TESTING MODELS OF STRATEGIC UNCERTAINTY: EQUILIBRIUM SELECTION IN REPEATED GAMES

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

Abstract. In repeated games, where both collusive and noncollusive outcomes can be supported as equilibria, it is crucial to understand the likelihood of selection for each type of equilibrium. Controlled experiments have empirically validated a selection criterion for the two-player repeated prisoner's dilemma: the basin of attraction for always defect. This prediction device uses the game primitives to measure the set of beliefs for which an agent would prefer to unconditionally defect rather than attempt conditional cooperation. This belief measure reflects strategic uncertainty over others' actions, where the prediction is for noncooperative outcomes when the basin measure is full, and cooperative outcomes when empty. We expand this selection notion to multi-player social dilemmas and experimentally test the predictions, manipulating both the total number of players and the payoff tensions. Our results affirm the model as a tool for predicting long-term cooperation, while also speaking to some limitations when dealing with first-time encounters.

1. Introduction

Identifying which of many possible equilibria best captures economic behavior is of central importance for applications with repeated interactions. For example, in models of oligopoly, both collusive and noncollusive equilibria can arise. To help guide assumptions over equilibrium selection, experimental work has sought to uncover simple theoretical criteria that can predict the likelihood of collusion based on primitives such as payoffs and discount rates. Thus far, the main body of experimental work on selection has focused on the canonical two-player indefinitely repeated prisoner's dilemma (RPD). However, it is unknown to what extent the predictive criteria from two-player environments can be used to predict selection in games with more than two players.

The *basin of attraction for always defect* (Blonski and Spagnolo, 2001, 2015) has been shown to simply organize experimental data in a meta-study of two-player RPD games (Dal Bó and Fréchette, 2018). The measure's calculation inputs are stage-game payoffs and the discount factor. The measure's output is the set of beliefs on the other player choosing to conditionally cooperate for which permanent defection is a best response. The wider the set of beliefs where defection is a best response, the greater the risk in attempting to cooperate, which is why this measure is thought of as a proxy for uncertainty over others' behavior (i.e., strategic uncertainty). Experimental data starting with Dal Bó and Fréchette (2011) show that when the theoretical size of the always defect basin is high (low), observed cooperation rates tend to be low (high). Furthermore, when the basin size is less

Date: February, 2024.

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

than (greater than) half, it aligns with the concept of risk dominance. Therefore, this simple ordinal property serves as a clear line-in-the-sand for predicting regions where we expect/do not expect collusive outcomes.

Our paper focuses on a simple and relevant extension of the model to more than two players. In environment with N players, an agent must assess the chances that multiple other players will cooperate. We develop two benchmarks to extend the measure of strategic uncertainty. A natural theoretical benchmark is an independent-belief extension, which considers symmetric and independent beliefs about each of the other players. At the other extreme, we also consider a setting where beliefs about other players are perfectly correlated. In the perfectly-correlated extension, the addition of another player does not impact strategic uncertainty, as the actions of other players in the game are assumed to be perfectly correlated. This extreme serves as a natural interpretation for a null hypothesis over the number of players, where the prediction for the N-player game will, ceteris paribus, be the same as for its two-player counterpart. With these two benchmarks, our experimental treatments allow us to isolate the effects of strategic uncertainty due to the higher N relative to the standard two-player game.

Our experimental design provides directional and null predictions for each extension. To achieve this goal, we introduce a second treatment variable that the prior RPD literature highlights as a clear driver of behavior: the stage-game payoff gain that a player gets from a defection, x. For illustration, consider the effect of reducing x. A lower temptation payoff reduces strategic uncertainty according to both extensions, predicting higher cooperation. In particular, this second parameter provides for a directional prediction under the correlated extension. Further, by shifting both N and x together, we can generate null-effect treatment comparisons under our independent extension. That is, a predicted increase in strategic uncertainty from higher N can be perfectly compensated by a reduction in x, generating a stronger test of this extension.

By manipulating N and x, we create a 2×2 between-subjects design across our two basin extensions, generating directional and null predictions for each. This design enables us to assess which of the two extensions more effectively captures behavior. Furthermore, it remains a possibility that both extensions lack predictive power, indicating that coordination in a game involving more than two players may not be associated with strategic uncertainty.

Our core results indicate that the independent-basin extension best organizes the longer-run (i.e., *ongoing*) behavior. Under the independent-belief extension we observe large behavioral shifts in the predicted direction when varying N in isolation. When manipulating x and N in opposite directions we observe no significant changes in behavior, which is in line with the prediction of the independent extension. However, while the independent extension succeeds in predicting the longer-run cooperation that may be the most relevant for applications, the measure is not a good predictor of intentions to cooperate, as captured by *initial* cooperation.

Our core findings suggest that equilibrium selection is driven by strategic uncertainty over the behavior of other players. Therefore, eliminating or minimizing doubts about others' actions should render the predictions from the basin model irrelevant. We explore

this hypothesis in an additional treatment with pre-play communication. Here, participants have the opportunity to exchange free-form messages before the repeated game begins, a feature designed to reduce uncertainty about the strategic intentions of others (see Kartal and Müller, 2022). In a parameterization where initial cooperation rates are below one percent in the treatment without communication, the introduction of pre-play chat shifts behavior to the other extreme, resulting in initial (ongoing) cooperation rates of 95 (80) percent.

While we observe that predictions based on strategic uncertainty lose validity with explicit collusion, the model performs well in several robustness treatments with tacit collusion. We first assess the extent to which selected equilibria are sticky when game parameters change. Specifically, we introduce a group-size change halfway through an experimental session, transitioning the same participants from a four-player game to a two-player game, and vice versa. If the selected equilibrium in the first half is sticky, then varying N will not affect behavior. As a consequence, the independent-basin extension, which better organizes results in our between-subjects comparison, would be irrelevant for comparative statics within a market. However, if strategic uncertainty changes with new parameters, then an increase (decrease) in N should decrease (increase) beliefs in others' cooperation after the parameter change, altering cooperation. Our findings indicate some stickiness in behavior in the short run, but we do not observe stickiness in longer-term behavior. Cooperation levels adjust after a change in N, moving with experience toward the levels observed in the sessions with fixed N. These results validate the independent-extension measure as a predictor for equilibrium selection even within a particular context.

In our second robustness treatment, we relax the condition for group success. Our previous treatments require joint cooperation from all N players for a group-wide success. Hence, increasing the number of players from two to four makes it mechanically harder to achieve success at any fixed rate of individual cooperation (i.e., two-from-two is easier than four-from-four). In this robustness treatment, however, we only require half of the players to cooperate for a group-wide success. Consider comparing the baseline reference (two-from two) to this robustness (two-from-four) treatment. The comparison increases N from two to four, but this change does not make it mechanically *harder* to achieve success. That is, achieving a success in the two-from-two game is harder than in the two-from-four case. Despite easing the conditions for success in the robustness treatment, strategic uncertainty increases, and the independent-extension predicts lower cooperation. The reason behind this shift is that coordination in the two-from-four treatment becomes more challenging, as it introduces the question of which players must cooperate and which can free-ride. Consistent with the counter-intuitive prediction, our experimental results indicate that cooperation rates are lower under the two-from-four requirement for an efficient outcome than the standard RPD where we require two-from-two.

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,[1] we generalize the equilibrium selection model by adding an additional source of strategic uncertainty: the number of players, N.[2] Where the literature has developed 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.[3]

Our environment also allows us to better distinguish between 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.[4] 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 ongoing collusion rather than initial intentions to collude.[5]

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, our measures of strategic uncertainty are 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 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

[1] As highlighted by Berry, Coffman, Hanley, Gihleb, and Wilson (2017), experimental replications can seem less frequent than they are if papers fail to advertise the features that are replications.

[2] The basin measure, detailed in Section 2, seeks to capture the intuition from Harsanyi and Selten (1988)'s risk dominance, and was initially proposed by Blonski and Spagnolo (2001, 2015). The basin measure was first empirically tested by Dal Bó and Fréchette (2011). See also Fudenberg, Rand, and Dreber (2012) for an examination of the effects with imperfect monitoring, Kartal and Müller (2022) for a test of a selection theory based on individual heterogeneity in preferences over dynamic strategies, and Mermer, Mueller, and Suetens (2021) for two-player games of strategic complements and substitutes.

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

[4] With two players, the introduction of sequential moves adds extra variability for identification. Ghidoni and Suetens (2022) also find that ongoing measures are better predicted than initial rates.

[5] 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."

environment changes across supergames, and the space of strategies grows exponentially. 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. That paper identifies a similar strategic uncertainty measure constructed around the most-efficient Markov perfect equilibrium and the best *symmetric* collusive equilibrium. A strategic-uncertainty model based on these strategies predicts behavior, where these strategies dovetail with repeated game strategies in the simpler environment studied here.[6]

An experimental literature on behavior in oligopolies documents that collusion 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.[7] We contribute to this literature on two margins. First, we examine how changes to N affect outcomes in an infinite horizon with collusive and noncollusive equilibria. Second, and crucially, we focus not only on the qualitative directional effects of N, but also, on validating the model suitability for studying strategic uncertainty. Specifically, the model, if validated, will help us understand the extent of substitutability between game primitives, which, in turn, can prove useful in predicting the directional effects of more-nuanced, multi-dimensional counterfactuals.

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 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, 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.[8]

The effects of communication devices as a support for collusion are well established in the experimental literature. As surveyed in Cason (2008) and Harrington, Gonzalez, and Kujal (2016), more-structured, limited forms of communication usually result in small,

[6]The applications of dynamic games are extensive, thanks to their inherent flexibility. The ongoing research on equilibrium selection in dynamic games builds upon recent work, among others, by Battaglini, Nunnari, and Palfrey (2012, 2016); Agranov, Fréchette, Palfrey, and Vespa (2016); Kloosterman (2019); Vespa and Wilson (2019); Rosokha and Wei (2020); Salz and Vespa (2020); Vespa (2020).

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

[8]Differences in behavior tend to be stickier when changes are small or introduced gradually. Weber (2006) shows that gradually increasing the number of players in a coordination game yields different results relative to situations where the game begins with a large group. The gradual introduction of changes to the payoff primitives has also been shown to have effects in repeated games; see Kartal, Müller, and Tremewan (2021). This suggests that the selection notions under examination are relevant for "large" counterfactual changes. Future research can help clarify how to integrate "large" into a predictive model of selection.

temporary collusive gains, where free-form communication generates large, long-lasting effects.[9] For these reasons, we also 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 predictable 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 the experimental literature on repeated games has largely focused on 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 marginal per capita return. 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).[10] Risk dominance has been shown to have substantial predictive content in simple coordination games with tradeoffs over payoff dominance and risk dominance (see Battalio, Samuelson, and Van Huyck, 2001; Brandts and Cooper, 2006; Dal Bó, Fréchette, and Kim, 2021, and references therein). Therefore, strategic uncertainty has demonstrated its usefulness as a theoretical selection device in both static and repeated games. We contribute to this literature with an experiment that explicitly tests and shows how the predictive effects extend further to multiplayer infinite-horizon settings.

Finally, our last robustness treatment provides a connection to coordination games with asymmetric-payoffs equilibria (such as the battle-of-the-sexes game). Coordination in this treatment requires at least two out of four players to cooperate, which relaxes the condition for success relative to the two-from-two treatment. But efficient equilibrium outcomes have two players coordinate on defecting (and getting a higher payoff) and two players cooperating (and getting a relatively lower payoff). The asymmetry means that

[9] 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 et al. (2016); Wilson and Vespa (2020).

[10] 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, participants are given the stage-game payoffs and actions, from which strategies (e.g., grim trigger or tit for tat) and gross payoffs are endogenously derived. Our use of risk dominance in a repeated game refers to the concept constructed by Spagnolo and Blonski (2001) inspired by Harsanyi and Selten (1988).

each player would prefer to be a free-rider. Similar tensions arise in one-shot coordination games with asymmetric payoffs like the battle-of-the-sexes game, where the literature has documented relatively high coordination failure rates that result in lower payoffs (cf. Cooper, DeJong, Forsythe, and Ross, 1990, 1993, 1994; Straub, 1995; Crawford, Gneezy, and Rottenstreich, 2008). The low cooperation rates in our two-from-four robustness treatment suggest that coordination challenges introduced by asymmetric payoffs already documented in the one-shot battle-of-the-sexes game can extend to a repeated-game setting like ours.[11]

2. Generalizing the Basin of Attraction

We begin this section by summarizing the progress made towards validating the basin of attraction for always defect as a theoretical prediction in the two-player RPD literature. A reader familiar with the literature can skip to Section 2.2, where we extend the framework by introducing a new parameter for strategic uncertainty, the number of players N.

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: $\Delta\pi := \pi(C,C) - \pi_0$.[12] 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 $\pi_i(C,D) = \pi_0 - s\Delta\pi$. The parameters $g > 0$ and $s > 0$ capture the relative temptation- and sucker-payoffs, respectively.

The strategic-uncertainty measure we focus on is based on two focal extensive-form RPD strategies:[13]

(i) *always defect*, $\alpha_{\text{All-D}}$, which plays the stage-game Nash in all rounds (the worst-case subgame-perfect equilibrium of the game).

(ii) *Grim trigger*, α_{Grim}, which begins by cooperating, but switches to always defect after any defection in past play (the collusive subgame-perfect equilibrium).[14]

[11] However, as Cooper and Weber (2020) argue, battle-of-the-sexes implementations with naturally-occurring strategy labels can display higher coordination rates (for instance Holm, 2000). Since in our setting actions represent abstract choices, we cannot assess the extent to which these findings extend to repeated games.

[12] The game payoffs π can also be transformed as $\tilde{\pi} = (\pi_i - \pi_0)/\Delta\pi$ to express all payoffs relative to joint defection (π_0) in units of the optimization premium ($\Delta\pi$).

[13] In Online Appendix G, we explain why focusing on these two strategies is both useful and minimally restrictive.

