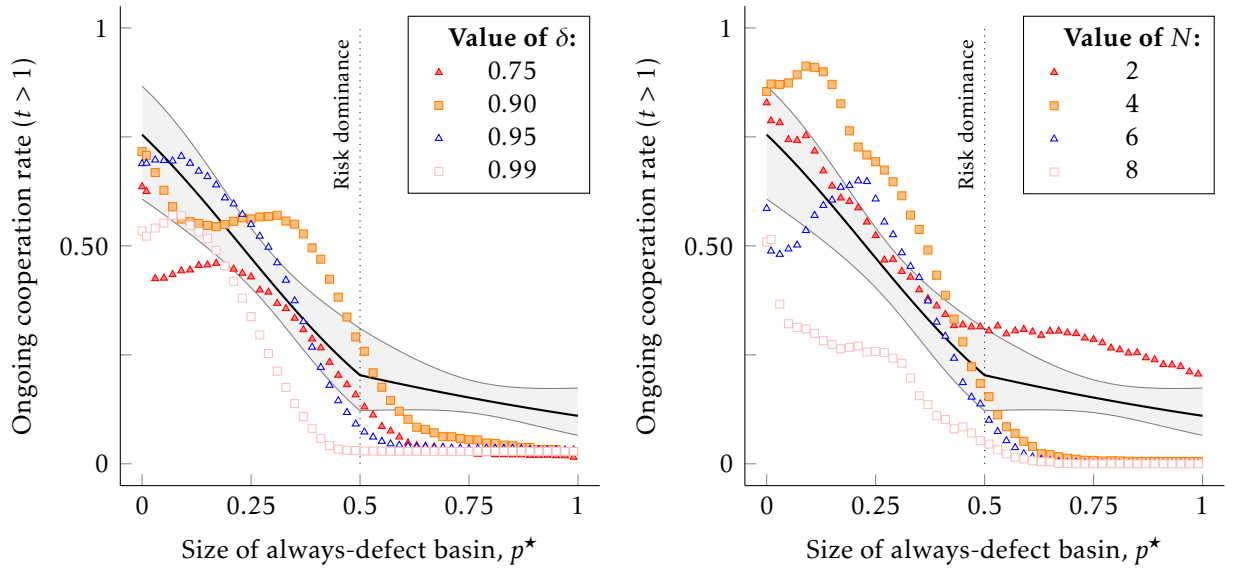


(A) Initial cooperation

(B) Ongoing cooperation

FIGURE A.3. Cooperation and the correlated basin-size model

Note: Circles indicate separate treatments and diamonds treatments pooled over each value of the correlated-basin measure. See Figure 2 in the main paper for the analogous figure under the independent basin.



(A) Varying discount factor δ

(B) Varying number of players N

FIGURE A.4. Cooperation behavior of asynchronous AIAs

Appendix B. Further Analysis of the Within-Subject Treatments

In the within-subject treatments we find evidence of hysteresis. Specifically, we observe a large and immediate jump in cooperation as N changes from $N = 4$ to $N = 2$, and no initial response as N moves in the opposite direction. This suggests that in the short run, subjects respond to a change in the environment with a strong intent to cooperate: cooperation rates of the within-treatment subjects in supergame 11 are significantly greater than the initial cooperation rates in supergame one.

Further examination of Figure 3(A) suggests greatly similar patterns in behavior of within-treatment subjects in the first ten and the last ten supergames. With $N = 4$ cooperation is initially high, but then falls rapidly as the subjects gain experience. With $N = 2$ the cooperation trends upward with each additional supergame. However, closer look at the levels of cooperation highlights differences between the two treatments. With no experience at all in the environment, 43.2 percent of subjects cooperate for $N = 2$ in the first round of the first supergame, compared to 29.4 percent for $N = 4$ (significantly different with $p = 0.005$ from a test of proportions). In contrast to the significant difference over N in the very first decision, in supergame 11 of our within-subject sessions (with prior experience at an alternate value of N) the initial cooperation rates at $N = 2$ and $N = 4$ are indistinguishable from one another in the plotted figure (at 60.0 and 59.7 percent cooperation, respectively), let alone statistically ($p = 0.974$).[B.1] We conclude that experience at another parameter value in the first half causes both treatments' cooperation rates to increase.

In Table B.1 we provide additional details, where we compare behavior of within- and between-treatment subjects after five rounds of experience. In the first two columns we present average behavior of between-treatment subjects (initial/ongoing cooperation and success, with subject-clustered standard errors for the individual choices) in supergames 6–10 for $N = 2$, $N = 4$. In the next two columns we present average behavior of within-treatment subjects in supergames 16–20. Examining the differences across the *within* and *between* cooperation levels, we find (i) no statistically significant differences in behavior for $N = 4$ ($p = 0.117/p = 0.539$ for initial/ongoing cooperation), (ii) statistically significant differences across the $N = 2$ cooperation rates ($p = 0.011$ for initial, $p < 0.001$ for ongoing). The significant differences reflect the substantially greater upward shift in the $4 \rightarrow 2$ treatment.

In the last three columns, we compute (for three different cases) the differences in average behavior between treatments with $N = 4$ and $N = 2$. In column Δ_{Btwn} we calculate the between-subject change using data from the $X = \$9$ treatment in supergames 16–20.[B.2] In the last two columns we present the within-subject change using data from the $2 \rightarrow 4$ and $4 \rightarrow 2$ treatments in supergames 6–10 and 16–20. While the three measures agree qualitatively—and exhibit economically large effects in N in the same direction—there

[B.1] Given the disjoint subject groups and identical treatment in supergames 1–10, we compare proportions using t-tests without clustering. We then compare the initial response under each value of N in the within-subject supergame eleven to all subjects at that N in supergame one. Using these tests, we reject equivalence with $p = 0.021$ for $N = 2$ and $p < 0.001$ for $N = 4$.

[B.2] These results are analogous to the marginal effects attributable to an increase in the independent basin of $\Delta p^{\star}_{\text{Ind.}} = +0.36$ in Table 3 once we remove the $X = \$1$ treatments.

