Experimental Economics (2025), **28**, 75–106
doi:10.1017/eec.2024.14

ORIGINAL PAPER

Incentives crowd out voluntary cooperation: evidence from gift-exchange experiments

Simon Gächter[1,2,3], Esther Kaiser[4] and Manfred Königstein[5]

[1]CeDEx, School of Economics, University of Nottingham, Nottingham, UK
[2]IZA, Bonn, Germany
[3]CESifo, Munich, Germany
[4]ZHAW School of Management and Law, Winterthur, Switzerland
[5]Professur für Angewandte Mikroökonomie, Universität Erfurt, Erfurt, Germany
Corresponding author: Simon Gächter; Email: simon.gaechter@nottingham.ac.uk

(Received 11 November 2024; revised 29 December 2024; accepted 29 December 2024)

Abstract
Explicit and implicit incentives and opportunities for mutually beneficial voluntary cooperation coexist in many economic relationships. In a series of eight laboratory gift-exchange experiments, we show that incentives can lead to crowding out of voluntary cooperation even after they have been abolished. This crowding-out also occurs in repeated relationships, which otherwise strongly increase effort compared to one-shot interactions. Using a unified econometric framework, we unpack these results as a function of positive and negative reciprocity, as well as the principals' wage offer and the incentive compatibility of the contract. Crowding-out occurs mostly due to reduced wages and not a change in reciprocal wage–effort relationships. Our systematic analysis also replicates established results on gift exchange, incentives, and crowding out of voluntary cooperation while being exposed to incentives. Overall, our findings show that the behavioral consequences of explicit incentives strongly depend on the features of the situation in which they are embedded.

Keywords: Crowding-out; Explicit incentives; Gift-exchange experiments; Incomplete contracts; Implicit incentives; Principal-agent games; Repeated games

JEL Codes: C70; C90

1. Introduction

Explicit and implicit performance incentives, as well as mutually beneficial opportunities for voluntary cooperation, coexist in many contractual and organizational settings. Explicit incentives ('pay for performance') and implicit incentives (strategic incentives enabled in repeated interactions) appeal to an agent's self-interest to exert high effort. Empirical and (field) experimental evidence (e.g., Anderhub, Gächter & Königstein, 2002; Bandiera, Barankay & Rasul, 2005; Gächter, Huang & Sefton, 2016; Lazear, 2000; Shearer, 2004) shows that effort behavior is often consistent with predictions of self-interest-based incentive theory. However, a large body of experimental evidence from trust games, gift-exchange games, and public goods games also shows that many people are willing to cooperate voluntarily – that is, to act against their self-interest to benefit others (for surveys, see, e.g., Bowles, 2016; Chaudhuri, 2011; Drouvelis, 2021; Fehr & Charness, 2024; Fehr & Fischbacher, 2003; Fehr & Schurtenberger, 2018; Gintis, Bowles, Boyd & Fehr, 2005). If both motivations – following explicit material incentives as set out in contracts and institutions and voluntary cooperation as motivated by social preferences – are behaviorally relevant and coexist in many economic relationships (e.g., Bewley, 1999; Fehr & Falk, 2002), the question arises how they influence agents' effort choices.

In this article, we study how incentives affect voluntary cooperation in the form of high levels of social value-enhancing effort choices. In naturally occurring contractual relations, present and past experience with incentives and trust-based voluntary cooperation coexist and might influence effort choice. Our goal is to separate the channels of present and past experience with incentives, as well as experience with trust-based voluntary cooperation, using a systematic and highly comparable series of eight laboratory gift-exchange experiments with designs inspired by previous evidence on the consequences of incentives for voluntary cooperation.

We are guided by three main questions. First, how does the presence of explicit incentives influence voluntary cooperation when incentives alone cannot achieve efficiency? Second, does the experience of explicit incentives have 'spillover effects' on subsequent voluntary cooperation even when there are no explicit incentives present anymore? Third, how does experience with voluntary cooperation before being exposed to incentive contracts influence crowding-out in the presence of incentives and their possible spillover effects on behavior under contracts without incentives?

Some answers to the first two questions already exist in the literature (see, e.g., Bowles, 2008; Bowles & Polania-Reyes, 2012), but the evidence does not come from comparable designs. For instance, laboratory evidence from gift-exchange market games (e.g., Fehr & Gächter, 2002), trust games (e.g., Bohnet, Frey & Huck, 2001), or common pool resource games (e.g., Cardenas, Stranlund & Willis, 2000) suggests that the presence of incentives may crowd out voluntary cooperation. Similarly, evidence from laboratory public goods games (e.g., Falkinger, Fehr, Gächter & Winter-Ebmer, 2000) and field evidence (e.g., Burks, Carpenter & Goette, 2009; Gneezy & Rustichini, 2000) suggest that incentives can have spillover effects on subsequent performance even after they have been abolished. To our knowledge, little is known about our third question. Our goals therefore are (i) to use the power of a comparable set of laboratory gift-exchange experiments to provide answers to the three questions posed above and (ii) to provide a unified econometric framework that explains effort choice in terms of incentives and positive and negative reciprocity, which are well-established behavioral motivations (e.g., Fehr & Gächter, 2000).