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

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 $p \cdot \alpha_{\text{Grim}} \oplus (1-p) \cdot \alpha_{\text{All-D}}$. The *basin for always defect* is defined as the set of beliefs p for which player i receives a higher expected payment from $\alpha_{\text{All-D}}$ than α_{Grim}. The always-defect belief basin is therefore the interval $[0, p^\star(g,s,\delta)]$ with the critical-point/interval-width given by:[15]

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

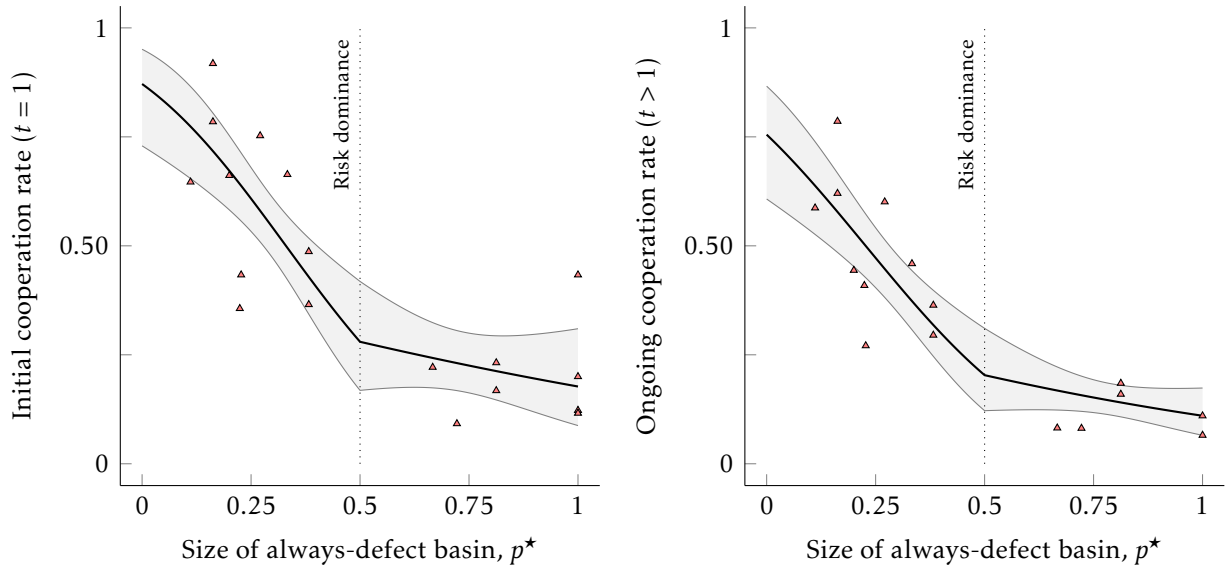
Consequently, the PD stage-game payoffs are used as primitive inputs into a risk/reward model of collusion based upon strategic uncertainty.

Equation (1) represents a theoretical relationship between the payoff primitives of the game and a critical strategic belief over the other player's likelihood of collusion. The hypothesized relationship is monotone, which allows unambiguous directional predictions for any counterfactual change in the primitives. 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.

Using results from the meta-study on the two-player RPD (Dal Bó and Fréchette, 2018), we illustrate the relationships between the scalar basin-size measure of strategic uncertainty and our two focal outcome measures: initial and ongoing cooperation rates. In both panels of Figure 1, the horizontal axis represents the theoretical measure of strategic uncertainty, while the vertical axes represent one of our outcome measures. In Panel (A) we present the results for *initial cooperation*; in Panel (B) we present results for *ongoing cooperation*. The solid line in both panels indicates $\hat{C}_{\text{Meta}}(p^\star)$, which we use to denote the predicted cooperation rate using meta-study data at each $p^\star$.[16] The shaded region represents the 95 percent confidence interval for $\hat{C}_{\text{Meta}}(p^\star)$.

[15] In the case that the strategy $(\alpha_{\text{Grim}}, \alpha_{\text{Grim}})$ is not a subgame-perfect equilibrium of the repeated game, the basin size for always defect is defined as $p^\star(g,s,\delta) = 1$.

[16] We estimate a probit regression using meta-study data from 996 participants clustered across 18 experimental treatments, where we focus on late-session cooperation (supergames 16-20). The individual-level cooperation decisions serve as the left-hand side variable, and the basin size is included on the right-hand side in a piecewise-linear fashion around the risk-dominance dividing point. Our econometric specification is inspired by Dal Bó and Fréchette (2018, Table 4). However, to maintain a continuous relationship, we modify their specification by eliminating a degree of freedom that allowed for a discontinuity at $p^\star = 1/2$.



(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.

For both initial and ongoing cooperation, we find essentially the same predicted relationship $\hat{C}_{\text{Meta}}(p^\star)$, consistently low levels of cooperation when always-defect is risk dominant ($p^\star > 1/2$); and a significantly decreasing relationship with $p^\star$ when collusion is risk dominant ($p^\star < 1/2$).

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 grim triggers, 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), data from RPD games do not provide enough variation to identify whether theoretical notions track more closely with either empirical measure. Consequently, with only two players it is challenging to identify the extent to which the strategic-uncertainty measure predicts initial intentions versus successful coordination.[17] However, as we will show below, adding more players provides additional variation that will allow us to differentiate between the two measures of cooperation.[18]

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

[18] In settings where collusion requires N agents to initially cooperate to produce ongoing cooperation, the relationship is given by initial cooperation rate to the N-th power. Separate identification between the two measures is possible by comparing treatments with different values of N.

2.2. **Extending to $N > 2$.** We now extend the strategic-uncertainty model to an N-player environment. To achieve this, we consider a family of symmetric social dilemmas that nest the standard two-player RPD. To maintain a constant 2×2 stage-game representation for all N, our family of dilemmas makes use of an aggregate 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, players' payoffs are determined by their own-action a_i and a deterministic binary signal $\sigma(a_{-i}) \in \{S(\text{uccess}), F(\text{ailure})\}$ of the others' actions, a_{-i}. In particular, the generic player i's stage-game payoff and signal function are given, respectively, by:

$$\pi_i\left(a_i, \sigma\right) = \begin{cases} \pi_0 + \Delta\pi & \text{if } a_i = C, \sigma = S, \\ \pi_0 + \Delta\pi\,(1+x) & \text{if } a_i = D, \sigma = S, \\ \pi_0 - \Delta\pi x & \text{if } a_i = C, \sigma = F, \\ \pi_0 & \text{if } a_i = D, \sigma = F; \end{cases} \tag{2}$$

$$\sigma(a_{-i}) = \begin{cases} S & \text{if } a_j = C \text{ for all } j \neq i, \\ F & \text{otherwise.} \end{cases} \tag{3}$$

These choices lead to a symmetric game, in which 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.[19]

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;[20] (ii) The number of players, N; and (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 of cooperating x 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 over α_{Grim} and $\alpha_{\text{All-D}}$.[21] Our family of social dilemmas requires cooperation from all N players for everyone to get an S signal. Thus, strategic uncertainty reduces to the probability that

[19] In Section 5, we present a treatment where cooperative outcomes require only two out of four players to cooperate. Introducing the possibility of achieving cooperative outcomes with some players not cooperating gives rise to a free-riding problem. However, in our main treatments, we sidestep this issue by assuming efficient ongoing cooperative outcomes only when all N players cooperate.

[20] In the meta-study notation this is implemented with $s = g = x$. This single-parameter formulation is equivalent to the Fudenberg, Rand, and Dreber (2012) benefit/cost formulation, where their benefit/cost ratio parameter (b/c) is given by $(1+x)/x$ here.

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

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

the other $N-1$ players *jointly* coordinate on the collusive strategy,

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

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

As in 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, $x\Delta\pi$, and the continuation gain from collusion, $\frac{\delta}{1-\delta}\Delta\pi$, obtained with probability:

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

where the right-hand-side 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 each individual other player attempts to collude. 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.[22]

Assuming perfect correlation for the other $N-1$ agents, $Q(N) = p$, so the critical belief is:

$$p^\star_{\text{Corr.}}(x) = \frac{1-\delta}{\delta}x. \tag{4}$$

In contrast, when beliefs are fully independent, $Q(N) = p^{N-1}$, so the critical belief is:

$$p^\star_{\text{Ind.}}(x,N) = \left(\frac{1-\delta}{\delta}x\right)^{1/N-1} \equiv \left(p^\star_{\text{Corr.}}(x)\right)^{1/N-1}. \tag{5}$$

Note that the correlated measure in Equation (4) is not a function of N, while the independent measure in Equation (5) increases in N. The two measures are identical only in the RPD case at $N = 2$.[23]

We focus on these two extreme cases of full independence and perfect correlation because: (i) They allow us to produce an experimental design that has stark behavioral predictions; and (ii) Both measures are simple to compute in settings beyond our environment.[24]

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

[23] Notice that both extensions of the measure capture beliefs over supergame strategies (full specifications of what action to play at any history). For the two strategies underlying the basin measures, actions are perfectly correlated in all rounds after the first one. For instance, consider α_{Grim}. Either all N players successfully coordinate on cooperation, or after an observed failure in round one, the punishment path is triggered with all N players choosing defect in all subsequent rounds. As such, the independent and correlated models will only differ in the potential for correlation in the very first round.

[24] 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 discuss this alternative in further detail and estimate the correlation parameter from the data.

TABLE 1. Experimental design

Panel A. Stage-game payoffs	$X = \$9$		$X = \$1$	
	$\sigma(a_{-i}) = S$	$\sigma(a_{-i}) = F$	$\sigma(a_{-i}) = S$	$\sigma(a_{-i}) = F$
Cooperate, $\pi_i(C, \sigma)$	\$20	\$2	\$20	\$10
Defect, $\pi_i(D, \sigma)$	\$29	\$11	\$21	\$11
Panel B. All-D Basin Size	$X = \$9$ $(x = 1)$		$X = \$1$ $(x = 1/9)$	
	$N = 2$	$N = 4$	$N = 4$	$N = 10$
Correlated, $p^\star_{\text{Cor.}}(x)$	$p^\star_0$ [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]
Independent, $p^\star_{\text{Ind.}}(x, N)$	$p^\star_0$ [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 prediction	$p^\star_0$	Marginal effect from basin:		
	[0.33]	Increase to [0.69]	Decrease to [0.04]	
Initial cooperation, $t = 1$	0.50	−0.26	+0.35	
Ongoing cooperation, $t > 1$	0.37	−0.21	+0.50	

Note: Meta-study predictions in Panel (C) correspond to the estimated relationship $\hat{C}_{\text{Meta}}\left(p^\star\right)$ illustrated in Figure 1.

3. Experimental Design

Based on the basin measures derived in Equations (4) and (5), our experimental design is founded on two competing hypotheses:

Correlated-Basin/Null-effect Hypothesis. *Cooperation decreases as we increase the cost of cooperation 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 $p^\star_{Ind.}$.*

In Panel (A) of Table 1 we illustrate our first treatment dimension, which manipulates the payoff cost of cooperating $X = x\Delta\pi$, where $\Delta\pi = \$9$. The two values of X—a high temptation of \$9 (a normalized temptation of $x = 1$) illustrated on the left, and a low temptation of \$1 (a normalized temptation of $x = 1/9$) illustrated on the right—lead to two payoff environments over own actions and signals.[25,26]

We also vary the number of players N as indicated in the column headings of Panel (B) in Table 1. The two rows of Panel (B) illustrate how choices regarding X and N influence

[25]See Figure E.1 in Online Appendix E for representative lab screenshots.

[26]Henceforth, we will focus on the payoff cost of cooperating X rather than the normalized parameter x.

the basin-size measures of strategic uncertainty under the correlated and independent extensions. In total, we create four treatments, each defined by an (N, X)-pair.

To independently manipulate each basin-size measure, we select (N=2; X=\$9) as our baseline treatment. When comparing (N=2; X=\$9) with (N=4; X=\$9) we keep the correlated-basin measure constant at $p_0^\star = 0.33$ and increase the independent-basin measure to $p_0^\star + \Delta p_{\text{Ind.}}^\star = 0.69$. Next, when comparing ($N$=2; X=\$9) with (N=4; X=\$1) we keep the independent-basin measure constant at $p_0^\star = 0.33$ and lower the correlated-basin measure to $p_0^\star - \Delta p_{\text{Corr.}}^\star = 0.04$. Finally, when comparing (N=4; X=\$1) with (N=10; X=\$1) we keep the correlated-basin measure constant at $p_0^\star - \Delta p_{\text{Corr.}}^\star = 0.04$ and increase the independent-basin measure to $p_0^\star + \Delta p_{\text{Ind.}}^\star = 0.69$.

By varying the primitives X and N, our 2×2 design yields four pairs of correlated/independent basin measures:[27]

$$\left(p_{\text{Corr.}}^\star, p_{\text{Ind.}}^\star\right) \in \left\{p_0^\star, p_0^\star - \Delta p_{\text{Corr.}}^\star\right\} \times \left\{p_0^\star, p_0^\star + \Delta p_{\text{Ind.}}^\star\right\} := \left\{0.33, 0.04\right\} \times \left\{0.33, 0.69\right\}.$$

Using the probit-model estimates illustrated in Figure 1 we can provide a *quantitative* prediction $\hat{C}_{\text{Meta}}(p^\star)$ for the cooperation rate under each basin-size measure $p^\star$. These predictions are outlined in Panel (C) of Table 1. The first column presents the initial and ongoing cooperation rates expected at $p^\star = 0.33$. The next two columns indicate the expected treatment effect resulting from a shift in strategic uncertainty from $p^\star = 0.33$ to either $p_0^\star - \Delta p_{\text{Corr.}}^\star = 0.04$ or $p_0^\star + \Delta p_{\text{Ind.}}^\star = 0.69$.

For illustration, consider the predictions under the standard independence-based extension. In the (N=2; X=\$9) and (N=4; X=\$1) treatments, the independent basin size is 0.33, and it increases to 0.69 in (N=4; X=\$9) and (N=10; X=\$1). If the strategic uncertainty relationship estimated from the two-player RPD meta-data were perfectly extrapolated 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.[28]

Note that our hypotheses do not specify whether initial cooperation, ongoing cooperation, or both are expected to align with the behavior of many players. Initial cooperation captures intentions to coordinate (with beliefs as a driver), while ongoing cooperation reflects successful coordination (with the interaction of the beliefs as a driver). In the case

[27] We chose $\Delta\pi$ = \$9 and $\delta = 3/4$ for simplicity of presentation to participants (i.e., integer values for both N and X). The precise design over the basin measures is as follows:

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

[28] 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 (N=2; X=\$9) and (N=4; X=\$9) treatment pair and the (N=4; X=\$1) and (N=10; X=\$1) pair. Based on the estimated relationship from the meta-study, this implies an increase in the initial cooperation rate of 35 percentage points and an increase in the ongoing cooperation rate of 50 percentage points, and null effects within each pair for fixed X.

of the two-player RPD, Figure 1 shows that the basin size closely follows both cooperation measures, making it challenging to disentangle the effects. By introducing variation in N, we add a channel that might help us distinguish between initial and ongoing cooperation, and identify which measure is better predicted by basin-size models.

Experimental Specifics. In our main experiments, we used a between-subject design over the four treatments outlined in Table 1. Participants for each treatment were recruited from the undergraduate population at the University of Pittsburgh, and each took part in exactly one session. We recruited a total of 584 participants, 252 for the first four main treatments, and 332 for the extensions discussed in Section 5. Each treatment comprised three sessions, aiming to enroll a minimum of 20 participants per session, except for the (N=10; X=\$1) treatment, for which we conducted two sessions with 30 participants each.[29] Sessions lasted between 55 and 90 minutes, and participants received 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.[30] The participants were randomly and anonymously matched in the 20 supergames in a stranger design.[31] The 20 supergames were divided into two parts of ten supergames.[32] 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.[33]

4. Results

We begin this section by describing the aggregate cooperation and success rates at the treatment level. Then, we proceed to discussing inferential tests of our two basin-extension hypotheses.