TABLE B.1. Cooperation: between vs. within

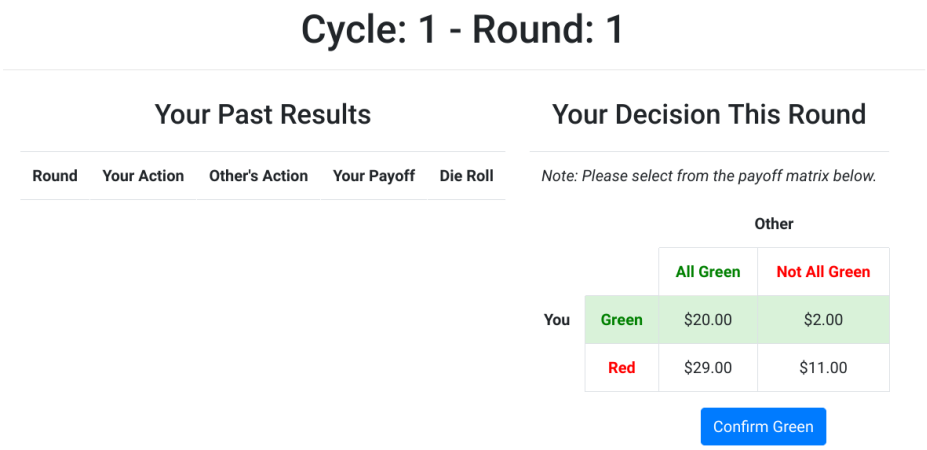
	Between (SG 6–10)		Within (SG 16–20)		$\Delta_{\text{Btwn.}}$	$\Delta_{\text{Wthn.}}$	
	$N = 2$	$N = 4$	$N = 2$	$N = 4$		$2 \to 4$	$4 \to 2$
Initial coop.	0.474	0.139	0.643	0.214	−0.469	−0.260	−0.504
	(0.036)	(0.025)	(0.056)	(0.041)	(0.060)	(0.042)	(0.056)
Ongoing coop.	0.299	0.054	0.598	0.042	−0.444	−0.258	−0.544
	(0.026)	(0.012)	(0.051)	(0.016)	(0.055)	(0.029)	(0.050)
Initial success	0.474	0.011	0.643	0.042	-0.503	-0.433	-0.632
Ongoing success	0.299	0.004	0.598	0.008	-0.450	-0.292	-0.594

Note: Comparisons at the same experience level are generated using supergames 6–10 across all sessions (fixing N, between and within sessions are identical until supergame 11). For the within change we measure the cooperation rates in supergames 16–20. All cooperation rates are raw proportions (with subject-clustered standard errors). The last three columns measure the corresponding cooperation rate when $N = 4$ minus the cooperation rate when $N = 2$.

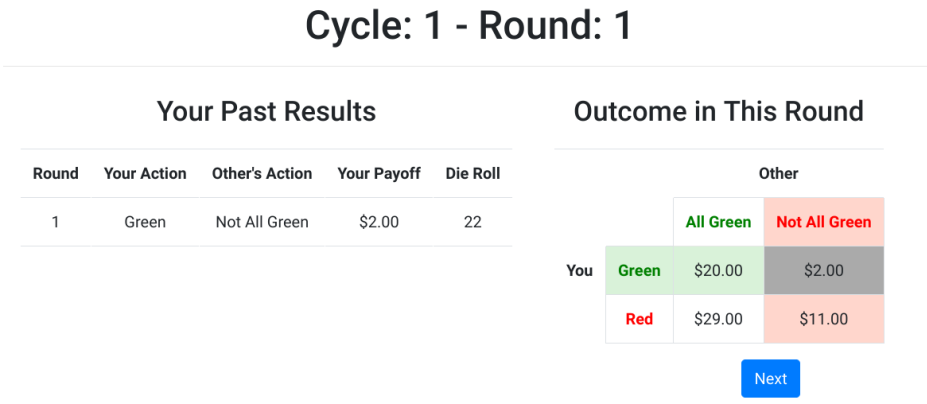
are differences, particularly in the comparisons to the $2 \to 4$ case. However, we note that there are two effects at play here. In the $2 \to 4$ comparison, reduced magnitudes are driven primarily by the fact that behavior in this treatment has not converged. To see this, consider the assessed between-subject effect if we used data from supergames 6–10: a -33.5 percentage point effect on initial cooperation, which is not significantly different from the -26.0 percent effect identified in the within comparison ($p = 0.117$).[B.3] In contrast, the greater assessed effect in the $4 \to 2$ comparison is the composite of the same *reduction* in the effect from looking at the still-converging data for $N = 4$, with a substantial increase in cooperation at $N = 2$ in the second half over the between-subject levels.

[B.3] Similarly for ongoing cooperation the between-effect assessed in supergames 6–10 is -24.6 percent compared to -25.8 percent within ($p = 0.539$).

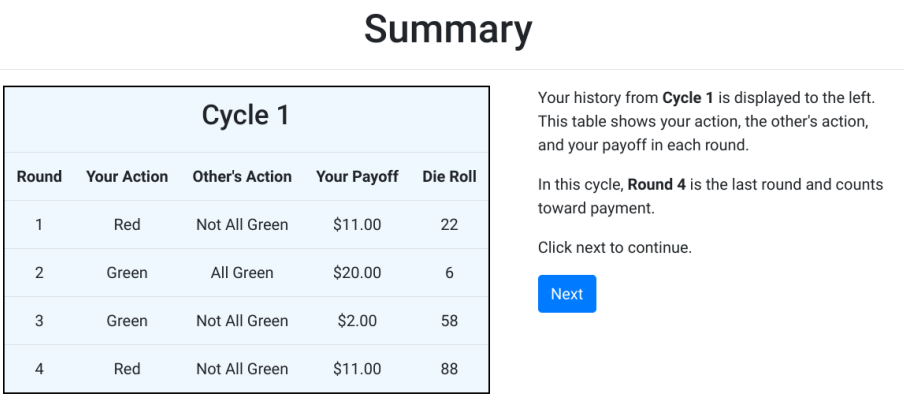
Appendix C. Interface Screenshots



(A) Action Selection



(B) Round Feedback



(C) Supergame Feedback

Figure C.1. Interface screenshots

Appendix D. Provided Instructions

Below we include the instructions given to participants. All language deltas/treatment-specific language are included in braces. Text in red is for the $N = 2$ treatment, in blue for the $N > 2$ treatments (here we provide the $N = 4$ implementation, where $N = 10$ has only minor changes). In green we provide the payoff text for $X = \$9$, in orange for $X = \$1$. Separate instructions for {Part two} are given to treatments in which N changes within a session. In the chat(1/2) treatment, the only changes are for the critical die rolls in the *Study Organization & Payment* section, where the supergame cutoff changes from 75 to 50. In the extension treatment in which only two of four players are needed for an Success signal, the description in the *Round Choices and Payoffs* are adjusted to accommodate the change.