The core argument explaining why incentives may crowd out voluntary cooperation is as follows. The psychological sources of cooperation are social preferences like concerns for fairness and equity, reciprocity and guilt aversion, loyalty and goodwill, or social norms and social esteem (all formalized in various theories).[1] By contrast, explicit incentives are, by design, a direct appeal to people's self-interest and, therefore, in conflict with other-regarding concerns. Incentives might also convey mistrust and trigger 'control aversion' (e.g., Falk & Kosfeld, 2006; Schmelz & Bowles, 2021; Schmelz & Ziegelmeyer, 2020; Ziegelmeyer, Schmelz & Ploner, 2012). The general point is that trust contracts and incentive contracts send psychologically conflictual signals to which agents may react differently.

There are at least three reasons why we believe that our research questions are important. First, the presence of explicit incentives in otherwise incomplete contracts raises the question whether 'material interests' and the 'moral sentiments' as expressed in voluntary cooperation are separable, that is, whether incentives and voluntary cooperation are independent of the levels of the other: Can we

[1] For *fairness and equity*, see Akerlof (1982); Bolton and Ockenfels (2000); Cox, Friedman and Sadiraj (2008); Fehr and Schmidt (1999); for *reciprocity*, see Dufwenberg and Kirchsteiger (2004); Falk and Fischbacher (2006); Rabin (1993); for *guilt aversion*, see Battigalli and Dufwenberg (2007); for *loyalty and goodwill*, see Bewley (1999); Simon (1991); and for *social norms and social esteem*, see Andreoni and Bernheim (2009); Bénabou and Tirole (2006); Ellingsen and Johannesson (2008); Sliwka (2007).

https://doi.org/10.1017/eec.2024.14 Published online by Cambridge University Press

add voluntary cooperation on top of what incentives induce the agent to do, or do incentives per se influence the extent of cooperation agents are willing to exert? More precisely, a *failure of separability* occurs if, under a given contract, voluntary cooperation – actual effort exceeding best-reply effort – is different than the voluntary cooperation agents are willing to provide under a comparable trust contract.

As Bowles and Hwang (2008) argue, separability is an often-invoked assumption, but the psychological differences between incentives and voluntary cooperation suggest separability might not hold. If separability fails (suggested by Bowles, 2008, 2014, 2016; Bowles & Polania-Reyes, 2012), incentives may be overused or underused, which has implications for mechanism design (e.g., Bowles & Hwang, 2008; for a discussion of these issues in a broader context, see Besley & Ghatak, 2018; Kranton, 2019).

Second, in many economic relationships, agents might have past experience with trust and reciprocity and/or with explicit incentives, either because of policy changes within an organization or because of different regimes agents might experience when moving to organizations with different incentive policies. Specifically, the experience of explicit incentives may have spillover effects on voluntary cooperation even if explicit incentives are no longer present. This possibility is suggested by literature on history dependence and learning (e.g., Cooper & Kagel, 2016; Cooper & Stockman, 2011; Rand & Peysakhovich, 2016). Because explicit incentives are salient appeals to self-interest, self-interested behavior may carry over into situations requiring voluntary cooperation even if explicit incentives are no longer present.[2] Similarly, principals' experience with agents' effort choices might influence their subsequent offered contracts. More precisely, *history dependence* in voluntary cooperation occurs if effort under a given trust contract is influenced by a previously experienced trust or incentive contract. History dependence may therefore support or hinder voluntary cooperation: If people experience voluntary cooperation, it may become salient and thereby support cooperation; similarly, experience with incentive contracts may make self-interest more salient, thereby undermining voluntary cooperation.

Finally, studying the behavioral consequences of performance incentives is important because, fundamentally, many real-world contracts are incomplete, which leaves important aspects unregulated and therefore non-enforceable. As has long been noted, voluntary cooperation is necessary to ensure efficiency under contractual incompleteness (see, e.g., Akerlof, 1982; Bewley, 1999; Bowles, 2003, 2016; Ellingsen, 2024; Fehr, Goette & Zehnder, 2009; Fehr, Klein & Schmidt, 2007; Williamson, 1985). Reciprocity-based voluntary cooperation can be a 'contract enforcement device' (Fehr, Gächter & Kirchsteiger, 1997), which, however, might be in conflict with performance incentives.

In summary, many contractual relationships require voluntary cooperation for their efficient fulfillment. Given this, the behavioral consequences of explicit incentives as appeals to self-interest – both their contemporaneous impact and their spillover effect – may depend on the salience of self-interest and be moderated by the experience people have with voluntary cooperation. Our experiments are designed to systematically test these arguments.