4.1. Main Treatment Differences.

In Table 2 we present average cooperation and success rates by treatment, for both initial and ongoing cooperation. Averages are computed for the last five supergames to capture late-session behavior, where subjects have accumulated experience in the environment.[34] Overall, the results reveal large shifts in cooperation as we manipulate the cost of cooperation X and/or the number of players N.

[29] While our design called for sessions to have at least 20 participants, we allowed sessions to grow by an additional group of size N depending on realized show ups. For (N=10; X=\$1) we instead opted to recruit 30 participants for each session so that we had three groups in each supergame.

[30] We employed common draws to maintain consistent supergame lengths at the session level for each treatment.

[31] All participants received both written and verbal instructions regarding the task and payoffs. Detailed instructions are available for interested readers in the Online Appendix F.

[32] Participants were provided with complete instructions for the first part and were informed that instructions for the second part would be given after completing supergame ten. For the four between-subject treatments outlined in Section 3, part two was identical to part one. In later sections of the paper, we describe an additional set of treatments with a within-subject change across the two parts. The decision to have two identical parts here enables direct comparisons in first-half play.

[33] This method, developed in Sherstyuk, Tarui, and Saijo (2013), is employed to induce risk neutrality across supergame lengths. Another advantage of this design choice is the absence of wealth effects within a supergame, where history serves only as an instrument for the future play of others.

[34] Including all rounds yields similar results (see Table A.1 in Online Appendix A).

TABLE 2. Cooperation and success rates across all supergames

Action and signal rates	$X = \$9$		$X = \$1$	
	$N = 2$	$N = 4$	$N = 4$	$N = 10$
	Cooperation			
Initial	0.503 (0.058)	0.035 (0.017)	0.792 (0.042)	0.357 (0.055)
	⟨0.50⟩	⟨0.24⟩	⟨0.50⟩	⟨0.24⟩
	|0.50|	|0.50|	|0.85|	|0.85|
Ongoing	0.450 (0.055)	0.006 (0.003)	0.409 (0.050)	0.184 (0.048)
	⟨0.37⟩	⟨0.16⟩	⟨0.37⟩	⟨0.16⟩
	|0.37|	|0.37|	|0.87|	|0.87|
	Success			
Initial	0.503	0.000	0.578	0.000
Ongoing	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 in parentheses. For comparison, we provide the meta-study prediction for the independent basin measure $\hat{\mathcal{C}}_{\text{Meta}}\left(p^{\star}_{\text{Ind.}}\right)$ in angle-brackets, $\langle\cdot\rangle$, and the prediction $\hat{\mathcal{C}}_{\text{Meta}}\left(p^{\star}_{\text{Cor.}}\right)$ for the correlated basin measure in vertical bars, $|\cdot|$ (cf. Panel (B) of Table 1 for details).

The first row in Table 2 provides a summary of initial cooperation. The 50.3 percent initial cooperation rate in our (N=2; X=\$9) treatment closely aligns with the 50.0 percent rate predicted by the meta-study. However, maintaining the cooperation cost at $X = \$9$ and doubling the group size to four virtually eliminates cooperative behavior, resulting in an initial cooperation rate of 3.5 percent in (N=4; X=\$9). In the first round of our low-temptation scenarios ($X = \$1$), groups of $N = 4$ exhibit highly cooperative behavior (79.2 percent) while groups of $N = 10$ display moderate cooperation (35.7 percent).

The next two rows in Table 2 summarize the ongoing cooperation rates. Across all treatments we observe a decrease in ongoing cooperation compared to initial cooperation. The most substantial quantitative drops are evident in the $X = 1$ treatments, where ongoing cooperation rates are halved in comparison to the initial cooperation rates.[35]

The last two rows in Table 2 present the fraction of rounds in which a success signal was observed.[36] The patterns that emerge for success rates are similar to those seen for

[35] In Online Appendix A, Table A.2 provides a more detailed breakdown of ongoing cooperation rates based on the observed history from the previous round. The findings suggest that individual cooperation is heavily conditioned on successful coordination in the preceding round. Interestingly, participants are markedly more forgiving after failed cooperation at $X = \$1$ than $X = \$9$.

[36] A success at the individual level requires joint cooperation from the other $N - 1$ participants. Success is directly linked to group-level cooperation, where the *expected* success rate, given an independent cooperation rate p, is p^{N-1}. In two-player games, the success rate is identical to the cooperation rate. Expected

ongoing cooperation, though with starker quantitative effects. Although success is the modal signal in the (N=2; X=\$9) and (N=4; X=\$1) treatments, in the (N=4; X=\$9) and (N=10; X=\$1) treatments there are *no* successes at all.[37]

The results presented in Table 2 speak to both the correlated- and independent-basin hypotheses. The collected evidence does not favor the correlated-basin hypothesis. For both the initial and ongoing cooperation rates we observe large changes in behavior as we move N for either fixed value of X. On the other hand, the data support the independent-basin predictions regarding directional shifts in both initial and ongoing cooperation rates as we vary X or N in isolation. However, for initial cooperation, we observe deviations from perfect substitution of strategic uncertainty as X and N move in opposing directions. The independent-basin hypothesis predicts a null effect when we compare either (N=2; X=9) to (N=4; X=1) or (N=4; X=9) to (N=10; X=1). But instead we observe substantial effects in the comparisons of initial cooperation, with 29 and 35 percentage point differences, respectively.

Meanwhile, ongoing cooperation in the (N=2; X=\$9) and (N=4; X=\$1) treatments are relatively close, at 45.0 and 40.0 percent, respectively. This finding aligns qualitatively with the independent-basin prediction of no difference due to perfect substitution of strategic uncertainty. For a similar comparison, however, we still note an 18 percentage point difference between (N=4; X=\$9) and (N=10; X=\$1). The difference is driven by a very stark finding of near-zero cooperation in (N=4; X=\$9). As we outline further below this is the main deviation in our data relative to the meta-study prediction.[38]

4.2. **Evaluation of the Independent- and Correlated-Basin Hypotheses.** Table 3 presents a direct statistical evaluation of our two competing hypotheses. The results are

success rates (given the cooperation rate and independent matching) in the initial round are, in the order of Table 2 columns: 0.503, 4.2×10^{-5}, 0.497 and 9.5×10^{-5}.

[37] As success directly aggregates individual-level cooperation, we refrain from reporting standard errors (where standard errors also cannot be calculated in cases where we have no variation). Nevertheless, the pronounced nature of the effect, in alignment with predictions for the independent-basin hypothesis, clearly illustrates the underlying economic relationship.

[38] Regarding inference, Online Appendix A presents two additional tests: (i) Tests examining the cardinal predictions from the meta-study, and (ii) Tests assessing the ordinal predictions across treatments. In the first set of tests (Table A.3), we evaluate the predicted cooperation levels $\hat{C}_{\text{Meta}}\left(p^{\star}\right)$ from the meta study. Our findings reveal the rejection of cardinal predictions for both initial and ongoing cooperation, irrespective of whether the basins are independent or correlated ($p < 0.001$ all F-tests). However, closer examination indicates that the meta-study aligns more closely with predicted behavior in two specific scenarios: (i) ongoing cooperation, and (ii) predictions using the independent basin. For ongoing cooperation under the independent-basin prediction, we do not reject the predicted cooperation levels in (N=2; X=\$9), (N=2; X=\$1) and (N=10; X=\$1) ($p > 0.150$ all comparisons, jointly $p = 0.400$). The only exception is the (N=4; X=\$9) treatment ($p < 0.001$), where the meta-study predicts a cooperation rate of 16 percent, but the observed rate is virtually zero.

For the ordinal comparisons, Table A.4 presents the six possible treatment comparisons in our design, along with the ordinal prediction from each basin notion. For ongoing (initial) cooperation, the independent basin correctly organizes five (four) of the six comparisons. While it orders the four treatments where a difference is predicted (for both ongoing and initial cooperation), the independent basin fails one of the two null tests for ongoing cooperation and both null tests for initial cooperation. Meanwhile, the correlated basin makes three successful predictions out of six, one incorrect directional prediction, and fails both null tests.

TABLE 3. Basin-effect decomposition: Main treatments

Experimental results	$p_0^\star$	Marginal effect in cooperation from:	
	[0.33]	Independent basin increase to $p_0^\star + \Delta p^\star_{\text{Ind.}} = [0.69]$	Correlated basin decrease to $p_0^\star - \Delta p^\star_{\text{Corr.}} = [0.04]$
Initial	0.464	−0.395	+0.357
	(0.058)	(0.048)	(0.053)
	⟨0.50⟩	⟨−0.26⟩	⟨+0.35⟩
Ongoing	0.366	−0.293	+0.115
	(0.051)	(0.051)	(0.061)
	⟨0.37⟩	⟨−0.21⟩	⟨+0.50⟩

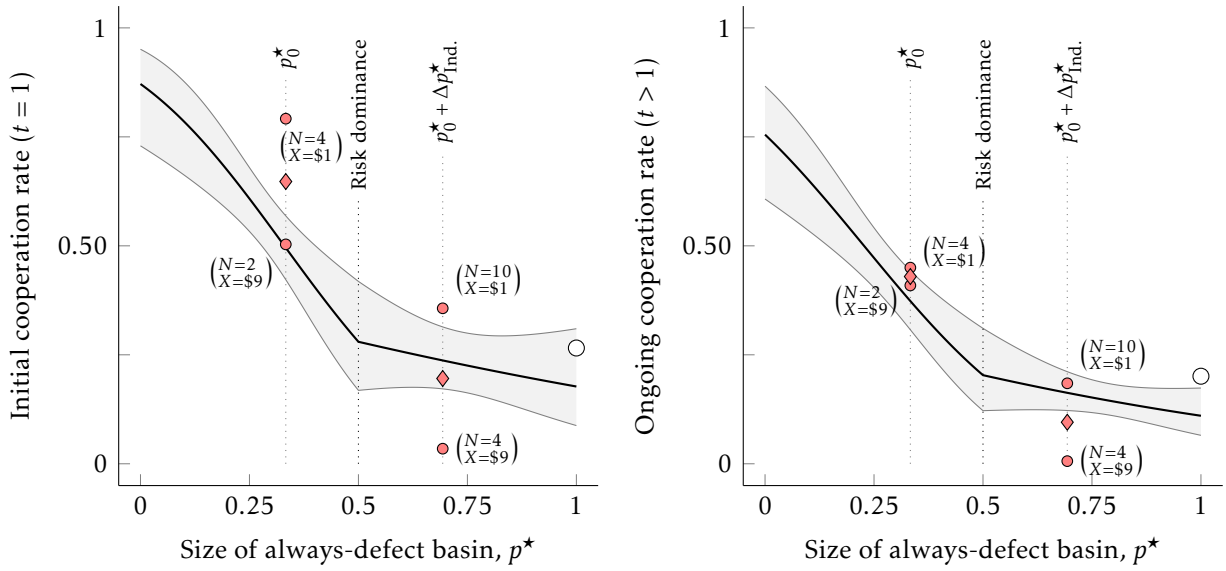
Note: Results are calculated using data from the last-five supergames. The cooperation decomposition runs two probits, one for initial, and one for the ongoing cooperation, with subject-clustered standard errors in parentheses. Right-hand-side variables are a constant and two dummies, one for a low-correlated–basin treatment (X = \$1, both N values), one for a high-independent–basin treatment (X = \$9/N = 4 and X = \$1/N = 10). Meta-study predictions given in angle brackets, ⟨·⟩, below each result.

based on probit regressions that examine subjects' cooperation decisions, with dummy variables corresponding to the 2×2 design outlined in Table 1 Panel (B). The dummy covariates include an indicator for the predicted $\Delta p^\star_{\text{Cor.}}$ decrease from the correlated basin (as we decrease X for any N) and an indicator for the predicted $\Delta p^\star_{\text{Ind.}}$ increase from the independent basin (as we increase N holding X constant).

Each row in Table 3 presents results from a distinct estimation, one focusing on initial cooperation and the other on ongoing cooperation. The first $p_0^\star$ column displays the estimated cooperation rate when both dummy variables are zero, representing the RPD cooperation rate with a basin size of $p_0^\star = 0.33$. The following two columns illustrate the estimated marginal effect on the cooperation rate for a shift in each basin measure, while holding the other basin constant. If either of the two basin hypotheses fully explained behavior, we would expect a significant estimate for the dummy on that basin shift and an insignificant effect on the other.

The estimation parallels the probit model we run to recover the meta-study prediction $\hat{C}_{\text{Meta}}(p^\star)$. The estimated baseline cooperation rates is for an RPD with $p_0^\star = 0.33$, where this baseline closely matches the meta-study prediction. Specifically, while the meta-study predicts the initial (ongoing) cooperation rate of 50.0 (37.0) percent, our data at $p_0^\star = 0.33$ reflects a very similar (and statistically indistinguishable) rate of 46.4 (36.6) percent. To illustrate this alignment, Figure 2 depicts the fitted relationships from the meta-study overlaid with our results from the four treatments using the independent-basin size on the horizontal axis. Filled circles represent individual treatments and filled diamonds treatments pooled over each value for the independent-basin measure. While there is notable divergence for initial cooperation, Figure 2 demonstrates quantitatively similar results for ongoing cooperation.[39]

[39]In Online Appendix A, Figure A.1 presents analogous results organized under the correlated-basin model. The figure illustrates much poorer organization of the data, both in terms of relative treatment comparisons and quantitatively.



(A) Initial cooperation

(B) Ongoing cooperation

FIGURE 2. Cooperation under 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.

To test our two competing hypotheses, we focus on the second and third columns of Table 3. In the scenario where the independent-basin measure comprehensively captured all pertinent aspects of behavior, we would expect a statistically significant and negative estimate in the second column, coupled with a zero effect in the third column. Meanwhile, if the correlated-basin captured behavior, we would expect a significantly negative effect in the third column and a zero effect in the second column.

In terms of initial cooperation, our results reveal that modifications to both basin measures yield significant effects ($p < 0.001$). Although the estimated effects are comparable in magnitude, they exhibit opposite directions, which is consistent with our predictions. Given that neither effect prevails over the other, we infer that both X and N contribute unique information to the prediction of initial cooperation, and this information is not entirely captured by either basin measure independently.