Instructions

Welcome. You are about to participate in a study on decision-making. What you earn depends on your decisions, and the decisions of others in this room. Please turn off your cell phones and any similar devices now. Please do not talk or in any way try to communicate with other participants. We will start with a brief instruction period. During the instruction period you will be given a description of the main features of the study. If you have any questions during this period, raise your hand and your question will be answered in private at your computer carrel.

Study Organization & Payment.

- The study has two Parts, where each Part has 10 decision-making **Cycles**. Each Cycle consists of a random number of **Rounds** where you make decisions.
- At the end of the study, one of the two Parts will be selected for payment with equal probability. For the selected Part, one of the 10 Cycles will be randomly selected for payment. Your payment for this randomly selected Cycle will be based on your decision's in that Cycle's last Round.
- The number of Rounds in each Cycle is random, where only the last Round in each Cycle counts for payment. Which Round is the last is determined as follows:
  - In every Round, after participants make their decisions, the computer will roll a fair 100-sided die. If the die roll is greater than 75 (so 76–100) the round just completed is the one that is used to determine the current Cycle's payment, and the Cycle ends. If instead the computer's roll is less than 75 (so 1– 75) then the Cycle continues into another Round.
  - Because of this rule, after every Round decision there is a 25 percent chance that the current Round is the ones that count for the Cycle's payment, and a 75 percent chance that the Cycle continues and the decisions in a subsequent round will count for that Cycle payment.
- Your final payment for the study will be made up of a $6 show-up fee, and your payment from the last Round in the randomly selected Cycle.

Part 1.

- In the first part of the study you will make decisions in 10 Cycles. In each Cycle you will be matched with {another participant}{a group of three other participants} in the room for a sequence of Rounds. You will interact with the same {other participant}{group of three other participants} in all rounds of the cycle.
- Once a Cycle is completed, you will be randomly matched to a new {participant}{group of three participants} for the next Cycle.
- While the specific {participant}{participants} you are matched to is fixed across all Rounds in the Cycle, the computer interface in which you make your decisions is

anonymous, so you will never find out which participants in the room you interacted with in a particular Cycle, nor will others be able to find out that they interacted with you.

Round Choices and Payoffs. For each Round in each Cycle, you and the matched {participant}{participants in your group} will make simultaneous choices. {Both}{All four} of you must choose between either the **Green** action or the **Red** action. After you and the other {participant}{three participants} have made your choices, you will be given feedback on the {other participant's}{other participants'} choices that Round, alongside the Computer's die roll to determine if that Round counts for the Cycle payment.

If a particular Round is the Cycle's last, and that Cycle is the one selected for final payment, there are four possible payoff outcomes.

(i) If both you and {the other participant}{all three of the other participants} choose the Green action, you get a round payoff of $20.

(ii) If you choose the Green action and {the other participant chooses}{any of the other participants choose} Red, you get a round payoff of {$2}{$10}.

(iii) If you choose the Red action and {the other participant chooses}{all of the three other participants choose} Green, you get a round payoff of {$29}{$21}.

(iv) If both you and {the other participant}{any of the other three participants} choose the Red action, you get a round payoff of $11.

These four payoffs are summarized in the following table:

		Other {Participant's Action:}{Participants' Actions:}	
		{Green}{All 3 Green}	{Red}{Any of 3 Red}
Your Action:	Green	$20	{$2}{$10}
	Red	{$29}{$21}	$11

Some examples of these payoffs:

Case 1. Suppose you choose Green and {the other participant}{all three of the other participants} in the Cycle also choose Green. If that Round is the final one in the Cycle {both}{all four} of you would get a payoff of $20.

Case 2. Suppose {you}{you and two of the other participants} choose Green while the other participant chooses Red. If that Round is the final one in the Cycle {you}{you and the other two participants who chose Green} would get a payoff of {$2}{$10}, while the other participant would get a payoff of {$29}{$21}.

Case 3. Suppose you chooseRed while {the other participant chooses}{all three of the other participants choose} Green. If that Round is the final one in the Cycle you would get a payoff of {$29}{$21}, while the other {participant}{three participants} would get a payoff of {$2}{$10}.

Case 4. Suppose you and {the other participant choose Red.}{another participant choose Red while the other two participants choose Green.} If that Round is the final one in the

Cycle {you}{you and the other participant that chose Red} would get a payoff of {$11}{$11, while the other two participants would get a payoff of{$2}{$10}} .

Part 2. After Part 1 is concluded, you will be given instructions on Part 2, which will have a very similar structure to the task in Part 1.

{END OF PART 1 HANDOUT}

Part 2 Instructions {Between Only, handed out Supergame 11}. Part 2 is identical to Part 1. In each of the 10 Cycles in Part 2 you will again be matched to {another participant}{three other participants} in the room.

Similar to Part 1, the Cycle payoff is determined by the last round in the Cycle, where the payoff depends on the action you chose and the {action chosen by the matched participant}{actions chosen by the three matched participants} for that Cycle. Similar to Part 1, the below Table summarizes the payoff based upon the choices made in the Cycle's last round.

		Other {Participant's Action:}{Participants' Actions:}	
		{Green}{All 3 Green}	{Red}{Any of 3 Red}
Your Action:	Green	$20	{$2}{$10}
	Red	{$29}{$21}	$11

{END OF PART 2 HANDOUT}

Part 2 Instructions {Within Only, handed out Supergame 11}. Part 2 is very similar to Part 1. However, in each of the 10 Cycles in Part 2 you will instead be matched to three other participants in the room for each Cycle.

Similar to Part 1, the Cycle payoff is determined by the last round in the Cycle, where the payoff depends on the action you chose and the actions chosen by the three matched participants for that Cycle. If a particular Round is the Cycle's last, and that Cycle is the one selected for final payment, there are four possible payoff outcomes.

(i) If both you and all three of the other participants choose the Green action, you get a round payoff of $20.

(ii) If you choose the Green action and any of the other participants chooses Red, you get a round payoff of $2.

(iii) If you choose the Red action and all three other participants choose Green, you get a round payoff of $29.