Our analyses are based on laboratory gift-exchange experiments (for surveys, see Charness & Kuhn, 2011; Cooper & Kagel, 2016; Fehr et al., 2009).[3] The gift-exchange game is a two-player game in which a principal offers a fixed wage to an agent. The agent can accept or reject the wage offer. If the agent accepts, they choose an effort level. Effort is costly for the agent and beneficial for the

[2] Related arguments are that (i) extrinsic incentives might crowd out intrinsic motivations such as pursuing activities for their own sake and 'not just for the money' (Deci, Koestner & Ryan, 1999; Frey, 1997) and (ii) that incentives can also change relationships, from goodwill based to a transactional, market-exchange based relationship (e.g., Bowles, 2016; Frey & Jegen, 2001; Gneezy & Rustichini, 2000; Sandel, 2012). An incentive contract may also provide an (unconscious) excuse to behave selfishly, which may allow people to abandon other-regarding concerns ('moral wiggle room' Dana, Weber & Kuang, 2007).

[3] We chose laboratory experiments for two reasons: (i) only the lab allows for the comprehensive investigation of all interaction effects we are interested in (Croson & Gächter, 2010; Falk & Heckman, 2009) and (ii) controlling for self-interest, which will be crucial for our approach, is hardly feasible in the field.

https://doi.org/10.1017/eec.2024.14 Published online by Cambridge University Press

principal. Efficiency requires maximal effort, whereas a self-interested agent will provide minimal effort irrespective of the accepted wage (no voluntary cooperation). Numerous experiments refute this prediction and demonstrate the relevance of voluntary cooperation – wages and effort are positively correlated even in one-shot games.[4] We replicate this finding in a version of the gift-exchange game we call the 'Trust contract.' This will provide the necessary benchmark for the comparisons we are mainly interested in.

The explicit incentives take the form of either a 'Fine contract' – that is, a contractually agreed wage reduction in case actual effort falls short of the desired effort or (in different experiments) a 'Bonus contract,' where the agent receives a contractually agreed additional wage payment if the actual effort is at least as high as the desired effort. Both contracts induce the same material incentives and hence any behavioral difference is a framing effect.

We design the set of feasible contracts such that the maximally enforceable effort (by means of incentive-compatible contracts) is substantially less than the efficient level. Thus, there is room for efficiency-enhancing voluntary cooperation beyond the maximally enforceable level. Our design also allows for an easy distinction between incentive-compatible and non-incentive-compatible (NIC) contracts; the latter are directly comparable to Trust contracts that are NIC by design.

Our research strategy is based on eight experiments organized in three sets. Some of our design elements are inspired by past research, and we will explain the connection in Sections 3 and 4. In the first set of three experiments, we establish some basic facts about history dependence and failure of separability. We investigate how voluntary effort provision is affected (i) after agents experienced explicit incentives (measuring history dependence) and (ii) while agents are exposed to Bonus or Fine contracts (measuring separability).

In the second set of two experiments, we investigate how experience with Trust contracts *before* being exposed to Bonus or Fine contracts affects behavior under incentives and after incentives. Experience with Trust contracts is an interesting contextual variable because the psychology of Trust contracts might set an important reference point before being exposed to incentives. Experience with Trust contracts before being exposed to incentives may blunt the salience of incentives and their focus on self-interest.

The third set of three experiments investigates how implicit incentives coming from repeated interaction affect history effects and separability observed in the first two sets of experiments, where we randomly change pairings across iterations to avoid confounds of separability issues with strategic incentives. Implicit incentives, which allow for sequential reciprocity across rounds of interactions, are arguably a very important feature of many ongoing relationships, and therefore it is important to understand how they, together with the explicit incentives, affect effort behavior.

Our most important results are as follows. We find that the experience of explicit incentives spills over to situations without incentives by 'crowding out' voluntary cooperation. This result establishes that the behavioral consequences of incentives can extend beyond their immediate presence. This effect is largest in repeated relationships, which otherwise strongly increase effort compared to one-shot interactions. Incentives also crowd out voluntary cooperation in the presence of incentives: There is no voluntary cooperation beyond the level induced by incentive-compatible contracts, even though agents are willing to provide higher levels without incentive-compatible contracts. Our econometric analysis shows that crowding-out mostly happens due to reduced wages and not due to changed wage–effort relationships.