Regarding ongoing cooperation, the increase from the independent basin shift is negative and significant ($p < 0.001$) and it is quantitatively close to the meta-study prediction. Meanwhile the estimated effect for the correlated basin is much smaller in magnitude and significant only at the 10 percent level ($p = 0.061$) after controlling for the independent basin. The small effect attributed to the correlated basin in our estimation could also be associated with other-regarding preferences. Part of the differences in cooperation are

driven by a higher fraction of unconditional cooperators at $X = \$1$ compared to $X = \$9$.[40] Because variation in X is associated with shifts in the correlated-basin value (invariant to N), this presents a difficulty in interpretation for the small positive effects for the correlated basin. While this could be driven by belief correlation, it could also be driven by other-regarding preferences.[41]

In addition to the qualitative directional effects, we observe that the quantitative shifts in ongoing cooperation under the independent-basin measure closely align with the predicted effects expected from the meta-study.[42] Specifically, the latter predicts a drop of 21 percentage points (last row in Table 3) in ongoing cooperation when the size of the basin increases from $p_0^\star = 0.33$ to 0.69, and our estimates indicate a decrease of 29 percentage points.[43]

The main difference in ongoing cooperation between our data and the independent-basin predictions from the meta-study arises from extreme behavior in the (N=4; X=\$9) treatment, where cooperation is essentially at the boundary. A two-player RPD with a basin size of $p^\star = 0.69$ has a predicted ongoing cooperation rate of 16.0 percent, and this prediction remains relatively stable for all other values of the basin where grim is risk-dominated (with 11.0 percent cooperation predicted at $p^\star = 1$). The very low late-session cooperation rates in (N=4; X=\$9) can be explained by considering the large payoff reduction from cooperation, coupled with unrelentingly negative feedback. That is, out of 1,145 supergame-rounds in this treatment where a group of four attempted to coordinate, only a single group was successful for a single round. However, though the observed level deviates from the prediction, a broader interpretation of the basin continues to hold: conditional cooperation is not expected when always-defect is risk dominant ($p^\star > \frac{1}{2}$), while the level of cooperation is predictably decreasing when grim is risk dominant ($p < \frac{1}{2}$).

Finally, we attempt to measure *how much* correlation is necessary to rationalize the data. To achieve this, we allow beliefs to be a convex combination of the independent and correlated models. With proportion σ, the $N-1$ agents collectively choose grim with

[40] In Tables G.1 and G.2 of Online Appendix G we present strategy frequency estimates from the first and last seven supergames, where we identify the fraction of choices that are consistent with unconditional cooperation in each treatment.

[41] By design, subjects receive coarse feedback in our environment. For example, a failure signal in a treatment with $N = 4$ indicates that at least one of the other three members did not cooperate. Such coarse feedback minimizes the possibility that early feedback is exacerbated as N increases. If we had provided subjects with disaggregated feedback, a treatment with $N = 10$ would provide effective feedback on everyone else in the session after a few supergames. This could translate into early choices having more of an impact in later choices in treatments with high N. While this type of effect is muted given our coarse feedback, it is still possible for it to arise. We do not see any clear evidence in this direction, but a definitive test would require treatments with a turnpike, perfect stranger designs, and/or larger sessions.

[42] Our measures of equilibrium selection aim to capture strategic uncertainty in a setting that differs from the two-player RPD. Discovering that our results align with findings in the two-player RPD literature is valuable. It implies that a measure of strategic uncertainty might be a robust predictor of collusion, irrespective of the specific details of the environment.

[43] On the contrary, the meta-study predicts an *increase* of 50 percentage points in initial cooperation when the size of the correlated basin decreases from $p_0^\star = 0.33$ to 0.04, and our estimates indicate an increase of roughly 11.5 percentage points, after controlling for the independent-basin effects.

probability p and always defect with probability $1-p$; with proportion $1-\sigma$, each agent independently chooses grim with probability p and always defect with probability $1-p$. Under this specification, the probability that the other $N-1$ players coordinate is given by

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

with the critical belief denoted by $p^\star(\sigma, x, N)$. The additional parameter σ nests the two extremes: $\sigma = 0$ for full independence, $\sigma = 1$ for perfect correlation.[44] Looking at the best fitting parameter, for (I)nitial cooperation, we estimate $\hat{\sigma}_I = 0.091$ (with standard error of 0.005), while the comparable estimate for (O)ngoing cooperation is $\hat{\sigma}_O = 0.031$ (with the standard error of 0.010). Both estimates are statistically different from zero ($p < 0.001$ and $p = 0.014$ for initial and ongoing, respectively). The conclusion from the exercise is that the estimated degree of correlation needed is quantitatively small.

We now summarize our main results:

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

Result 2 (Correlated-Basin Measure)**.** *Our data are inconsistent with the predictions from the correlated-basin hypothesis, for both initial and ongoing cooperation. In particular, where 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 required to rationalize the results is small.*

5. Beyond the main results

Our analysis thus far has abstracted away other features of the coordination problem to focus on the pure effects of the stage-game primitives. In this section, we introduce additional treatments to study possible limitations of the strategic-uncertainty model in predicting changes in equilibrium selection.

5.1. **Between vs. Within Identification.** Here, we explore the extent to which behavior after a policy change might not align with corresponding changes to the basin. Consider a policy change that alters the underlying strategic environment—the temptation and/or the number of players for our experiments—and therefore the collusive prediction. The underlying idea from the model is that beliefs about others' strategies drive behavior. But if beliefs are shaped by experiences formed prior to the policy change, a strategic uncertainty model might fail to predict behavioral changes *within* population. In the previous section, our treatments employed a *between*-subjects design, where identification relied on comparisons of late-session behavior between different populations, each with experience in a fixed strategic setting. In our modified treatments, we investigate the effects on collusive behavior following a change in the number of players N *within* the same session.

[44] Full details of the estimation procedure are provided in Online Appendix D.

We examine two within-session treatment shifts: one with (N=2; X=\$9) in the first half of the session, and (N=4; X=\$9) in the second half; and the reverse treatment shift with (N=4; X=\$9) in the first half, and (N=2; X=\$9) in the second. Given that we keep the temptation parameter constant at $X = \$9$, we label these two treatments as $2 \rightarrow 4$ and $4 \rightarrow 2$, respectively. In both treatments, the change in N comes as a surprise: subjects are aware of a second part, but they do not receive instructions for the second part until the end of supergame ten. In terms of the independent-basin model, this creates a shift across the session from a low basin size of 0.33 when $N = 2$ to a high basin size of 0.69 when $N = 4$. In particular, this is a shift in N that generates a substantial treatment effect for the between-subject treatments.

In Figure 3(A) we present the initial cooperation rates by supergame and type of treatment. The between-subject treatments with $N = 2$ and $N = 4$ are indicated by gray dashed lines, 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.

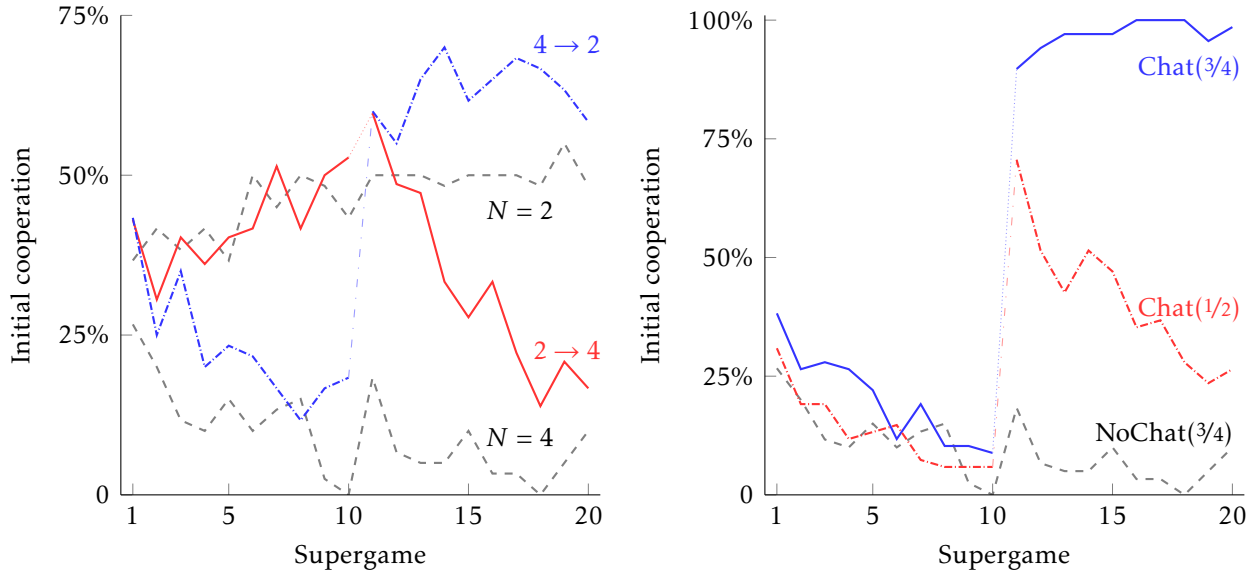
The figure illustrates a substantial between-subject effect, with more cooperation in $N = 2$ than $N = 4$ across all twenty supergames. Pooling the between and within treatments in supergames 6–10, we arrive at an initial cooperation rate of 47.4 percent for $N = 2$ and 13.9 percent for $N = 4$.[45] As we move into supergames 11–20 for our within-subject treatments, the strategic environment changes, specifically, the number of players an individual is matched with either decreases or increases. For the $2 \rightarrow 4$ treatment (the solid red line), initial cooperation remains high as N increases. While there is no immediate drop in cooperation, we observe that as participants gain experience at $N = 4$, the cooperation rate continues to fall, reaching 16.7 percent by supergame 20. In contrast, moving from $N = 4$ to $N = 2$ (the blue dash-dotted line), we observe an immediate jump in cooperation as N decreases: the initial-round cooperation in the last supergame with $N = 4$ is 18.3 percent, but after the reduction to $N = 2$ the cooperation rate immediately jumps up to 60.0 percent. This jump in cooperation as N decreases is then sustained across the remaining supergames, with 58.3 percent cooperation by supergame 20.

Inspecting the results illustrated in Figure 3(A) it is clear that there is minimal evidence for the hypothesis that selected equilibrium is sticky to a within-population shift in N. Despite exposure to a prior environment in the first half of the session, 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.[46]

Overall, we find that:

[45] When testing differences in initial cooperation rates in supergames 6–10 within each N (comparing between and within sessions with identical treatment up to this point), we find $p = 0.150$ for $N = 2$ and $p = 0.981$ for $N = 4$ using t-tests. A joint test across both values of N yields $p = 0.353$.

[46] In Online Appendix B, we offer a more detailed like-with-like comparison of the between-subject and within-subject results. These additional findings do not indicate differences with the between-subject results as we move from $N = 2$ to $N = 4$. However, contrary to the hypothesis that the selected equilibrium is sticky, we observe a significant *increase* in responsiveness to changes in N (relative to the between-subject treatments) in the $4 \rightarrow 2$ treatment.



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

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

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

Result 3 (Between vs. Within)**.** *Changing N within subjects as opposed to between does not substantially alter the qualitative results. We find no evidence that the selected equilibrium is sticky in the long run as we shift a primitive within the population.*

5.2. **Explicit Coordination.** In this second set of extension treatments, we examine the strategic-uncertainty mechanism underlying the basin-size model. Specifically, we study the extent to which our results may be influenced by explicit coordination, as free-form communication can diminish strategic uncertainty by enabling players to reveal their strategic intentions.[47] This analysis is motivated by an empirical finding indicating that instances of detected collusion in the industry often originate from explicit collusion—despite the illegality of such meetings.[48]

We design our "chat" treatments by modifying an environment with the least-collusive outcomes, represented by the ($N{=}4; X{=}\$9$) treatment. In our first chat treatment, Chat(3/4), the initial ten supergames replicate the conditions of the ($N{=}4; X{=}9$) treatment. However, in supergames 11–20, we introduce pre-supergame chat between all four players. The second chat treatment, Chat(1/2), mirrors Chat(3/4) in terms of timing of when the chat is introduced but reduces the continuation probability to $\delta' = 1/2$ (this continuation is used across all twenty supergames). The Chat(1/2) treatment keeps constant the stage-game payoffs and number of players, but lowers the continuation probability δ to the point that

[47] Our design is not tailored to pinpoint the exact channel through which strategic uncertainty is reduced. It could be that messages convey the opponent's reasonableness and understanding of the game's tensions. Alternatively, messages might not directly convey information on rationality but simply reduce social distance, making it easier to trust the other player.

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

the grim-trigger strategy is only a knife-edge subgame perfect equilibrium, requiring a critical belief of $p^\star(\delta') = 1$ on the other three players cooperating (and so Equations (4) and (5) also coincide). Therefore, Chat(1/2) serves as a litmus test for whether explicit coordination can implement outcomes that are *not* supportable as a robust equilibrium (that is, with arbitrarily small trembles in others' behavior).

In Figure 3(B) we depict initial cooperation rates by supergame, using the (N=4; X=\$9) treatment as a baseline, here labeled NoChat(3/4).[49] The figure highlights an unambiguous result on the power of explicit coordination under $\delta = 3/4$: providing pre-play chat takes the near-zero initial cooperation rate in NoChat(3/4) to almost full cooperation (99.8 percent, with 80.0 percent ongoing cooperation) in Chat(3/4). Such high levels of cooperation with communication are inconsistent with the predictions of the independent-basin model. Therefore, once explicit coordination devices are allowed for and strategic uncertainty dissipated, our independent basin-size model becomes redundant. That is, the independent model is only intended for implicit/tacit coordination.

However, as we shift the continuation probability to the $\delta' = 1/2$ boundary, even with pre-play communication, participants find it challenging to sustain cooperation. While initial cooperation is substantially higher than the baseline without chat (30.0 percent), ongoing cooperation falls to 4.4 percent (with an ongoing success rate of 0.2 percent). As such, our second chat treatment indicates that for explicit communication to play a role, collusion needs to be at least supportable as non-knife edge equilibrium outcome.

Result 4 (Implicit vs. Explicit)**.** *In a multi-player setting, where implicit cooperation results in near-zero cooperation, explicit coordination leads to very high levels of cooperation. However, in the limiting case, where cooperation is a knife-edge subgame perfect equilibrium outcome, even pre-play chat fails to support cooperation.*

5.3. **Easing Requirements for a Success.** In our prior treatments, we find a clear reduction in coordination on the efficient group outcome as we increase N. However, in our experiments, as we increase the number of players to four, we are mechanically making it harder to coordinate, as requiring four cooperators out of four is more stringent than requiring two cooperators out of two. In our final robustness exercise, we explore an alternative construction of a four-player game. Specifically, we make it mechanically *easier* to coordinate by allowing for an efficient group-wide outcome even if only two out of four players cooperate. At first sight, relaxing the bar for success in this way suggests that groups will successfully coordinate at much higher rates. However, as we will show, the basin of attraction makes the reverse prediction. The reason for this counter-intuitive prediction is that making it mechanically easier to attain a cooperative outcome introduces a new coordination challenge: *who* will cooperate and *who* gets to freeride? In fact, this final extension adds so much strategic uncertainty that our new treatment has a full basin for always defect. As such, this robustness treatment provides a stark test of the basin-of-attraction notion as we increase N.

[49] Late-session cooperation and success rates (in supergames 16–20 with subject-clustered standard errors) are provided in Table A.5 in Online Appendix A.