(iv) If both you and any of the other three participants choose the Red action, you get a round payoff of $11.

These four payoffs are summarized in the following table:

		Other Participant's Action:	
		All 3 Green	Any of 3 Red
Your Action:	Green	$20	$2
	Red	$29	$11

Some examples of these payoffs:

Case 1. Suppose you choose Green and all three of the other participants in the Cycle also choose Green. If that Round is the final one in the Cycle all four of you would get a payoff of $20.

Case 2. Suppose you and two of the other participants choose Green while the other participant chooses Red. If that Round is the final one in the Cycle you and the other two participants who chose Green would get a payoff of $2, while the other participant would get a payoff of $29.

Case 3. Suppose you choose Red while all three of the other participants choose Green. If that Round is the final one in the Cycle you would get a payoff of $29, while the other three participants would get a Round payoff of $2.

Case 4. Suppose you and another participant choose Red while the other two participants choose Green. If that Round is the final one in the Cycle you and the other participant that chose Red would get a payoff of $11, while the other two participants would get a payoff of $2.

{END OF PART 2 HANDOUT}

Part 2 Instructions {Chat Only, handed out Supergame 11}. Part 2 is identical to Part 1 except for the beginning of each cycle where we will now allow the matched participants to chat to one another before the cycle begins. In each of the 10 Cycles in Part 2 you will again be matched to three other participants in the room.

Similar to Part 1, the Cycle payoff is determined by the last round in the Cycle, where the payoff depends on the action you chose and the actions chosen by the three matched participants for that Cycle. Similar to Part 1, the below Table summarizes the payoff based upon the choices made in the Cycle's last round.

		Other Participants' Actions:	
		All 3 Green	Any of 3 Red
Your Action:	Green	$20	$2
	Red	$29	$11

In contrast to Part 1 though, at the beginning of each new cycle, a chat window will be given to you, which will stay open for two minutes, or until all group members close it.

You may not use the chat to discuss details about your previous earnings, nor are you to provide any details that may help other participants in this room identify you. This is important to the validity of this study and will be not tolerated. However, you are encouraged to use the chat window to discuss the upcoming Cycle.

If at any point within the two-minute limit you wish to leave the chat, you can click the "Finish Chat" button. The other participants will be informed that you left.

{END OF PART 2 HANDOUT}

Appendix E. Strategies and the selection index

In the infinitely repeated prisoner's dilemma (RPD) the set of possible strategies is very large. However, the survey/meta-study of RPD lab experiments of Dal Bó and Fréchette (2018) shows that a small set of strategies rationalizes choices well for a large number of parameterizations. The five strategies that capture most choices are: (i) always cooperate, (ii) always defect; and three strategies in which cooperation is conditional, (iii) grim trigger, (iv) tit for tat, and (v) suspicious tit for tat. The difference between tit for tat and suspicious tit for tat takes place only in the first interaction, where tit for tat starts with cooperation and suspicious tit for that starts with defection. In subsequent rounds, both strategies cooperate if there the signal is that the other cooperated and defect otherwise.

The selection index that is used in the RPD literature and in this paper focuses on two strategies: always defect and grim trigger. A first reason to focus on these two strategies is that they capture the two very distinct types of behavior that may be supported in equilibrium, non-cooperative and conditionally cooperative behavior. There is one non-cooperative strategy that is a subgame-perfect equilibrium (always defect), but there are many conditionally cooperative strategies that depending on δ can be subgame perfect. However, amongst the set of strategies that empirically rationalizes the data, grim trigger is the only conditionally cooperative strategy that is subgame perfect. Tit for tat is a Nash equilibrium of the supergame, but there can be incentives to deviate from the punishment path. In addition, notice that if one selects two strategies and one is always defect, then selecting grim trigger or tit for tat for the other, does not change the path of play when combining these strategies. If one subject uses always defect and the other uses tit for tat, the outcome is cooperation in the first round and defection from the second round on, which is exactly what would happen if the other subject used grim trigger instead. Meanwhile, if both subjects were to use tit for tat, the outcome is cooperative in every round, which is also what would happen if both subjects used the grim trigger instead. In some sense there is little loss in focusing solely on the grim trigger as the conditionally cooperative strategy.

The main purpose of this appendix is to show that constraining to the five strategies identified as focal in Dal Bó and Fréchette (2018) also extends to our setting. The appendix starts with a brief description of the Strategy Frequency Estimation Method (SFEM), which was introduced in Dal Bó and Fréchette (2011).[E.1] From a big-picture perspective, the method takes the choices made by subjects and contrasts them against the choices that each strategy in a set of given strategies would have made had subject been using each of these other strategies. Using a mixture model that allows for errors in choices, the procedure reports the proportion of choices that are better rationalized by each strategy. In other words, the procedure works by inferring strategies that better rationalize choices. Dal Bó and Fréchette (2019) use an alternative procedure to study strategies, in particular, an experimental design that essentially familiarizes subjects with a set of strategies so that in the end subjects end up selecting a strategy directly to be implemented to make

[E.1] Further details on the procedure are available in the online appendix of Embrey, Fréchette, and Stacchetti (2013), and a Monte-Carlo-style analysis was also performed in Fudenberg, Rand, and Dreber (2010). The procedure has also been used to study strategies in other repeated-game experiments, for example, Aoyagi, Bhaskar, and Fréchette (2019), Vespa (2020), and Vespa and Wilson (2020).

choices for them. They contrast this elicitation procedure where the role of strategies is explicit to the strategies inferred from choices by the SFEM. They find consistency across the two methods.

Strategy Frequency Estimation Method. The goal of the procedure is to recover ϕ_k, which represents the frequency attributed to strategy k in the data. To illustrate how the procedure works, consider a set of strategies $\mathcal{K}$ that subjects may follow. Let $d^i_{gr}(\mathbf{h})$ be the choice of subject i and $k^i_{gr}(\mathbf{h})$ the decision prescribed for that subject by strategy $k \in \mathcal{K}$ in round r of supergame g for a given history $\mathbf{h}$. Strategy k is a perfect fit for round r if $d^i_{gr}(\mathbf{h}) = k^i_{gr}(\mathbf{h})$. The procedure models the probability that the choice (d) corresponds to the prescription of strategy k as:

$$\Pr\left(d^i_{gr}(\mathbf{h}) = k^i_{gr}(\mathbf{h})\right) = \frac{1}{1+\exp\left(\frac{-1}{\gamma}\right)} = \beta. \tag{4}$$

In Equation (4), $\gamma > 0$ is a parameter to be estimated. One interpretation of Equation (4) is that subjects can make mental errors in the implementation of a strategy, with β capturing the probability that the subject does not make such an error. To provide some intuition it is useful to consider the limit values of β. On the one hand, as $\gamma \to 0$, $\beta \to 1$ and the fit is perfect. On the other hand, as $\gamma \to \infty$, $\beta \to \frac{1}{2}$. In this case, the estimate of γ is so high that the prediction of the model is no better than a random draw.