[4] Some early gift exchange experiments are discussed by Charness (2004); Charness, Frechette and Kagel (2004); Fehr et al. (1993); Fehr et al. (1997); Fehr et al. (1998); Hannan et al. (2002). Gift exchange has been observed not only in abstract but also in real-effort experiments (e.g., Gächter et al., 2016; Gneezy, 2004; Kujansuu & Schram, 2021). Evidence on gift exchange is not confined to the laboratory. See Barr and Serneels (2009); Englmaier and Leider (2020); Falk (2007); Gneezy and List (2006); Kirchler and Palan (2018); Kube, Maréchal and Puppe (2012) and as examples of field studies on gift exchange. Cohn, Fehr and Goette (2015) show that people who exert gift exchange in the lab also exert it in the field.

https://doi.org/10.1017/eec.2024.14 Published online by Cambridge University Press

Table 1 Games and parameters

Offered contract:	Trust game	Fine game	Bonus game
Fixed wage	$w \in [-700, 700]$	$w \in [-700, 700]$	$w \in [-700, 700]$
Desired effort (=output)	$e^d \in [1, 20]$	$e^d \in [1, 20]$	$e^d \in [1, 20]$
Fine/Bonus	–	$f \in \{0, 24, 52, 80\}$	$b \in \{0, 24, 52, 80\}$
Agent's payoff	$w - c(e)$	$w - c(e)$ if $e \geq e^d$	$w - c(e) + b$ if $e \geq e^d$
		$w - c(e) - f$ if $e < e^d$	$w - c(e)$ if $e < e^d$
Principal's payoff	$35e - w$	$35e - w$ if $e \geq e^d$	$35e - w - b$ if $e \geq e^d$
		$35e - w + f$ if $e < e^d$	$35e - w$ if $e < e^d$

Effort cost: $c(e) = 7e - 7$; Payoff if contract rejected: 0 for both.

2. The stage games and benchmark solutions

2.1. The games

Our tools are gift-exchange games (Fehr & Gächter, 2002; Fehr et al., 1997), summarized in Table 1, and incentive games inspired by Anderhub et al. (2002) and adapted for present purposes. Each game has three stages. The principal first offers the agent a contract. In the Trust game, the contract comprises a fixed wage w and a desired effort e^d (effort can also be interpreted as output). Desired effort can be seen as a minimal form of communication with which the principal sends a message about what they expect from the agent.[5] The contract must obey the restrictions $1 \leq e^d \leq 20$ and $-700 \leq w \leq 700$ in integers (negative wages are allowed because in a benchmark solution (see Section 2.2), wages can become negative). In the Fine and Bonus game, the contract, in addition to w and e^d, also specifies a fine or bonus (see Table 1).

Second, the agent can accept or reject the contract. If the agent rejects, the game ends and both earn nothing. If the agent accepts, he or she enters the third stage and chooses effort e in integers (where $1 \leq e \leq 20$). The agent is not restricted by e^d. This reflects contractual incompleteness because e^d is not enforceable. The stage game ends after the effort choice.

In all games, the principal's return from effort is $35e$ and the agent's cost function is increasing and, for simplicity, linear in effort: $c(e) = 7e - 7$. Each player knows the rules, including all payoff functions, and is informed about all choices made in the game.

In the Trust game, the offered contract only consists of w, e^d. Because w cannot be conditioned on effort, we refer to this game as the 'Trust game.' The principal earns $35e - w$ and the agent earns $w - c(e)$.

The offered contract in the Fine game consists of w, e^d, f, where f represents a fine (it can be interpreted as an announced wage reduction if $e < e^d$). The principal can announce one of four lump-sum fine levels: $f \in \{0, 24, 52, 80\}$. If e is less than e^d, f is subtracted from the agent's wage and the principal's wage bill is reduced accordingly. If e is greater than or equal to e^d, the fine is not imposed.

In the Bonus game, the offered contract contains w, e^d, b, where b is a bonus (an announced wage increase if $e \geq e^d$) with $b \in \{0, 24, 52, 80\}$. If e is greater than or equal to e^d, the bonus is added to the agent's payoff and subtracted from the principal's payoff. If e is less than e^d, the bonus is not due.

We use lump-sum Fine and Bonus as incentives because they are simple and easy to understand. Moreover, they have attractive properties for our purposes, as we show in Section 2.2.

[5] Stipulating a desired effort can be seen as a very minimal form of communication. The literature on communication (e.g., Cardenas et al., 2000) suggest that richer communication than what we allow in our experiments could lead to more voluntary cooperation.

2.2. Stage game benchmark solutions for money-maximizing agents

Trust game: A money-maximizing agent will choose $e = e^{min} = 1$ irrespective of w and therefore the principal will offer the wage that just ensures the agent's acceptance: $w = 1$ (or $w = 0$). The resulting payoffs are 34 money units for the principal and 1 money unit for the agent. This solution is inefficient because the efficient surplus is 567 at $e = e^{max} = 20$.

Fine game and Bonus game: In choosing effort, the agent must consider two alternatives, e equals e^d or e equals 1. Effort e greater than e^d is suboptimal since it causes higher cost without increasing payment. Conditional on e lower than e^d, minimal effort e equals 1 is best because Fine and Bonus payments are independent of e. Hence, the optimal effort level is:

$$e^* = \begin{cases} e^d \text{ if } w - c(e^d) \geq w - f - c(1) \;\Leftrightarrow\; f \geq c(e^d) \text{ or } w + b - c(e^d) \geq w - c(1) \;\Leftrightarrow\; b \geq c(e^d); \\ 1 \text{ otherwise.} \end{cases} \tag{1}$$

Notice that the best-reply efforts are the same in the Fine game and the Bonus game; any behavioral difference for a given contract is therefore due to a framing effect.