We hold constant the RPD's 2×2 stage-game representation but weaken the requirement for a success signal to the case where $M-1$ or more other players cooperate, with $1 \leq M \leq N$. Defining the count of cooperative actions for the other players as $\text{CoopCount}_{N-1}(a_{-i}) := \sum_{j \neq i} \mathbf{1}_{a_j = C}$, an agent's signal is given by:

$$\sigma(a_{-i}; M, N) = \begin{cases} S & \text{if } \text{CoopCount}_{N-1}(a_{-i}) \geq M-1, \\ F & \text{otherwise,} \end{cases} \tag{6}$$

where for our original treatments $M = N$.

Easing a requirement for success makes it structurally easier to generate a group-wide success. Define $Q_p(M,N)$ as the probability of having M cooperators among N players, where each player chooses to cooperate with probability p. For any fixed $p \in (0,1)$ we have:[50]

$$Q_p(M,N) > Q_p(M,M) > Q_p(N,N). \tag{7}$$

Although it is mechanically easier to achieve joint success for any fixed cooperation rate p, weakening the success requirement introduces additional strategic uncertainty. If an individual believes that the other players select a conditionally cooperative strategy with probability p, the agent will focus on the following pivotal probability:

$$q_p(M,N) = \Pr\{M-1 \text{ from } N-1 \text{ others choose } \alpha_{\text{Grim}}; p\}.$$

In all other situations the agent's action will not affect the long-run outcome: (i) There will either be fewer than $M-1$ others cooperating (with a miscoordination cost of $(1-\delta)x$ to the agent); or (ii) There will be M or more cooperators and group-wide success will be guaranteed (with a miscoordination cost of x to the agent for the unnecessary coordination).[51] Therefore, best-responding agents will only cooperate at intermediate values of p. The basin of attraction for always-defect will either be full (all $p \in [0,1]$) or spread across two disjoint regions ($p \in [0, \underline{p}^\star] \cup [\overline{p}^\star, 1]$). A strategic agent will not want to conditionally cooperate: (i) When others cooperate with low probability, $p \in [0, \underline{p}^\star]$, coordination is likely to fail even if they cooperate; or (ii) When others cooperate with high probability, $p \in [\overline{p}^\star, 1]$, coordination is likely to succeed even if they defect. For this reason, if $1 < M < N$, perfectly correlated beliefs result in an always-defect basin size of one for any $x > 0$ and $\delta \in (0,1)$.

Keeping the other parameters constant at $X = \$9$ and $\delta = 3/4$, our final treatment requires $M = 2$ cooperators from the group of $N = 4$ for joint success (which we will call the *two-from-four* treatment). While this change makes successes mechanically easier to obtain

[50] The second inequality is just $p^M > p^N$ which follows as $M < N$. The first inequality comes from decomposing the probability for a group-wide success to the chance the first M players jointly cooperate (meaning it must succeed) and a remaining positive probability, so $Q_p(M,N) = Q_p(M,M) + (1 - Q_p(M,M))\Pr\{M \text{ cooperate from } N \,|\text{First } M \text{ not all cooperators}\}$.

[51] The condition for grim to be a best response is:

$$\Pr(\text{Exactly } M-1 \text{ choose grim}) \geq x\frac{(1-\delta)}{\delta} + x\Pr(\text{More than } M-1 \text{ choose grim}).$$

For full derivation see Online Appendix C.

TABLE 4. Basin-effect decomposition: Two-from-four treatment

Group-wide success	Two-from-four	Compared to:	
		$(N{=}2; X{=}\$9)$	$(N{=}4; X{=}\$9)$
All rounds	0.255	0.360 (p=0.006)	0.000 (p<0.001)
Initial rounds	0.302	0.266 (p=0.006)	0.000 (p<0.001)
Ongoing rounds	0.222	0.330 (p=0.006)	0.000 (p<0.001)

Note: Results are calculated using data from the last-five supergames. The values in parentheses correspond to p-values testing differences between the two-for-four treatment and each of the reference treatments. For the $(N{=}2; X{=}\$9)$ comparisons we use standard tests of proportion; however, because we have no outcome variation in $(N{=}4; X{=}\$9)$, for those tests we use likelihood ratio tests over binomial probabilities.

than in the $(N{=}2; X{=}\$9)$ and $(N{=}4; X{=}\$9)$ treatments, our basin measure of strategic uncertainty (either independent or correlated) makes the opposite prediction. In fact, at $X = \$9$ and $\delta = 3/4$, the coordination problem is exacerbated to such a degree that weakening the requirement for a success leads to theoretically full basin for always defect.[52,53]

Results from the two-from-four treatment yield a cooperation rate of 22.7 percent across all rounds, and a group-wide success rate of 25.5 percent, compared to a group-wide success rate in the baseline $(N{=}2; X{=}\$9)$ treatment of 36.0 percent. Hence, as predicted by our basin calculations, easing the requirement for success significantly reduces successful coordination ($p = 0.006$).[54]

This directional success in a counter-intuitive direction certainly suggests that part of the additional difficulty in coordinating is captured by the basin. However, the basin measure fails to order the success rates for the two-from-four treatment relative to the $(N{=}4; X{=}\$9)$ treatment. While this certainly motivates further research, some caution is warranted. As shown in Figure 2(B) the ongoing cooperation rate in the two-from-four treatment (marked with an empty circles) is not far from what we might expect from the meta-study in RPD games where cooperation is not an equilibrium (basin size of one). In contrast, as we highlighted earlier, the $(N{=}4; X{=}\$9)$ treatment with an ongoing cooperation rate that is almost at zero represents the only treatment that is notably far from the meta-study.

One possible explanation for the result is the stark nature of feedback and learning in the $(N{=}4; X{=}\$9)$ game. With a cooperation rate of 25 percent (approximately the expected value from the meta-study basin), the anticipated group-wide success rate with four players

[52] At $\delta = 3/4$ the always-defect independent-extension basin is smaller than one for $X < X^{\star} = \$7.91$. For all greater temptations the always-defect basin is full.

[53] We conducted three sessions for the two-from-four treatment (with 64 unique participants). Instructions for this treatment are identical to the four-from-four treatment, except for the explanation of the success/failure signals.

[54] Cooperation at the individual level is also significantly lower in the two-from-four robustness treatment ($p < 0.001$), compared to the 44.6 percent cooperation rate observed in the two-player RPD. However, as we weaken the cooperation requirement for efficiency, our focus is on a more-comparable measure, successful coordination in the group.

is merely 0.4 percent. In contrast, even with a lower cooperation rate of 20 percent in the two-from-four treatment, we would expect a considerably less extreme group-wide success rate of 18.1 percent.

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 ($M = 2$) is lower than the group size ($N = 4$), the basin-of-attraction extension predicts reduced coordination due to an increase in strategic uncertainty. The treatment results indicate low cooperation rates in line with empirical rates observed for extreme basin-values 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 relative to the treatment where $M = N = 2$. However, we also find that coordination is higher than in the $M = N = 4$ treatment, which runs counter to the prediction. This finding accentuates the extreme results in our high-tension ($X = 9$) multiplayer ($M = N = 4$) game.*

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 that captures 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 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. In particular, the model performs well in trading off the competing effects from the two distinct sources of strategic uncertainty. Meanwhile, we also document that the model is less successful in predicting initial cooperation rates. Outside of the laboratory, observing the initial round of cooperation can be challenging, but there is more hope that policymakers can observe features of ongoing interactions. Naturally, our game is highly stylized, but it suggests that further research that tests this model of equilibrium selection in more realistic settings may be useful for policy. Given the primitives of an environment, the basin-of-attraction model may be able to predict for what situations ongoing collusion is more likely to emerge. This information might be useful for antitrust authorities to decide which industries to allocate more attention to.

After illustrating the theoretical power of the model for implicit coordination, we turn to several application-motivated extensions that probe the model's limitations. We first show that results continue to hold when manipulations are introduced within the same population. We next document that if subjects are provided with a tool that reduces

strategic uncertainty (pre-play chat), the selection model is inappropriate for predicting behavior. That is, the model fails to predict when collusion can be explicitly coordinated. We finally demonstrate that easing cooperation requirements, specifically by stipulating that a subset of players is sufficient to achieve the efficient outcome, does not necessarily promote collusion. This can occur because, with a subset of players being adequate for the efficient outcome, additional strategic uncertainty arises regarding which individuals will cooperate and who may free-ride. The model captures this extra source of strategic uncertainty, predicting decreased cooperation

Taking a step back, a shortcoming of any experimental paper 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 now outline how we plan to address this in a companion paper (Boczoń, Vespa, and Wilson, 2024) that lays out a possible path for future research in this area. The idea is that one can evaluate the performance of artificial intelligence algorithms (AIAs) that companies use for pricing decisions (Calvano, Calzolari, Denicolo, and Pastorello, 2020; Asker, Fershtman, and Pakes, 2021) within the RPD setting. The companion paper shows that the experimental results for both the previous RPD literature and our main environments with $N > 2$ can be replicated with AIAs. 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, such AIAs can be used to expand the scope of experimental studies if partially validated on the narrower domains studied within the laboratory.

References

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

Aoyagi, Masaki, Venkataraman 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*.

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

Boczoń, Marta, Emanuel Vespa, and Alistair J Wilson (2024), "Lab to algorithm: Prediting AIs with humans, and vice versa." *In preparation*.

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, 3267–3297.

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.

Cooper, David J and Roberto A Weber (2020), "Recent advances in experimental coordination games." *Handbook of experimental game theory*, 149–183.

Cooper, Russell, Douglas V DeJong, Robert Forsythe, and Thomas W Ross (1993), "Forward induction in the battle-of-the-sexes games." *American Economic Review*, 1303–1316.

Cooper, Russell, Douglas V DeJong, Robert Forsythe, and Thomas W Ross (1994), "Alternative institutions for resolving coordination problems: experimental evidence on forward induction and preplaycommunication." *Problems of coordination in economic activity*, 129–146.

Cooper, Russell W, Douglas V DeJong, Robert Forsythe, and Thomas W Ross (1990), "Selection criteria in coordination games: Some experimental results." *The American Economic Review*, 80, 218–233.

Crawford, Vincent P, Uri Gneezy, and Yuval Rottenstreich (2008), "The power of focal points is limited: Even minute payoff asymmetry may yield large coordination failures." *American Economic Review*, 98, 1443–1458.

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, 3929–3952.

Dal Bó, Pedro, Guillaume R Fréchette, and Jeongbin Kim (2021), "The determinants of efficient behavior in coordination games." *Games and Economic Behavior*, 130, 352–368.

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 R Fréchette, and Ennio Stacchetti (2013), "An experimental study of imperfect public monitoring: Efficiency versus renegotiation-proofness." NYU working paper.

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, Drew, David G Rand, and Anna Dreber (2012), "Slow to anger and fast to forgive: Cooperation in an uncertain world." *American Economic Review*, 102, 720–749.

Ghidoni, Riccardo and Sigrid Suetens (2022), "The effect of sequentiality on cooperation in repeated games." *American Economic Journal: Microeconomics*, 14, 58–77.

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

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*. The MIT Press, Cambridge, MA.

Holm, Håkan J (2000), "Gender-based focal points." *Games and Economic Behavior*, 32, 292–314.

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 (2022), "A new approach to the analysis of cooperation under the shadow of the future: Theory and experimental evidence." Working Paper.

Kartal, Melis, Wieland Müller, and James Tremewan (2021), "Building trust: The costs and benefits of gradualism." *Games & Economic Behavior*, 130, 258–275.

Kloosterman, Andrew (2019), "Cooperation in stochastic games: A prisoner's dilemma experiment." *Experimental Economics*, 23, 447–467.

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*. The MIT Press, Cambridge, MA.

Mermer, Ayse Gül, Wieland Mueller, and Sigrid Suetens (2021), "Cooperation in infinitely repeated games of strategic complements and substitutes." *Journal of Economic Behavior & Organization*, 188, 1191–1205.

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.

Spagnolo, Giancarlo and Matthias Blonski (2001), "Prisoners' other dilemma." *Working Paper*.

Straub, Paul G (1995), "Risk dominance and coordination failures in static games." *The Quarterly Review of Economics and Finance*, 35, 339–363.

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.

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, 75–98.

Appendix A. Additional Tables and Figures

Table A.1. Cooperation and success rates across all supergames

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

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

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

History	$X = \$9$		$X = \$1$		Chat ($X = \$9, N = 4$)	
	$N = 2$	$N = 4$	$N = 4$	$N = 10$	$\delta = 3/4$	$\delta = 1/2$
(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 in parentheses. 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}).

TABLE A.3. Cardinal comparisons to meta-study predictions

Treatment	Independent basin		Correlated basin		Cooperation	
	Initial	Ongoing	Initial	Ongoing	Initial	Ongoing
$(N{=}2; X{=}\$9)$	0.495	0.373	0.495	0.373	0.503	0.450
	($p = 0.954$)	($p = 0.148$)	($p = 0.954$)	($p = 0.148$)		
$(N{=}4; X{=}\$9)$	0.237	0.163	0.495	0.373	0.035	0.006
	($p < 0.001$)	($p < 0.001$)	($p < 0.001$)	($p < 0.001$)		
$(N{=}4; X{=}\$1)$	0.495	0.373	0.842	0.718	0.792	0.409
	($p < 0.001$)	($p = 0.436$)	($p = 0.164$)	($p < 0.001$)		
$(N{=}10; X{=}\$1)$	0.495	0.373	0.842	0.718	0.357	0.187
	($p < 0.001$)	($p = 0.612$)	($p < 0.001$)	($p < 0.001$)		
$F-$test, all	49.2	643.0	204.6	3,661.7		
	($p < 0.001$)	($p < 0.001$)	($p < 0.001$)	($p < 0.001$)		
$F-$test, all but $(N{=}4; X{=}\$9)$	17.72	0.99	27.1	95.6		
	($p < 0.001$)	($p = 0.398$)	($p < 0.001$)	($p < 0.001$)		

Note: Results are calculated using data from the last five supergames using subject-clustered standard errors. In the first column, we present our four main treatments. The second and third columns display corresponding rates of initial and ongoing cooperation as predicted by the meta-study for the independent-basin measure. The fourth and fifth columns show cooperation rates predicted by the meta-study for the correlated basin measure. In the last column, we present observed cooperation rates in our main treatments. The p-values listed in the first four rows result from testing the null hypothesis of statistical equivalence between the observed and predicted cooperation rates. In the last two rows, we report F-statistics and p-values from testing the null hypothesis that observed and predicted cooperation rates across the treatments are statistically equal. The first F-test is conducted across all treatments, and the second is performed for all treatments except $(N{=}4; X{=}\$9)$.