With the specification for the mental error in Equation (4), the procedure uses maximum likelihood to estimate the frequency of strategy k in the data (ϕ_k) and parameter γ. Let y^i_{gr} be an indicator that takes value one if the subject's choice matches the decision prescribed by the strategy. Since Equation (4) specifies the probability that a choice in a specific period corresponds to strategy k, the likelihood of observing strategy k for subject i is given by:

$$p_i(k) = \prod_g \prod_r \beta^{y^i_{gr}} (1-\beta)^{1-y^i_{gr}}. \tag{5}$$

Aggregating over subjects we get: $\sum_i \ln\left(\sum_k \phi_k p_i(k)\right)$.[E.2] The procedure maximizes the likelihood function to obtain estimates for γ and the frequencies ϕ_k.[E.3]

An example may serve to clarify some aspects of the approach. Consider a case in which the set of included strategies is always defect (All D) and always cooperate (All C). The fit will be good (high β) if the population is composed of subjects who either almost always select D or almost always select C. The estimated frequency $\phi_{\text{All D}}$ would be the maximum likelihood estimate of the proportion of subjects who almost always select D. If a large proportion of subjects shifts between C and D within the supergame, neither

[E.2] To construct $p_i(k)$, consider a subject who is implementing the prescriptions of strategy k with mistake rate given by $1-\beta$. The case that the subject's choice matches the prescription ($y^i_{gr} = 1$) would be observed with probability β. If $y^i_{gp} = 0$, then the subject's choice does coincide with the one prescribed by the strategy.

[E.3] Since $\sum_k \phi_k = 1$, the procedure provides $|\mathcal{K}| - 1$ estimates and the $|\mathcal{K}|$-th strategy is computed by difference. The procedure also estimates γ. Following Equation (4) there is a one-to-one mapping between γ and β, so we will refer to the estimate of γ directly as an estimate of β.

strategy would accommodate their choices and the procedure will rationalize it with a low estimate of β.

The method depends on the pre-specified set $\mathcal{K}$ of included strategies. The information that subjects receive at the end of each round in our environments with $N > 2$ is similar to the information that subjects receive in an indefinitely repeated prisoner's dilemma. The reason is that subjects do not learn the specific choices of others, but an aggregate signal: either enough other people cooperated (a success) so that a cooperative outcome depended on their own choice or not. Because the history information is similar, we will focus on the same set of five strategies identified in Dal Bó and Fréchette (2007) and in fact, as we will show later, these five strategies are enough to obtain relatively high goodness of fit estimates (as captured by β).

Strategy Estimation: Results. In this section, we present results of the strategy estimation method. We divided the 20 supergame session into three parts. The last part consists of the last seven supergames, and the first part consists of the first seven supergames. We will first show estimates for the last supergames of the session and later provide results for the first supergames of the session.[E.4]

Final Supergames. Table E.1 presents the estimation results for each of our nine treatments. For each treatment and each of the five strategies the table reports the estimate and (whenever possible) the bootstrap-estimated standard errors.[E.5] The table also reports the β estimates that derive from the estimate of γ and the observations used in each estimation.[E.6]

Finally, the table reports goodness-of-fit estimates for a model in which we reduce the set of included strategies. $\beta^{\dagger}$ corresponds to the β estimate when we exclude tit for tat and suspicious tit for tat. In this case, the only conditional-cooperation strategy in the set is the grim trigger. Since the model uses maximum likelihood and the restricted model is nested, we use a likelihood-ratio test to evaluate the null hypothesis that the restriction does not bind. The row with the heading 'p-value†' reports the p-value corresponding to the test. The last two rows (referenced with ‡) perform a similar exercise but when the set of included strategies involves only always cooperate and always defect; that is, there are no history-dependent strategies included in this case.

A first observation is that all goodness-of-fit measures (the β estimates) are quite high. All estimates are at around 0.9 or higher. This indicates that five included strategies do a good job of rationalizing the data in all treatments. We now describe the results treatment by treatment.

[E.4] The results that we report qualitatively do not depend on having seven supergames among the early and late samples. The focus on seven is intended for two reasons. First, it allows for six supergames in between, so that it is possible to see if behavior early on changes relative to behavior much later in the session. Second, there is enough data in each seven-supergame sample.

[E.5] Recall that the procedure recovers standard errors for all the strategies but one. (See footnote E.3 for details).

[E.6] Observations for the chat treatment with $\delta = \frac{1}{2}$ are lower than in other treatments because in this case with a higher termination probability after each round, supergames are shorter.

In the treatment that corresponds to an indefinitely repeated prisoner's dilemma $\binom{N=2}{X=\$9}$, the strategies that carry most of the mass involve always defect (45.2 percent) and grim trigger (28.9 percent). There is, however, a non-negligible yet not-significant mass captured by tit for tat (18 percent). In the estimation that excludes tit for tat and suspicious tit for tat, there is a small reduction in terms of goodness of fit: from 0.929 to 0.912, which indicates that including only the Grim-Trigger strategy for conditional cooperation does not lead to the dataset to be rationalized with substantial additional error. However, the p-value of the likelihood ratio test in this case rejects the null, so that the restriction does impose a loss from that perspective. As a comparison reference, we point out that a SFEM estimation that only includes always cooperate and always defect, does lead to a relatively large decrease in the goodness of fit, which decreases from 0.929 to 0.804 (in $\beta^{\ddagger}$). In this case, the null in the likelihood ratio test is also rejected, but the increase in terms of the noise component needed to rationalize the data in this case is substantially larger relative to the case where the grim trigger is also included ($\beta^{\dagger}$).