The agent's best-reply function (1) is the incentive-compatibility constraint for the principal's contract design problem. For each level of f or b, a maximal level of desired effort exists that satisfies f and b are greater than or equal to $c(e^d)$. Given our parameters, the maximally enforceable effort is 12. Before choosing effort, the agent must accept an offered contract. With the parameters from Table 1, it is optimal for the principal to set w such that the agent is just compensated for his or her effort cost $c(e^*)$; furthermore, the solution to the principal's problem is f and $b = 80$, $e^d = 12$ and $w_f = c(12) = 77$ or $w_b = b - c(12) = -3$ (where w_f (w_b) denotes the wage in the Fine (Bonus) game). Accordingly, the agent will accept the contract and choose e equals 12. This solution is more efficient than the solution without incentives, but it does not generate the maximal surplus (the surplus is 343 money units, which goes entirely to the principal).

We set the maximally enforceable effort under incentive-compatible contracts at 12 because this leaves room for voluntary cooperation beyond what incentives can achieve. This design feature reflects contractual incompleteness that characterizes many real-world contracts, even if some aspects can be contractually regulated. By allowing for different Fine and Bonus levels (including zero), we give the principal the possibility to set the strength of the incentives he or she wants to apply to the agent (we included zero because there is evidence that deliberately abstaining from using incentives when incentives could have been used induces more cooperation; Fehr & List, 2004; Fehr & Rockenbach, 2003). Moreover, different combinations of f, b, and e^d that satisfy the incentive-compatibility constraint can induce different best-reply efforts, and this variation allows for a sharper test of whether agents choose best-reply efforts than a more restricted (e.g., binary) set would have allowed for.

Also notice that if the offered contract violates incentive compatibility, e^* equals 1, like in the Trust game. This property will be important in our analysis because it makes Trust contracts and NIC contracts directly comparable.

3. Research questions, experimental design, and procedures

3.1. Research questions and experimental design

Table 2 lists our eight between-subjects experiments. Here, we only describe our experiments; we discuss our behavioral hypotheses in Section 4.

Experiment #1 is TTT, our benchmark. In TTT, participants play three phases, each of which comprises ten one-shot Trust games played in randomly matched pairs. If effort is higher than predicted according to the benchmark solution (i.e., $e > e^*$), we refer to this as 'voluntary cooperation.' This is a standard gift-exchange experiment (see, e.g., Fehr et al., 1997; Fehr, Kirchler, Weichbold &

https://doi.org/10.1017/eec.2024.14 Published online by Cambridge University Press

Table 2 Main research questions and experimental design

Experiment label	Phase 1 (period 1–10)	Phase 2 (period 11–20)	Phase 3 (period 21–30)	# Participants	# Independent matching groups
0. *Establishing a benchmark of voluntary cooperation*					
1. TTT	Trust	Trust	Trust	78	6
A. *Establishing the effects of explicit incentives without prior experience of Trust contracts*					
2. FT	Fine	Trust	–	80	6
3. BT	Bonus	Trust	–	78	6
B. *Explicit incentives after experiencing Trust contracts*					
4. TFT	Trust	Fine	Trust	86	6
5. TBT	Trust	Bonus	Trust	84	6
C. *Explicit incentives after experiencing Trust contracts under implicit incentives in repeated relations*					
6. TTT-R	Trust	Trust	Trust	24	12
7. TFT-R	Trust	Fine	Trust	36	18
8. TBT-R	Trust	Bonus	Trust	34	17

Experiments #1 to #5 are one-shot interactions ('Strangers'), whereas experiments #6 to #8 are run as repeated games in fixed pairs ('Partners,' indicated by the suffix -R). The two-phase experiments are a conceptual replication of Fehr and Gächter (2002).

Gächter, 1998), adapted for our purposes and repeated over thirty periods. Because we observe behavior under Trust contracts across three phases of ten periods each, TTT allows us to observe whether learning leads to the erosion of voluntary cooperation across the phases or whether history dependence (as triggered by the experience of gift-exchange contracts in phase 1) can also lead to stable gift-exchange in later phases.

Against the TTT benchmark, we conduct three sets of experiments. The first set of experiments (the two-phase experiments #2 and #3, labeled FT and BT in panel A in Table 2) are a conceptual replication of similar two-phase experiments by Fehr and Gächter (2002). These experiments aim at (i) measuring the impact of explicit incentives on effort choices, (ii) investigating how incentives affect effort choice with and without incentive-compatible contracts and without prior experience of Trust contracts before being exposed to incentive contracts, (iii) studying a spillover effect of experiencing explicit incentives on effort choice in subsequent Trust contracts and (iv) measuring the role of framing (Fine vs. Bonus contracts). To avoid confounds with strategic incentives, participants play 1-shot experiments in 2 phases of 10 periods each. In phase 1, principals can design either Fine or Bonus contracts (in between-participants treatments), whereas in phase 2 (within-participants), only Trust contracts are feasible. Notice that FT and BT allow us to isolate the behavioral consequences of explicit incentives both while they are present (in phase 1) and after they have been abolished (in phase 2) when the behavioral salience of explicit incentives is unconfounded with prior experience of gift exchange under Trust contracts.