TABLE A.4. Ordinal pairwise treatment comparisons

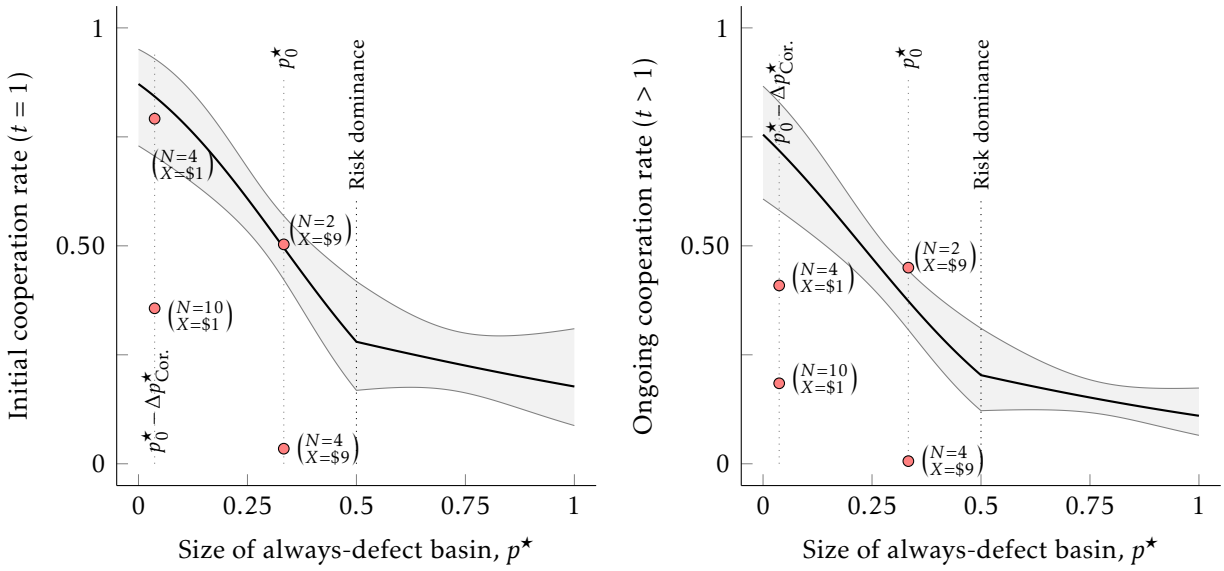
Treatment pair	Basin		Cooperation	
	Independent	Correlated	Initial	Ongoing
$(N{=}2; X{=}\$9)$ vs. $(N{=}4; X{=}\$9)$	$\succ$	$\sim$	$\hat{\succ}$ ($p < 0.001$)	$\hat{\succ}$ ($p < 0.001$)
$(N{=}2; X{=}\$9)$ vs. $(N{=}4; X{=}\$1)$	$\sim$	$\prec$	$\hat{\prec}$ ($p < 0.001$)	$\hat{\sim}$ ($p = 0.585$)
$(N{=}2; X{=}\$9)$ vs. $(N{=}10; X{=}\$1)$	$\succ$	$\prec$	$\hat{\succ}$ ($p = 0.070$)	$\hat{\succ}$ ($p = 0.002$)
$(N{=}4; X{=}\$9)$ vs. $(N{=}4; X{=}\$1)$	$\prec$	$\prec$	$\hat{\prec}$ ($p < 0.001$)	$\hat{\prec}$ ($p < 0.001$)
$(N{=}4; X{=}\$9)$ vs. $(N{=}10; X{=}\$1)$	$\sim$	$\prec$	$\hat{\prec}$ ($p < 0.001$)	$\hat{\prec}$ ($p < 0.001$)
$(N{=}4; X{=}\$1)$ vs. $(N{=}10; X{=}\$1)$	$\succ$	$\sim$	$\hat{\succ}$ ($p < 0.001$)	$\hat{\succ}$ ($p = 0.002$)
# correct directional predictions (out of 4)				
Independent basin			4	4
Correlated basin			3	3
# correct null predictions (out of 2)				
Independent basin			0	1
Correlated basin			0	0

Note: Results are calculated using data from the last five supergames. In the first column, we present six treatment comparisons. The next two columns indicate which of the two treatments in each pair has a higher predicted cooperation rate under the independent basin (column 2) and the correlated basin (column 3). The symbol $\succ$ signifies that the treatment listed first has a higher predicted cooperation rate, $\prec$ indicates that the treatment listed second has a higher predicted cooperation rate, and $\sim$ denotes that the two treatments have the same predicted cooperation rate. In the last two columns, we conduct tests to assess statistically significant differences in observed cooperation rates within each treatment pair, both for initial (column 4) and ongoing (column 5) cooperation, separately. All tests utilize subject-clustered standard errors and adhere to the 10 percent significance level. $\hat{\succ}$ indicates that the treatment listed first has empirically higher cooperation rate, $\hat{\prec}$ indicates that the treatment listed second has empirically higher cooperation rate, and $\hat{\sim}$ indicates that the cooperation rates observed in the two treatments are statistically indistinguishable. In the bottom half of the table, we present summary statistics for the total number of correct directional and null predictions under each basin extension.

TABLE A.5. Cooperation and success rates with implicit vs. explicit coordination

Action and signal rates	Implicit	Explicit	
	NoChat(3/4)	Chat(3/4)	Chat(1/2)
	Cooperation		
Initial	0.035 (0.017)	0.988 (0.007)	0.300 (0.037)
Ongoing	0.006 (0.003)	0.806 (0.030)	0.044 (0.018)
	Success		
Initial	0.000	0.971	0.094
Ongoing	0.000	0.756	0.002

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



(A) Initial cooperation

(B) Ongoing cooperation

FIGURE A.1. Cooperation under the correlated basin-size model

Note: Filled circles indicate separate treatments and filled diamonds treatments pooled over each value of the independent-basin measure. See Figure 2 in Section 4 for analogous results under the independent basin.

Appendix B. Further Analysis of the Within-Subject Treatments

In the within-subject treatments we find evidence of hysteresis. 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. 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 significantly greater than the initial cooperation rates in supergame one.[B.1] This suggests that in the short run, subjects respond to a change in the environment with a strong intent to cooperate after accumulating experience at another parameter value.

In the first four columns of Table B.1 we compare behavior of within- and between-treatment subjects after five rounds of experience holding the payoff cost of cooperating fixed at $X = \$9$. In the first two columns we present average behavior of between-treatment subjects in supergames 6–10 for $N = 2$ and $N = 4$. In the next two columns we present average behavior of within-treatment subjects in supergames 16–20 in the $N = 4 \to 2$ and $N = 2 \to 4$ treatments. Examining the differences across the *within* and *between* subjects, we find: (i) No statistically significant differences for $N = 4$ and $N = 2 \to 4$ ($p = 0.117/p = 0.539$ for initial/ongoing cooperation), (ii) Statistically significant differences for $N = 2$ and $N = 4 \to 2$ ($p = 0.011$ for initial, $p < 0.001$ for ongoing). The significant differences reflect the substantially greater upward shift in the $4 \to 2$ treatment.

In the last three columns of Table B.1 we compare changes in behavior of within- and between-treatment subjects in response to a change in N. In column Δ_{Btwn} we present the change in average behavior of between-treatment subjects in supergames 16–20 as N increases from $N = 2$ to $N = 4$.[B.2] In the last two columns (jointly labeled as$\Delta_{\text{Wthn.}}$) we present the within-subject change in average behavior for the $2 \to 4$ and $4 \to 2$ 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 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.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.

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

TABLE B.1. Cooperation and success rates with between vs. within identification

Action and signal rates	Between (SG 6–10)		Within (SG 16–20)		$\Delta_{\text{Btwn.}}$	$\Delta_{\text{Wthn.}}$	
	$N = 2$	$N = 4$	$N = 2$	$N = 4$		$2 \to 4$	$4 \to 2$
	Cooperation						
Initial	0.474 (0.036)	0.139 (0.025)	0.643 (0.056)	0.214 (0.041)	−0.469 (0.060)	−0.260 (0.042)	−0.504 (0.056)
Ongoing	0.299 (0.026)	0.054 (0.012)	0.598 (0.051)	0.042 (0.016)	−0.444 (0.055)	−0.258 (0.029)	−0.544 (0.050)
	Success						
Initial	0.474	0.011	0.643	0.042	-0.503	-0.433	-0.632
Ongoing	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 in parentheses. The last three columns measure the corresponding cooperation rate when $N = 4$ minus the cooperation rate when $N = 2$.

Appendix C. General Basin Calculation

Consider the N-player environment where we require cooperation from $M-1$ other players in order to receive a success signal. As such, if M players cooperate, then all N players will be guaranteed to receive a success signal. In the main body of the paper we consider a boundary case where $N = M$. In this appendix, we consider a generalized basin of attraction calculation where $M \leq N$.

Define $F(k)$ as the CDF over k, which is the number of other players choosing grim trigger. An agent will prefer grim trigger over always defect as long as:

$$(1 - F(M-2)) - F(M-2)\delta x \geq (1 - F(M-1))(1+x) - (F(M-1) - F(M-2))\delta(1+x),$$

which is equivalent to:

$$\Pr(\text{Exactly } M-1 \text{ choose } \alpha_{\text{Grim}}) \geq x\frac{(1-\delta)}{\delta} + x\Pr(\text{More than } M-1 \text{ choose } \alpha_{\text{Grim}}).$$

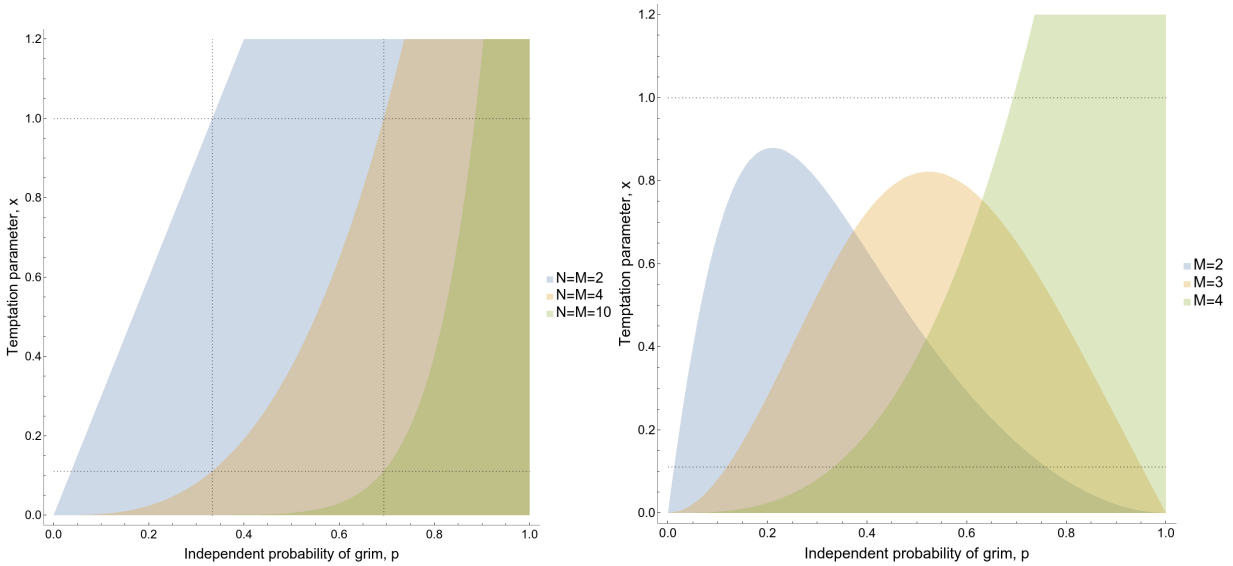
Given independent beliefs and probability p of others playing grim trigger, F is the CDF of a Binomial$(N-1, p)$. As such, we can simplify the preference for conditional cooperation to:

$$\binom{N-1}{M-1} p^{M-1}(1-p)^{N-M} \geq x\frac{(1-\delta)}{\delta}\sum_{k=0}^{M-1}\binom{N-1}{k} p^k (1-p)^{N-k-1}.$$

For $M = N$ cooperation is preferred for $p \in [p^\star, 1]$.

For $1 < M < N$ cooperation is preferred for $p \in \left[\underline{p}^\star, \overline{p}^\star\right]$, where $0 < \underline{p}^\star \leq \overline{p}^\star < 1$. This is the case because cooperation is never a best response if no one else cooperates (impossible to get a success, as $M \geq 2$) or if everyone else cooperates (success regardless of own action, $M < N$).

In Figure C.1 we outline the values of p and x for which cooperation is preferred at $\delta = 3/4$. In Panel (A) we do so for $N = M \in \{2, 4, 10\}$, and in Panel (B) for $N = 4$ and $M \in \{2, 3, 4\}$. The figure illustrates that for $M < N$ the basin size is given by $\underline{p}^\star + 1 - \overline{p}^\star$ instead of $p^\star$. Moreover, the figure illustrates that the basin size is full for our extension treatment with $M = 2$ and $X = \$9$.



(A) $N = M$ (B) $M < N = 4$

FIGURE C.1. Grim best response to independent belief p and temptation x

Appendix D. Estimation of σ

Our measure of how much belief correlation is necessary to rationalize the data uses a convex combination of the independent and correlated models. Specifically, the probability that the other $N-1$ players coordinate is given by

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

with the critical belief denoted by $p^{\star}(\sigma, x, N)$. The additional parameter σ nests the two extremes: $\sigma = 0$ for full independence, $\sigma = 1$ for perfect correlation.