The $\binom{N=4}{X=\$9}$ treatment is the treatment with the least amount of cooperation and the estimation reflects it. About a third of the mass corresponds to always defect and the rest essentially corresponds to suspicious tit for tat, a strategy that starts by defecting. In this case, the goodness-of-fit measure is extremely high (at 0.981) and essentially there is no loss when the estimation excludes tit for tat and suspicious tit for tat, as the $\beta^{\dagger}$ is unchanged up to the third decimal. Moreover, the likelihood ratio test does not reject the hypothesis that the constrained model is not restrictive.

The treatment with highest degrees of initial cooperation in our data is $\binom{N=4}{X=\$1}$, and little less than a quarter of the data is consistent with always cooperate. While this suggests a relatively large amount of unconditional cooperation, we point out that in this dataset we cannot distinguish between always cooperate and conditional cooperative strategies if the subjects do not experience others defecting.[E.7] The broader evidence, however, suggests that when cooperation takes place is conditional given that in treatments with more frequent defection there is essentially no evidence of large amounts of unconditional cooperation. In fact, in this treatment the most popular strategy is tit for tat, capturing more than 50 percent of the mass. There is also a close to 20 percent that corresponds to always defect. We note that while the likelihood ratio test in this case rejects the null that eliminating tit for tat and suspicious tit for tat is not restrictive, the loss in terms of

[E.7] While a strategy in an infinitely repeated game specifies what to do at each possible decision node (an infinite-dimensional object), the observed set of choices for a subject correspond to a specific path of play. To increase possible identification Vespa (2020) uses a one-period-ahead strategy method (OASM) in which subjects make choices in round r without knowing what the other did in round r-1. That is, the subject makes a choice in round r for each possible choice that the other could have taken in round r-1. After making these choices, the subject learns the actual history of play for round r-1, and their choice for round r is implemented for the actual choice that the other took in round r-1. In this way, it is possible to retrieve in an incentivized manner choices that subjects would have made off the path of play. Implementing OASM is costly particularly in terms of time with instructions, but also because it reduces the number of supergames that subjects can reasonably play within a session given that they must make more decisions per round. Since the goal of the current paper does not lie in identifying strategies, we decided not to include it in our design.

goodness-of-fit is rather small. Specifically, the goodness of fit drops from 0.939 to 0.931. The drop in goodness of fit is much larger when grim trigger is also excluded. The $\beta^{\ddagger}$ coefficient is at 0.840, a relatively much larger drop from 0.939.

In the last core treatment in our dataset $\binom{N=10}{X=\$1}$ all strategies for which standard errors can be computed have statistically significant estimates. About a quarter of the mass corresponds to always defect and slightly above ten percent for always cooperate. Grim trigger, tit for tat and suspicious tit for tat account for close to 60 percent of the mass jointly. However, when the estimation excludes tit for tat and suspicious tit for tat, the goodness-of-fit estimate does not change up to the third decimal and remains at a relatively high level. Consistent with this, the likelihood ratio test does not suggest that the restriction leads to a loss.

Overall, we now summarize the main takeways from the perspective of focusing on always defect and grim trigger in the index measure:

- The estimates in the four core treatments that are used to test the extensions of the basin suggest that focusing on these two strategies does not lead to a substantial loss. Either because a likelihood ratio test directly points towards the restriction not binding or because when it binds the relative loss is small (as measured by the goodness-of-fit estimates).
- Meanwhile, when the estimation is further restricted to exclude grim trigger, in all treatments we see that the likelihood ratio test indicates that the restriction does bind and in most cases it in fact leads to a relatively large loss in goodness of fit.

We now discuss the estimates for our extension treatments. Treatment ($N = 2$ to $4/X = \$9$) the last seven supergames correspond to a part of the session where $N = 4$, so that estimates can be compared to treatment ($N = 4/X = \$9$). The estimates in this treatment are noisier (relative to other treatments), which is likely due to the fact that subjects are not as experienced with this environment in the within treatment relative to ($N = 4/X = \$9$), where by supergame 14 subjects had experienced ten supergames more with this parameterization by this point.[E.8] However, the big picture is similar. First, there is a very small reduction in the goodness-of-fit estimate when the set of strategies excludes tit for tat and suspicious tit for tat (from 0.950 to 0.948).[E.9] Second, most subjects appear to use strategies that are captured either by directly not cooperating (always defect) or by starting in a non-cooperative manner (suspicious tit for tat).

In the second half of the session, treatment ($N = 2$ to $4/X = \$9$) as shown in Figure 3(A), shows a slow adjustment towards the low cooperation rates seen in ($N = 4/X = \$9$). This is a treatment where in the first half of the session $N = 2$ and by the last seven supergames with $N = 4$ cooperation rates are lower, but the reduction takes place at a slow pace. In this case, the restriction that excludes both tit for tat strategies leads to a larger loss in

[E.8] In fact, the estimates for ($N = 2$ to $4/X = \$9$) in Table E.1 are closer to the estimates for treatment ($N = 4/X = \$9$) using the *first* seven supergames, which are reported in Table E.2. In both cases, the largest mass corresponds to always defect (around sixty percent) and the next strategy in popularity is suspicious tit for tat.

[E.9] The likelihood ratio test also leads to the same result using a 95 percent confidence level.

goodness-of-fit relative to other treatments, from 0.921 to 0.883. However, additionally excluding grim trigger leads to a much larger loss, with $\beta^{\ddagger}$ at 0.819.

We conclude that both within treatments share some similar features with the corresponding between treatments in their second half, but there is an adjustment given that the first ten supergames displayed different primitives and the adjustment is also reflected in the SFEM estimates.

Both chat treatments were conducted with the same background primitives ($N = 4/X =$ \$9) and in the chat treatment that keeps $\delta = \frac{3}{4}$, there is a large difference relative to ($N = 4/X =$ \$9), where no such communication was available. With chat, grim trigger and tit for tat strategies essentially capture almost all the mass. This is consistent with the large cooperation rates that were documented in this case. The other side of this is that the coefficients of always defect and the conditional cooperation strategy that stats by defecting (suspicious tit for tat) are at zero, while they essentially captured all the mass in ($N = 4/X =$ \$9).

A big shift happens when there is chat and $\delta = \frac{1}{2}$. In this case, always defect captures more than sixty percent of the mass and Suspicious Tit for Tat adds another 11 percent. In other words, these findings are consistent with what we reported in the text, meaning that even if coordination is eased in this treatment, supporting cooperation is very difficult.[E.10]

The final extension treatment corresponds to primitives ($N = 4/X =$ \$9), but where only two players are needed for cooperation to result. Consistent with our reports in the text, the strategy that captures most of behavior in this treatment is always defect, with close to three-quarters of the mass. Hence, the evidence from the SFEM is consistent with the fact that cooperation in this treatment is quite unlikely despite the fact that the number of players needed for a cooperative outcome is smaller than N.