The second set of experiments (#4 and #5 in panel B, labeled TFT and TBT, respectively) extends the basic setting of experiments #2 and #3 (with their respective research questions) by adding a prior experience with gift exchange under Trust contracts. Therefore, TFT and TBT allow for the study of how a history of experience with Trust contracts (in phase 1) influences behavior under incentive contracts (in phase 2) and subsequent Trust contracts (in phase 3).

The third set of experiments is finitely repeated games with the same Partner in all phases (#6 to #8 in panel C, labeled TTT-R, TFT-R, and TBT-R, respectively). These experiments add implicit incentives to the designs and research questions of experiments #4 and #5. There are theoretical and empirical reasons why there are implicit (i.e., strategic) incentives to cooperate: If selfishness and rationality are not common knowledge, cooperation can be sequentially rational

https://doi.org/10.1017/eec.2024.14 Published online by Cambridge University Press

(Kreps, Milgrom, Roberts & Wilson, 1982). Bounded rationality can also lead to cooperation (Selten & Stoecker, 1986). Previous experimental evidence also suggests that cooperation in repeated games of cooperation (including gift-exchange games) is higher than in one-shot games (e.g., Brown, Falk & Fehr, 2004; Embrey, Fréchette & Yuksel, 2018; Falk, Gächter & Kovacs, 1999; Reuben & Suetens, 2012).

3.2. Procedures

We conducted 20 sessions at the University of St. Gallen (Switzerland) with a total of 500 participants (first-year undergraduates of business, economics, or law). We recruited participants by drawing a random selection from a database of volunteer participants and invited them by email. In a typical session, 28 participants were present at the same time.

After arrival at the lab, participants read the instructions (see Appendix A; the same for all) and then were asked to answer control questions on payoff calculations. The experiment did not start before all participants had answered all questions. Roles were assigned at random and fixed throughout the session. We explained that all decisions would be anonymous during the whole experiment. At the beginning of each session, we told participants that there would be different parts and that they would learn about them one after the other.

The experiments were computerized and conducted with 'z-Tree' (Fischbacher, 2007). Participants were separated by partitions and matched anonymously. In sessions with random matching, we formed 2 independent matching groups of 14 participants each. Participants were not informed about the matching groups and were told they would be randomly matched with another person in the room. Participants also never learned the identity of their opponent. Each session lasted two hours. Participants earned on average CHF 45 (approximately €30 at the time of the study).

4. Hypotheses

There is ample evidence for voluntary cooperation under Trust contracts even with Stranger matching (see, e.g., Cooper & Kagel, 2016; Drouvelis, 2021; Fehr & Charness, 2024; Fehr & Gächter, 1998; for overviews). This violates the assumption of selfish (money-maximizing) rationality, which predicts that effort and offered wage are minimal in one-shot games as well as finitely repeated games. However, if one assumes social preferences, a 'trust-reciprocity' or 'gift-exchange' mechanism is possible (Akerlof, 1982; shown in, e.g., Berg, Dickhaut & McCabe, 1995; Fehr, Kirchsteiger & Riedl, 1993; Fehr et al., 1997): The principal offers a substantial fixed wage trusting that the agent will respond in kind by choosing an above-minimal effort level.

A positive wage–effort relationship is a well-established empirical regularity that we expect to replicate with Stranger and Partner matching (see, e.g., Brown et al., 2004; Falk et al., 1999; Gächter & Falk, 2002). In Partner matching, some trust by the principal is necessary in the beginning, but in later periods the principal as well as the agent might respond in kind to each other's previous choice. We refer to this as the 'sequential reciprocity' mechanism. We expect sequential reciprocity across periods to be more powerful than one-shot trust reciprocity in inducing higher wages and obtaining higher effort conditional on wage.

We also expect to find strong behavioral effects of explicit incentives (e.g., Anderhub et al., 2002; Dickinson, 1999; Dickinson & Villeval, 2008; Gächter et al., 2016). In our setting, the higher the fines or bonuses are, the higher the effort will be because the higher effort is the best reply provided the contract is set up to be incentive compatible (Section 2.2). Our design allows for a sharp test of best-reply predictions because contracts can be incentive-compatible or not. If a contract is not incentive compatible, e^* equals 1 regardless of other features of the contract (the same as for Trust contracts; see equation (1)). If a contract is incentive compatible, there are 12 possible best-reply effort levels, and we can compare behavior against them.

https://doi.org/10.1017/eec.2024.14 Published online by Cambridge University Press

Assuming we replicate these well-established psychological mechanisms, we investigate how explicit incentives interact with voluntary cooperation under Trust contracts. We focus on three main dimensions of this question: (i) How does the experience of incentive contracting influence voluntary cooperation in subsequent Trust contracting? (ii) How do explicit incentives affect effort choices when contracts are or are not incentive compatible? (ii) How does experience with Trust contracts before being exposed to incentive contracts change the results obtained to questions (i) and (ii)?