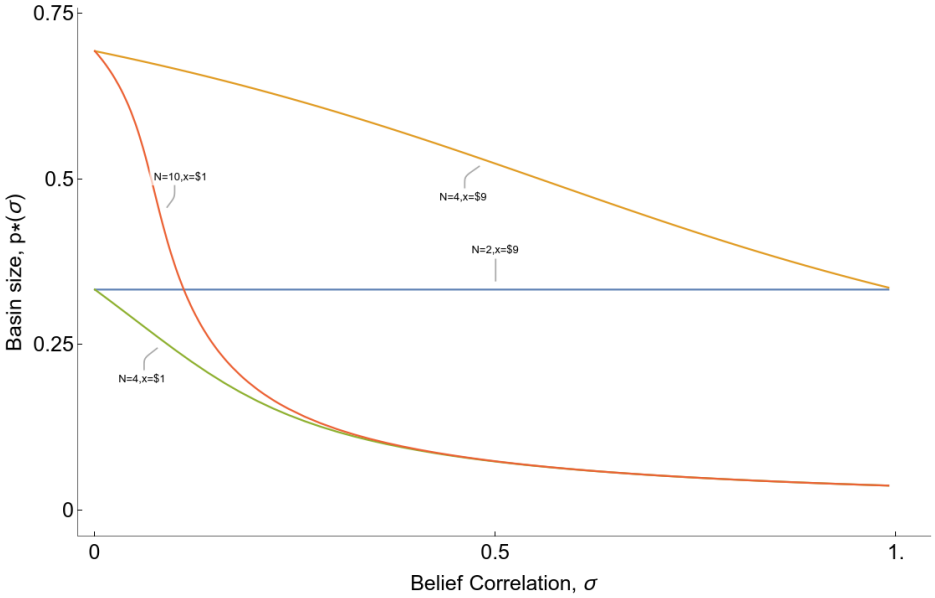


Figure D.1. Belief correlation and basin size

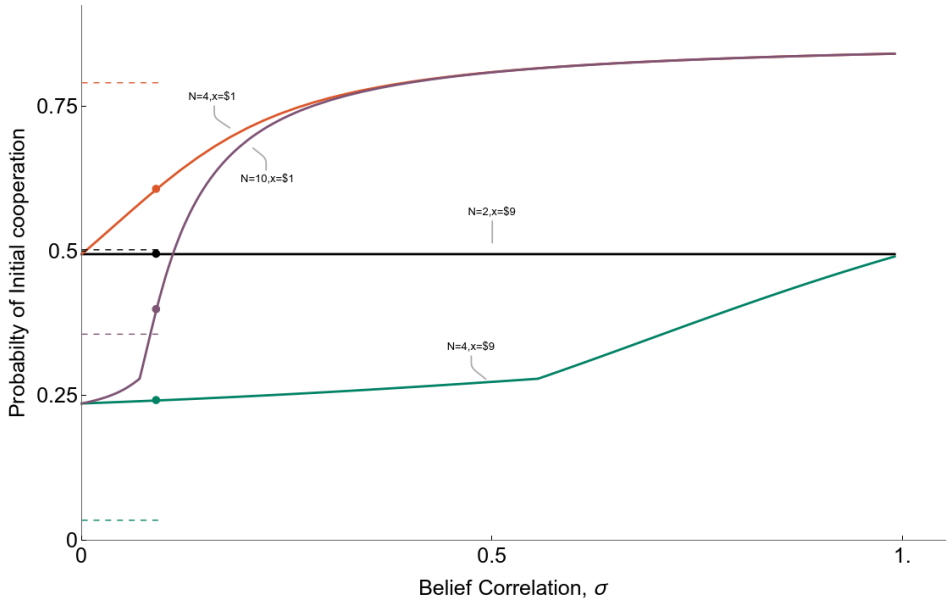
Note: The figure displays analytic solutions for $p^{\star}(\sigma)$ for any value of $\sigma \in [0,1]$

To provide some insight into the estimator of σ, consider Figure D.1. The figure outlines the analytic solution for $p^{\star}(\sigma)$, the basin of attraction in each treatment, as a function of σ. For instance, when $\sigma = 0$ ($\sigma = 1$) the value of $p^{\star}$ for each treatment coincide with the predictions for the independent (correlated) extension discussed in Table 1 Panel B. Between the two extremes, the figure displays the predicated intermediate values of $p^{\star}$ for each σ within the range of $(0,1)$.

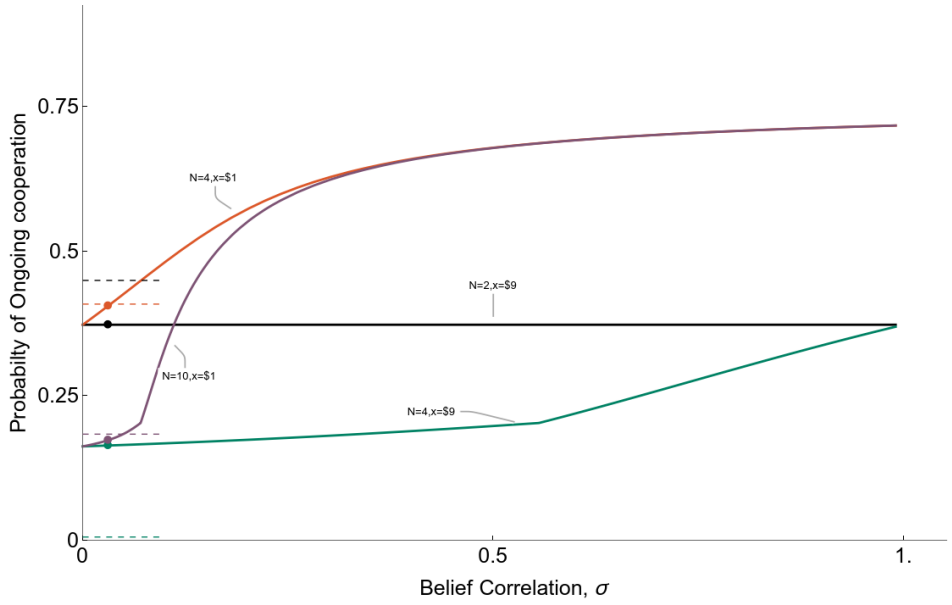
For each $p^{\star}$, we can use the probit estimates from the meta-study, and obtain predicted cooperation, $\hat{C}_{\text{Meta}}(p^{\star})$; see Section 2 for details. Figure D.2(A) shows the predictions for the case of initial cooperation and Figure D.2(B) for ongoing cooperation. In addition, notice that both figures display in dashed lines the observed cooperation rates in our treatments.

With this background we can construct a log-likelihood equation across our four treatments:

$$\begin{aligned} l(\sigma) = & \, l\left(\hat{C}_{\text{Meta}}\left(p^{\star}(\sigma,1,2)\right)\right) + l\left(\hat{C}_{\text{Meta}}\left(p^{\star}(\sigma,1,4)\right)\right) \\ & + l\left(\hat{C}_{\text{Meta}}\left(p^{\star}(\sigma,1/9,4)\right)\right) + l\left(\hat{C}_{\text{Meta}}\left(p^{\star}(\sigma,1/9,10)\right)\right). \end{aligned}$$



(A) Initial cooperation



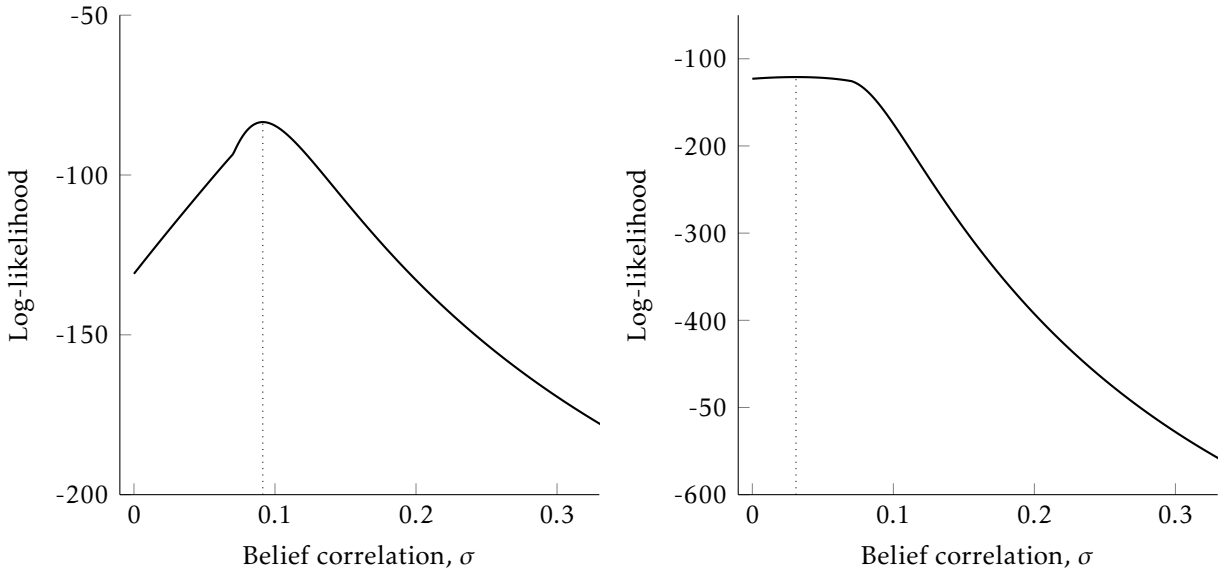
(B) Ongoing cooperation

FIGURE D.2. Belief correlation and cooperation

Note: Using the estimated initial (ongoing) cooperation rates from the meta study as a function of basin-size $\hat{C}_{\text{Meta}}\left(p^{\star}\right)$ we can therefore indicate the expected cooperation level for any value of σ, which we illustrate in panel A (B). The dashed lines (with color matching the corresponding treatment) mark the observed cooperation rates in our treatments. The dots indicate the corresponding value of σ that maximizes the likelihood function.

This is a single equation in σ, where the likelihood of our data from each treatment is measured as a binomial under a cooperation probability of $\hat{C}_{\text{Meta}}\left(p^{\star}\left(\sigma, x, N\right)\right)$.[D.1] Finally, we estimate σ via maximum likelihood.

[D.1] Notice that the expression uses the normalized value x, so that $x = 1, N = 2$ corresponds to the $(N{=}2; X{=}\$9)$ treatment.



(A) Initial cooperation

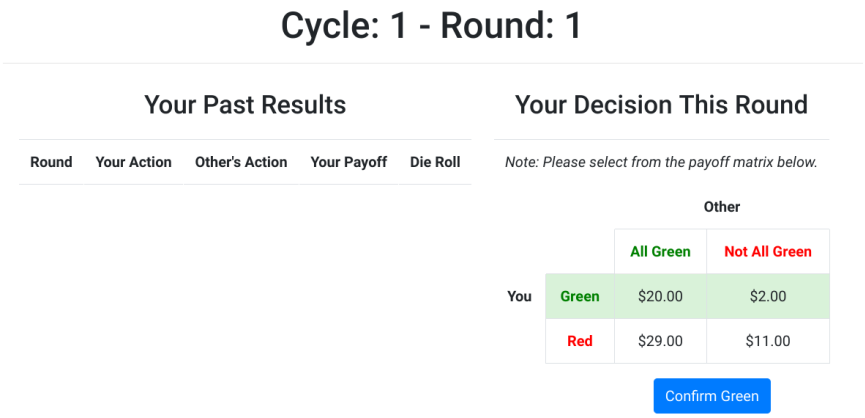
(B) Ongoing cooperation

FIGURE D.3. Belief correlation

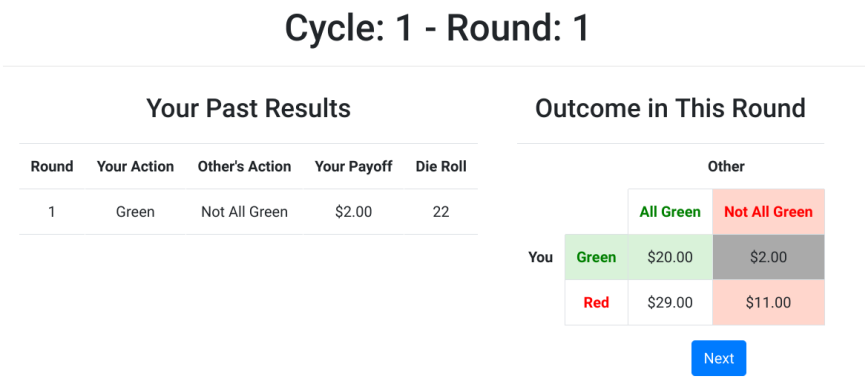
Note: Data are taken from the last five supergames in each between-subject treatment. Log-likelihoods are calculated using the imputed cooperation rate from the Dal Bó and Fréchette (2018) meta-study with belief correlation σ calculated as σ proportion of independent beliefs and $(1-\sigma)$ proportion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

Figure D.3 shows the log-likelihood as a function of σ for initial and ongoing cooperation.

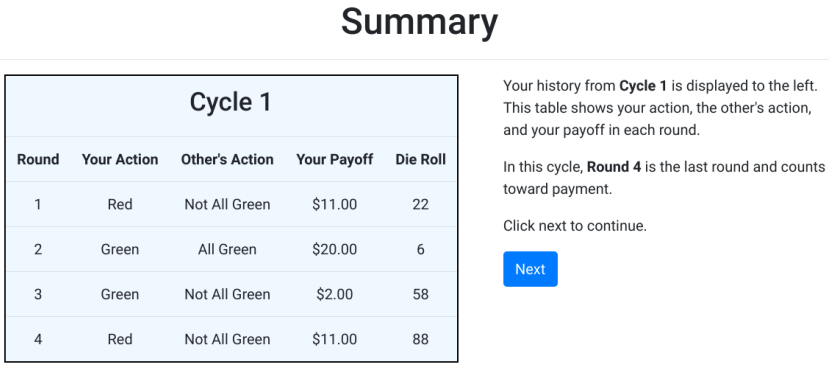
Appendix E. Interface Screenshots



(A) Action selection



(B) Round feedback



(C) Supergame feedback

Figure E.1. Interface screenshots

Appendix F. Provided Instructions

Below, we include the instructions provided to participants. All language deltas/treatment-specific language is enclosed in braces. Text in red pertains to the $N = 2$ treatment, while text in blue pertains to the $N > 2$ treatments (here we provide the $N = 4$ implementation, where $N = 10$ has only minor changes). Payoff text for $X = 9$ is presented in green, and for $X = 1$ in orange. Separate instructions for {Part two} are provided to treatments where 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 out of four players are needed for a success signal, adjustments are made to the description in the *Round Choices and Payoffs* to accommodate this 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 G. Strategies and the selection index

An RPD is characterized by a large number of possible strategies. However, the 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 is limited to the first interaction, where tit for tat starts with cooperation and suspicious tit for that starts with defection. In all subsequent rounds, both players cooperate as long as their opponents cooperate and defect otherwise.

In this paper, and more broadly in the RPD literature, the selection index focuses on two strategies: always defect and grim trigger. A first reason to focus on these two strategies is that they capture very distinct types of behavior (non-cooperative and conditionally cooperative) that both may be supported in equilibrium. Always defect is the only non-cooperative strategy that is subgame perfect, and grim trigger is the only conditionally cooperative *and* empirically-relevant strategy that depending on δ can be subgame perfect. Note that for tit for tat, which is a Nash equilibrium of the supergame, there can be incentives to deviate from the punishment path. In addition, if one player chooses always defect and the other plays tit for tat, the outcome is cooperation in the first round and defection from the second round on. This is the same outcome that will realize if the other player chooses grim trigger instead. Similarly, if both players choose tit for tat, the outcome is the same (cooperation in every round) as when the two players use grim trigger instead. Therefore, one can begin to argue that there is little loss in focusing solely on grim trigger as the conditionally cooperative strategy.

In this appendix we show why focusing on always defect and grim trigger also extends to our setting. Here, we start with a brief description of the Strategy Frequency Estimation Method (SFEM), which was introduced in Dal Bó and Fréchette (2011).[G.1] From a big-picture perspective, the method takes choices made by subjects and contrasts them against hypothetical choices that would have been made were the subjects using a different strategy from a pre-determined strategy set. Using a mixture model that allows for errors in choices, the procedure reports the proportion of choices that are better rationalized by each strategy. Dal Bó and Fréchette (2019) also use an alternative procedure to study strategies: an experimental design that familiarizes subjects with a set of strategies and asks them to select one to be played in their name. The authors contrast this elicitation procedure with the SFEM and 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 $\mathcal{K}$. Let $d^i_{gr}(\mathbf{h})$ be the choice of subject i and $k^i_{gr}(\mathbf{h})$ the decision prescribed for subject i by strategy $k \in \mathcal{K}$ in round r of supergame g for

[G.1] Further details on the procedure are available in the online appendix of Embrey, Fréchette, and Stacchetti (2013). A Monte-Carlo-style analysis was also performed in Fudenberg, Rand, and Dreber (2012). 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).