Early Supergames. We conclude this section of the appendix by describing SFEM estimates for the first seven supergames; essentially a period of the session in which learning is more likely to be taking place. Results for all treatments are presented in Table E.2.

Starting with our four core treatments, a first observation is that the goodness-of-fit estimates (β) are still quite far from random, with the smallest at 0.812. This suggests that even if constrained to few strategies, most of the data can be rationalized by the small set. However, there is a large gain in goodness-of-fit estimates when we when compared to the last seven supergames, described earlier and reported in Table E.1. In all four cases the drops are relatively large in magnitude, with the smallest involving a reduction from 0.929 to 0.874 and the largest from 0.934 to 0.812. This suggests that as subjects gather experience, their behavior becomes more consistently captured by the five strategies included in the estimation.

Second, comparing between the first seven supergames and the last seven supergames the broad picture treatment by treatment is not very different. There can be adjustment in terms of what strategy best captures behavior, but the changes do not appear to be

[E.10] While there is a non-negligible mass for Grim Trigger, any time a subject playing that strategy is matched with a subject playing always defect or suspicious tit for tat, will end up defecting in round 2. Given the high estimates for these two strategies, the likelihood of long-term cooperation is very small.

TABLE E.1. SFEM output: last seven supergames

Strategies	$\binom{N=2}{X=\$9}$	$\binom{N=4}{X=\$9}$	$\binom{N=4}{X=\$1}$	$\binom{N=10}{X=\$1}$	$\binom{N=2\to4}{X=\$9}$	$\binom{N=4\to2}{X=\$9}$	$\binom{N=4}{X=\$9}$ Chat	$\binom{N=4}{X=\$9}$ Chat, $\delta=\frac{1}{2}$	$\binom{N=4}{X=\$9}$ Success if 2 Coop
All C	0.017	0.000	0.231*	0.133***	0.014	0.073	0.083	0.016	0.017
	(0.024)	(0.009)	(0.124)	(0.046)	(0.018)	(0.070)	(0.070)	(0.022)	(0.031)
All D	0.452***	0.313***	0.182**	0.265***	0.549	0.218**	0.000	0.613***	0.735***
	(0.154)	(0.010)	(0.09)	(0.017)	(0.240)	(0.075)	(0.003)	(0.118)	(0.065)
Grim trigger	0.289**	0.009	0.046	0.094***	0.185**	0.140*	0.669**	0.262***	0.101*
	(0.123)	(0.010)	(0.126)	(0.025)	(0.127)	(0.085)	(0.281)	(0.092)	(0.057)
Tit for tat	0.180	0.009	0.535***	0.094***	0.000	0.458***	0.248	0.000	0.148**
	(0.112)	(0.010)	(0.151)	(0.025)	(0.063)	(0.086)	(0.300)	(0.009)	(0.061)
Suspicious tit for tat	0.063	0.670	0.006	0.414	0.252	0.111	0.000	0.110	0.000
β	0.929	0.981	0.939	0.934	0.950	0.921	0.975	0.873	0.899
# Observations	1,360	1,320	1,632	1,500	1,296	1,304	1,560	884	1,152
$\beta^{\dagger}$	0.912	0.981	0.931	0.934	0.948	0.883	0.974	0.871	0.896
p-value†	<0.000	1.000	<0.000	1.000	0.096	<0.000	0.041	0.597	0.003
$\beta^{\ddagger}$	0.804	0.978	0.840	0.897	0.895	0.819	0.883	0.809	0.863
p-value‡	<0.000	<0.000	<0.000	<0.000	<0.000	<0.000	<0.000	<0.000	<0.000

Note: (i) Bootstrapped standard errors in parentheses. Level of significance: *** 1 percent; ** 5 percent; * 10 percent. (ii) $\beta^{\dagger}$ corresponds to the β estimate in case tit for tat and suspicious tit for tat are excluded. (iii) p-value† reports the p-value of a likelihood ratio test in which the restricted model excludes tit for tat and suspicious tit for tat. (iv) $\beta^{\ddagger}$ corresponds to the β estimate in case grim trigger, tit for tat, and suspicious tit for tat are excluded. (v) p-value‡ reports the p-value of a likelihood ratio test in which the restricted model excludes grim trigger, tit for tat, and suspicious tit for tat.

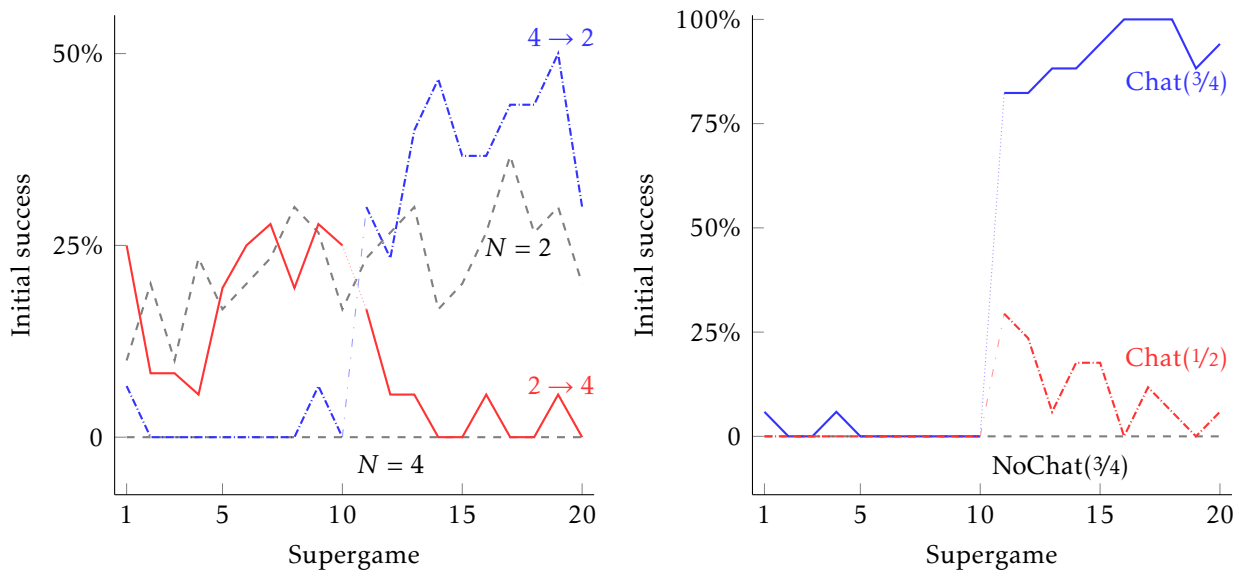
conceptually meaningful. For instance, in ($N = 4/X = \$9$) the largest mass in the first seven supergames corresponds to always defect (at 60 percent) and the second largest is captured by suspicious tit for tTat (at 33.8 percent). In the last seven supergames, the order is reversed, with 67 percent for suspicious tit for tat and slightly more than thirty percent for always defect. In both cases, however, these two strategies jointly capture more than ninety percent of the mass and both are strategies that start by selecting Defect. That is, the odds of a second period in which there is a cooperative outcome when the majority of the population plays one of these strategies is slim.