Regarding our first main question, we consider the effects induced by history dependence and learning. For instance, in the context of a step-level public goods game, Cooper and Stockman (2011) have shown that cooperation was influenced by experience in the first half of the experiment that manipulated either monetary concerns or fairness concerns (see also Cooper & Kagel, 2016). The paths of cooperation were different depending on the starting experience, but behavior converged over time. In our context, prior experience with Trust contracts can create a different history dependence than prior experience with incentive contracts. According to evidence on cooperation in gift-exchange games, in our TTT setting, to which we will compare FT/BT and TFT/TBT, agents' effort choices will only depend on the wage offer. This should hold at least for phase 1, whose length of ten periods is comparable to most gift-exchange experiments. If phase 1 of TTT creates a precedence of gift exchange, history dependence in TTT may result in a wage–effort relationship that is stable in phases 2 and 3. By contrast, if agents in phase 1 of FT/BT learn their self-regarding incentives, gift exchange may dissipate, and effort in phase 2 of FT/BT may approach minimal levels.

There are three reasons why prior experience with incentive contracts can create a different history dependence than prior experience with Trust contracts: First, because an incentive contract appeals to the agent's self-interest by communicating the monetary consequences for the agent of complying and violating the contract, it could shift agents' attention to monetary incentives, thereby inducing a larger fraction of agents to choose minimal effort. Second, experience with incentive contracts might induce agents who still cooperate voluntarily to respond with somewhat lower effort – that is, it might weaken the reciprocal wage–effort relationship. Third, prior experience with incentive contracts, and the consequences this has for agents, could diminish principals' trust in the agents' willingness to respond in kind. Consequently, the principal might offer lower wages, which in turn reduces the agent's effort response. Thus, our first hypothesis is:

Hypothesis 1 *Compared to experiencing Trust contracts, experiencing incentive contracts reduces the amount of voluntary cooperation under subsequent Trust contracting.*

We refine our first hypothesis depending on different experimental conditions as follows: We predict lower effort in phase 2 of treatments FT and BT than in phase 2 of TTT (Hypothesis 1a). The reason is that the explicit incentives of phase 1 of FT/BT may focus the agent on their self-interest, which then spills over into phase 2, whereas agents in phase 1 of TTT likely experience gift exchange where history dependence sustains gift exchange in phase 2 as well. Similarly, we predict smaller effort in phase 3 of TFT and TBT (after experiencing incentives in phase 2) than in phase 3 of TTT (Hypothesis 1b).

Because in TFT and TBT, participants can experience cooperation under Trust contracts in phase 1, we predict higher effort levels in phase 3 of TFT and TBT than in phase 2 of FT and BT (Hypothesis 1c). The reason is that in TFT/TBT, the salience of self-interest in phase 2 is now potentially moderated by the experience of gift exchange in phase 1. Finally, we predict lower effort levels in phase 3 of treatments TFT-R and TBT-R than in phase 3 of TTT-R (Hypothesis 1d).

Since the main hypothesis hinges fundamentally on reciprocal behavior, documented amply in previous gift-exchange experiments, we formulate:

Hypothesis 2 *Under Trust contracting, effort choices respond reciprocally to the offered wage.*

Specifically, a higher wage offer by the principal will reduce the probability of rejecting the contract (Hypothesis 2a; also found by Anderhub et al., 2002) and the probability of minimal effort given the contract has been accepted (Hypothesis 2b). Furthermore, given the contract has been accepted and effort is higher than minimal, effort correlates positively with wage (Hypothesis 2c). These effects will be at least as strong under Partner matching than under Stranger matching because in Partner matching, strategic incentives (sequential reciprocity) across periods exist, and they will likely strengthen reciprocity.

We also expect that explicit incentives strongly influence effort in our settings (see, e.g., Anderhub et al., 2002; Gächter et al., 2016) and formulate this as our third hypothesis:

Hypothesis 3 *Stronger monetary incentives induce higher effort.*

We also investigate whether the framing of incentives as fine or bonus is important. Existing evidence is mixed. For instance, Fehr and Gächter (2002) and Fehr et al. (2007) find that, in settings similar to ours, bonus contracts induce higher efforts than fine contracts. de Quidt, Fallucchi, Kölle, Nosenzo and Quercia (2017) find mixed evidence for framing effects in a small survey of the existing literature and find no difference between Fine and Bonus contracts in their real-effort experiment. Hence, because our design is closest to two studies that find evidence for contract framing, we formulate the following:

Hypothesis 4 *Framing incentives as fine or bonus matters. Effort will be higher under Bonus contracts than under Fine contracts.*