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 prescribed decision (k) as:

$$\Pr\left(d^i_{gr}(\mathbf{h}) = k^i_{gr}(\mathbf{h})\right) = 1/1+\exp\left(\frac{-1}{\gamma}\right) = \beta, \tag{8}$$

where β captures the probability that the subject does not make make mental errors in the implementation of a strategy and $\gamma > 0$ is a parameter of interest. In the limit, as $\gamma \to 0$ and $\beta \to 1$ the model fully rationalizes the data. On the other hand, as $\gamma \to \infty$, $\beta \to \frac{1}{2}$, the model has no explanatory power.

Now, let y^i_{gr} be an indicator that takes value one if the subject's choice matches the decision prescribed by the strategy. It follows from Equation (8) that the likelihood of observing strategy k for subject i is given by:

$$p_i(k) = \prod_g \prod_r \beta^{y^i_{gr}} (1-\beta)^{1-y^i_{gr}}. \tag{9}$$

Aggregating over subjects we arrive at the log-likelihood of the following form: $\sum_i \ln\left(\sum_k \phi_k p_i(k)\right)$.[G.2,G.3]

For illustration, consider a case in which the set of strategies includes always defect and always cooperate. The fit of the model will be good (high β) if the population is composed of subjects who either almost always defect or almost always cooperate. If a large proportion of subjects shifts between cooperation and defection within a supergame, neither strategy will accommodate their choices and the estimation will return a low estimate of β.

The SFEM depends on the pre-specified set of strategies $\mathcal{K}$. 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 a two-player RPD. The reason is that in our multi-player game subjects do not learn the specific choices of others, but instead, receive an aggregate signal of either a success or a failure. Therefore, to study behavior of multiple players we focus on the same set of five strategies identified in Dal Bó and Fréchette (2011) for a two-player game. In fact, as we will show later, these five strategies suffice to obtain relatively high goodness of fit estimates (as captured by β).

[G.2] To construct $p_i(k)$, consider a subject who is implementing the prescriptions of strategy k with mistake rate given by $1-\beta$. If $y^i_{gr} = 1$, the subject's choice matches the prescription; if $y^i_{gp} = 0$, the subject's choice does not coincide with the one prescribed by the strategy.

[G.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 (8) 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.

Results. In this section, we will present results obtained using the SFEM. First, we will show estimates for the last seven supergames of the session and then, we will provide results for the first seven supergames of the session.[G.4]

Final Supergames. In Panel (A) of Table G.1 we present estimation results for each of our nine treatments, including the estimates of β. For each treatment and each of the five strategies identified as focal in Dal Bó and Fréchette (2018) the table reports the estimates and (whenever possible) the bootstrap-estimated standard errors.[G.5,G.6]

In Panel (B) of Table G.1 we report goodness-of-fit estimates obtained after restricting the set of available strategies. In particular, $\beta^{\dagger}$ corresponds to the β estimate after eliminating tit for tat and suspicious tit for tat from the strategy set. In this case, the only conditional-cooperation strategy remaining in the set is grim trigger. Since the model uses maximum likelihood and the restricted and unrestricted models are nested, we use a likelihood-ratio test with a $p-$value referenced with † to evaluate the null hypothesis that the restriction does not bind. In the last two rows in Table G.1 Panel (B) we report results of an analogous exercise but where the only two strategies included in the strategy set are always cooperate and always defect. The corresponding β estimate and $p-$value are marked with ‡. In what follows, we discuss the estimation results across different strategy sets treatment by treatment.

In the (N=2; X=\$9) treatment, two strategies with the most mass are always defect (45.2 percent) and grim trigger (28.9 percent). However, there is a non-negligible yet not-significant mass captured by tit for tat (18.0 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: a drop from 0.929 to 0.912. On the one hand, the reduction in goodness of fit seems marginal. On the other hand, the likelihood ratio test suggests that excluding tit for tat and suspicious tit for tat from the strategy set leads to a statistically significant loss. To put this result into perspective, consider a SFEM estimation that only accounts for always cooperate and always defect. Here, we observe a relatively large decrease in the goodness-of-fit measure: a drop from 0.929 to 0.804. This suggests that the noise component needed to rationalize the data without grim trigger is substantially larger than with grim trigger in the strategy set.

The (N=4; X=\$9) treatment is one with the least amount of cooperation, as confirmed by our estimation results. About a third of the mass corresponds to always defect and about two thirds to suspicious tit for tat. Here, the goodness-of-fit measure is close to one (at 0.981). Moreover, there is no evidence of a loss in carrying out the estimation without tit

[G.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.

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

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

for tat and suspicious tit for tat, as shown by the $\beta^{\dagger}$ estimates and the likelihood ratio test statistics.

The treatment with the highest degree of initial cooperation in our data is (N=4; X=\$1). Here, about a little less than a quarter of the mass corresponds to always cooperate. While this suggests a relatively large amount of unconditional cooperation, whether a player cooperates unconditionally cannot be determined unless their opponent defects.[G.7] However, the broader evidence suggests that cooperation is conditional. In treatments with more frequent defection there is essentially no evidence of large amounts of unconditional cooperation. In fact, in the (N=4; X=\$1) treatment the most popular strategy is tit for tat, which captures 53.5 percent of the mass. There is also a close to 20.0 percent of the mass that corresponds to always defect. We note that while the likelihood ratio test rejects the null, the loss in terms of goodness of fit is rather small: a drop from 0.939 to 0.931. A reduction in goodness of fit is much larger in the absence of grim trigger: a drop from 0.939 to 0.840.

In our last core treatment, (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 a tenth to always cooperate. Grim trigger, tit for tat, and suspicious tit for tat, jointly, account for close to 60.0 percent of the mass. 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. Consistently, the likelihood ratio test rejects the null at any typical significance level. However, the goodness-of-fit measure decreases from 0.934 to 0.897 without grim trigger in the strategy set.

Overall, we find that:

Result G.1. *Focusing on two strategies, always defect and grim trigger, when testing the extensions of the basin does not lead to a substantial loss. This is the case 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).*

Result G.2. *Further restricting the strategy set to exclude grim trigger, in most cases, leads to a relatively large loss in goodness of fit.*

We now discuss the estimates for our extension treatments. In the $2 \rightarrow 4$ treatment the last seven supergames are played with $N = 4$, so that the results of this extension are comparable to those obtained in our core (N=4; X=\$9) treatment. However, the former

[G.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 corresponds 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 others' choices in round $r-1$. That is, the subject makes a choice in round r for each possible choice of the other player in round $r-1$. After making these choices, the subjects learn the actual history of play for round $r-1$, and their choices for round r are implemented. 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 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.

estimates are noisier (relative to the latter). This is likely due to the fact that in the last seven supergames of the $2 \rightarrow 4$ treatment subjects are less experienced relative to their counterparts in (N=4; X=\$9).[G.8] However, the big picture is similar: (i) 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 (a drop from 0.950 to 0.948);[G.9] (ii) There is a non-negligible loss in the goodness-of-fit measure after excluding grim trigger from the strategy set (a drop from 0.950 to 0.895); (iii) 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). Meanwhile, in the more competitive $4 \rightarrow 2$ treatment excluding both tit for tat and suspicious tit for that leads to a larger loss in goodness of fit relative to other treatments (a drop from 0.921 to 0.883). However, were we to additionally exclude grim trigger, it would lead to a much larger loss, with $\beta^{\ddagger}$ at 0.819.

In a treatment with chat and $\delta = 3/4$, grim trigger and tit for tat essentially capture almost all the mass. Excluding tit for that and suspicious tit for tat leads to a statistically significant yet negligible reduction in goodness of fit. A reduction in goodness of fit is much larger when grim trigger is also excluded: a drop from 0.975 to 0.883. On the contrary, in a chat treatment with $\delta = 1/2$, always defect captures more than 60.0 percent of the mass and suspicious tit for tat accounts for 11.0 percent of the mass. Here, excluding tit for that and suspicious tit for tat has virtually no effect on the goodness of fit (a drop from 0.975 to 0.974). However, were we to additionally exclude grim trigger, it would lead to an economically and statistically significant loss, with $\beta^{\ddagger}$ at 0.809.[G.10]

In our final extension treatment always defect captures close to three-quarters of the mass. This evidence is consistent with what we have observed earlier that cooperation in our treatment with weakened cooperation requirement is quite unlikely despite the fact that the number of players needed for a cooperative outcome is smaller than N. Furthermore, we observe a small reduction in the goodness-of-fit estimate when the set of strategies excludes tit for tat and suspicious tit for tat (a drop from 0.899 to 0.896). The reduction in goodness of fit becomes much larger once we also exclude grim trigger (a drop from 0.899 to 0.863).

Early Supergames. We conclude this section of the appendix by describing SFEM estimates for the first seven supergames. We first note that the goodness-of-fit estimates (β) are still quite far from random, with the smallest estimate at 0.812. This suggests that even if constrained to few strategies, most of the data can be rationalized. However, in all but one treatment there are large reductions in goodness-of-fit estimates when compared to the last seven supergames (see Table G.1). This suggests that as subjects gather experience, their behavior becomes more consistently captured by the five focal strategies identified in Dal Bó and Fréchette (2018).

[G.8] In fact, the estimates for $2 \rightarrow 4$ in Table G.1 are closer to the estimates for (N=4; X=\$9) using the *first* seven supergames, which are reported in Table G.2. In both cases, the largest mass corresponds to always defect (around 60 percent) and suspicious tit for tat.

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

[G.10] While there is a non-negligible mass for grim trigger, any time a subject who plays a grim-trigger strategy is matched with a player who uses always defect or suspicious tit for tat they will start to defect in round two. Therefore, the likelihood of long-term cooperation is very small.

TABLE G.1. SFEM output in the last seven supergames

Strategies	$(N{=}2; X{=}\$9)$	$(N{=}4; X{=}\$9)$	$(N{=}4; X{=}\$1)$	$\binom{N=10}{X=\$1}$	$2 \rightarrow 4$	$4 \rightarrow 2$	Chat(3/4)	Chat(1/2)	*Two from four*
				Panel A.					
Always cooperate	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)
Always defect	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
				Panel B.					
$\beta^{\dagger}$	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.

Second, comparing between the first- and the last seven supergames we observe differences in strategies that best capture the observed behavior. For instance, in (N=4; X=\$9) the largest mass in the first seven supergames corresponds to always defect (60.0 percent) and the second largest to suspicious tit for tat (33.8 percent). In the last seven supergames, the order is reversed, with 67.0 percent for suspicious tit for tat and 31.1 percent for always defect. However, in both cases, the two strategies jointly capture more than 90.0 percent of the mass. That is, the odds of a cooperative outcome in the second round among inexperienced and experienced players are equally slim.

Another example is the (N=10; X=\$1) treatment, in which in the first seven supergames close to 50.0 percent of the mass corresponds to strategies that start by cooperating. This changes in the last seven supergames in which almost 70.0 percent of the mass corresponds to strategies that start by defecting. These results suggest that subjects in this treatment start by cooperating but in time switch to defect.

Moving to our extension treatments, we note that behavior in the first seven supergames of our within treatments does not coincide with that observed in the last seven supergames among within-treatment subjects. In the first seven supergames of $2 \rightarrow 4$ close to 40.0 percent of the mass corresponds to strategies that start by cooperating. Meanwhile, in the last seven supergames, 80.0 percent of the mass is captured by strategies that start by defecting. This change in behavior is even more striking in our second within-subject treatment. In the first seven supergames of $4 \rightarrow 2$, slightly more than 80.0 percent of the mass corresponds to strategies that start by defecting. However, in the last seven supergames, only about a third of strategies start by defecting.

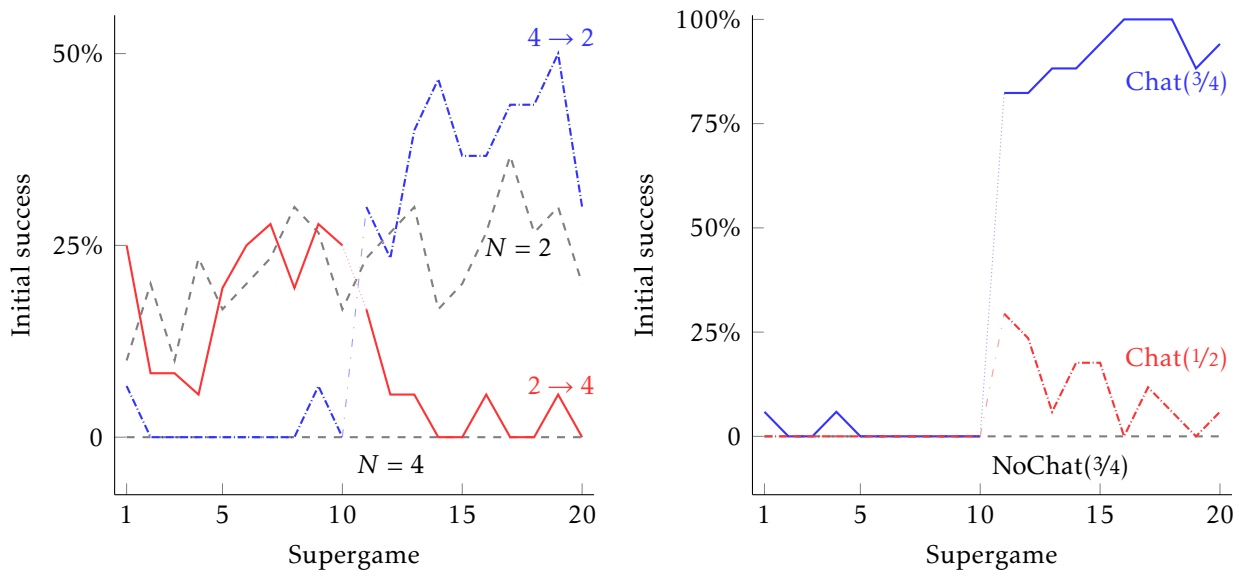
Similarly, we observe large differences between the the first and the last seven supergames in chat treatments in which the possibility to exchange messages is only introduced in the second half of the session. For example for $\delta = 3/4$, in the first seven supergames more than 80.0 percent of the mass corresponds to strategies that start with defection. In the last seven supergames more than 90.0 percent of the mass is captured by strategies that start with cooperation.

Finally, in the extension treatment where only two of four players are needed for a cooperative outcome we see less of an effect of learning as the session progresses. Between the first and the last seven supergames the mass associated with strategies that start with defecting decreases by less than 8 percentage points.

TABLE G.2. SFEM output in the first seven supergames

Strategies	$(N{=}2; X{=}\$9)$	$(N{=}4; X{=}\$9)$	$(N{=}4; X{=}\$1)$	$\binom{N=10}{X=\$1}$	$2 \rightarrow 4$	$4 \rightarrow 2$	Chat(3/4)	Chat(1/2)	*Two from four*
Always cooperate	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)
Always defect	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.



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

(B) Explicit vs. implicit (N=4; X=\$9)

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