Perhaps a slightly different picture is painted by treatment ($N = 10/X = \$1$), where the coefficients for the first seven supergames suggest that close to fifty percent of the mass corresponds to strategies that start by cooperating. This declines in the last seven supergames where almost seventy percent of the mass corresponds to strategies that start by defecting. This suggests that subjects in this treatment start by trying to cooperate but in time learn to defect.

Moving to our extension treatments, we first note that ($N = 2/X = \$9$) and ($N = 2$ to $4/X = \$9$), which in the first seven supergames have the same parameterization have comparable estimates for all strategies, and, in fact, goodness-of-fit estimates that coincide up until the third decimal point. A similar picture emerges in the comparison between ($N = 4/X = \$9$) and ($N = 4$ to $2/X = \$9$). The joint mass of strategies that start by defecting in ($N = 4/X = \$9$) is close to 94 percent. In ($N = 4$ to $2/X = \$9$), the number is 81 percent. While the number is lower in the latter, in both cases share behavior is overwhelmingly captured by strategies that are at least initially non-cooperative.

When we compare behavior in the first seven supergames of within treatments to behavior in the last seven supergames, there can be large changes. In the first seven supergames of ($N = 2$ to $4/X = \$9$), where $N = 2$ a mass of close to forty percent corresponds to strategies that start by cooperating. Meanwhile, in the last seven supergames, where $N = 4$ eighty percent of the mass is captured by strategies that start by defecting. The change is even more striking when N is reduced in the second half of the session. In the first seven supergames of ($N = 4$ to $2/X = \$9$), slightly more than eighty percent of the mass corresponds to strategies that start by defecting. However, in the last seven supergames (when N is reduced to 2), only about a third of strategies start by defecting. This illustrates that the patterns captured by the SFEM estimates are consistent with the patterns described in the text.

An even larger change is documented between the beginning and end of the session in chat treatments, where the possibility to exchange messages is only introduced in the second half of the session. The first half of the session in ($N = 4/X = \$9$) with chat is identical to treatment ($N = 4/X = \$9$) and treatment ($N = 4$ to $2/X = \$9$). The estimates reflect that with more than eighty percent of the mass corresponding to strategies that start by defecting in all three treatments. However, as we reported earlier, the large masses that correspond to always defect and suspicious tit for tat essentially disappear once chat is introduced. In the case with $\delta = \frac{1}{2}$ there is an effect in the same direction, but of a much



(A) Between vs. Within ($X = \$9$)

(B) Explicit vs. Implicit $\left(\begin{smallmatrix} N=4 \\ X=\$9 \end{smallmatrix}\right)$

FIGURE E.1. Initial success rates in extensions (by supergame)

smaller magnitude. Initially, more than 90 percent of the mass being rationalized by always defect, but strategies that start by defecting capture close to three-quarters once chat is introduced.

Finally, in the extension treatment where only two of four players are needed for a cooperative outcome we see some patterns that are common to most other treatments. For instance, the goodness-of-fit estimates is higher in the last seven supergames. However, the effect of learning is relatively smaller as the session evolves: most subjects use strategies that start by defecting early and late in the session.

TABLE E.2. SFEM output: first seven supergames

Strategies	$\binom{N=2}{X=\$9}$	$\binom{N=4}{X=\$9}$	$\binom{N=4}{X=\$1}$	$\binom{N=10}{X=\$1}$	$\binom{N=2\to4}{X=\$9}$	$\binom{N=4\to2}{X=\$9}$	$\binom{N=4}{X=\$9}$ Chat	$\binom{N=4}{X=\$9}$ Chat, $\delta=\frac{1}{2}$	$\binom{N=4}{X=\$9}$ Success if 2 Coop
All C	0.048	0.017	0.289^{***}	0.228^{***}	0.028	0.036	0.015	0.016	0.013
	(0.034)	(0.024)	(0.082)	(0.055)	(0.037)	(0.026)	(0.019)	(0.020)	(0.017)
All D	0.517^{***}	0.600^{**}	0.287^{***}	0.440^{**}	0.511^{***}	0.320	0.315	0.921^{***}	0.716^{***}
	(0.086)	(0.270)	(0.063)	(0.211)	(0.080)	(0.210)	(0.267)	(0.253)	(0.070)
Grim trigger	0.160^{*}	0.000	0.000	0.308^{***}	0.224^{**}	0.000	0.179^{**}	0.032	0.098^{**}
	(0.093)	(0.021)	(0.065)	(0.080)	(0.106)	(0.024)	(0.083)	(0.024)	(0.046)
Tit for tat	0.147^{**}	0.045	0.392^{***}	0.024	0.121^{**}	0.152	0.000	0.032	0.081^{*}
	(0.057)	(0.043)	(0.115)	(0.092)	(0.054)	(0.113)	(0.047)	(0.023)	(0.043)
Suspicious tit for tat	0.129	0.338	0.032	0.000	0.117	0.492	0.491	0.000	0.092
β	0.874	0.922	0.851	0.812	0.874	0.894	0.912	0.900	0.844
# Observations	1,840	1,840	1,992	1,380	2,088	1,820	2,080	1,124	1,868

Note: Bootstrapped standard errors in parentheses. Level of significance: *** 1 percent; ** 5 percent; * 10 percent.