Finally, with incentive contracts, issues of incentive compatibility and whether the offered contract satisfies the agent's participation constraint arise. We expect that agents will reject contracts that violate the participation constraint. Regarding our second main question, because incentive contracts appeal to agents' self-interest, we expect that incentive contracts diminish ('crowd out') voluntary cooperation both when they are incentive compatible and when they are not (in which case they are incentive equivalent to Trust contracts; see, e.g., Fehr & Gächter, 2002). In other words, we expect separability to fail. We thus have formulated the following hypothesis:

Hypothesis 5 *Under Stranger matching, contracts that do not satisfy the participation constraint are rejected (Hypothesis 5a). Accepted incentive contracts induce an effort level that is at least as large as the theoretical best-reply effort implied by selfish rationality (Hypothesis 5b) but lower voluntary cooperation (the difference between effort and best-reply effort) than in Trust contracts (Hypothesis 5c). We also predict that NIC contracts – which are incentive equivalent to Trust contracts – perform worse than incentive-compatible contracts and comparable Trust contracts (Hypothesis 5d).*

Because real-world relationships are often ongoing with the same partners, we study all issues raised above in a Stranger versus Partner comparison. Based on (i) the theoretical argument that strategic incentives in the form of sequential reciprocity available in Partner but not in Stranger should increase cooperation in Partner cooperated to Stranger, and (ii) evidence consistent with (i) (Falk et al., 1999; Gächter & Falk, 2002), we predict:

Hypothesis 6 *Partner matching induces higher effort than Stranger matching.*

In the data analyses we elaborate further on these hypotheses and specify the subsets of data we use to study them.

https://doi.org/10.1017/eec.2024.14 Published online by Cambridge University Press

Table 3 Mean statistics by contract type and matching condition

	Stranger condition			Partner condition		
Contract type	Trust	Fine	Bonus	Trust	Fine	Bonus
Wage	79.9	140.7	80.5	256.8	286	194.1
Fine/Bonus	–	74	70.6	–	58.8	67.3
Desired effort	7.3	10.8	10.8	14.8	16.2	14.8
Incentive compatible	–	72.8%	69.6%	–	19.4%	33.5%
Contract acceptance	77.2%	84.6%	82.7%	89.8%	87.8%	81.2%
Actual effort	3.7	7.1	6.9	13.3	13.9	13.5
Profit principal	4.7	95	76.6	146.1	168.2	150.1
Profit agent	81.1	79.4	89.6	193.4	179.6	162
Total profit	85.8	174.4	166.2	339.4	347.7	312
# Observations	3,660	830	810	1,060	180	170

The data used here are all experimental decisions under the respective contract type (Trust, Fine, Bonus), separately displayed by matching protocol (Stranger, Partner).

5. Results

This results section is structured as follows. Before we investigate our hypotheses (see Section 4), we start by providing an overview of mean statistics of wages, bonus and fine, desired and actual effort, and profits across the three main contract types of Trust, Fine, and Bonus contracts, as well as the two matching protocols Strangers and Partners, followed by an initial analysis of the main treatment outcomes in terms of average effort levels (Section 5.1). Section 5.2 investigates the behavioral mechanisms behind effort *after* experiencing Fine or Bonus contracts, and Section 5.3 studies determinants of effort choice under Fine and Bonus contracts, distinguishing whether the contracts are incentive compatible or not. Section 5.4 provides a comparison of treatments in terms of predicted effort choices.

5.1. A descriptive overview

5.1.1 Mean statistics by contract type and matching condition

Table 3 shows mean statistics for all experimental decisions taken by principals and agents, as well as the resulting profits for each type of contract and each matching condition. At this stage, we do not distinguish between phases. The purpose is to provide an overview of how the main treatment conditions affect the main decision variables in our experiment.

Table 3 documents that wage, desired effort, actual effort, and profits are substantially higher under Partner rather than Stranger matching. Contract acceptance is 81.2% across conditions. Mean fine and bonus are 74 and 70.6 under Stranger conditions (and the rate of incentive-compatible contracts is 72.8% and 69.6%, respectively), indicating that if incentives are available, most principals choose high-powered incentives (recall that maximum incentives are 80). Furthermore, the mean fines and bonuses are smaller under Partner matching (means are 58.8 and 67.3), the rate of incentive-compatible contracts is much lower than in Stranger matching (19.4% and 33.5%), and the desired effort is above 14 in all Partner conditions. These facts reflect that explicit incentives are less important in repeated games than in one-shot games. Wages are higher for Fine contracts than Bonus contracts to compensate for the fact that a fine reduces payment while a bonus increases it. In Stranger, the effort is higher under Fine and Bonus contracts than under Trust contracting.

The relative increase of effort in incentive contracts is larger in Stranger than Partner, which suggests that explicit incentives are especially effective in short-term relationships. As expected, effort

https://doi.org/10.1017/eec.2024.14 Published online by Cambridge University Press

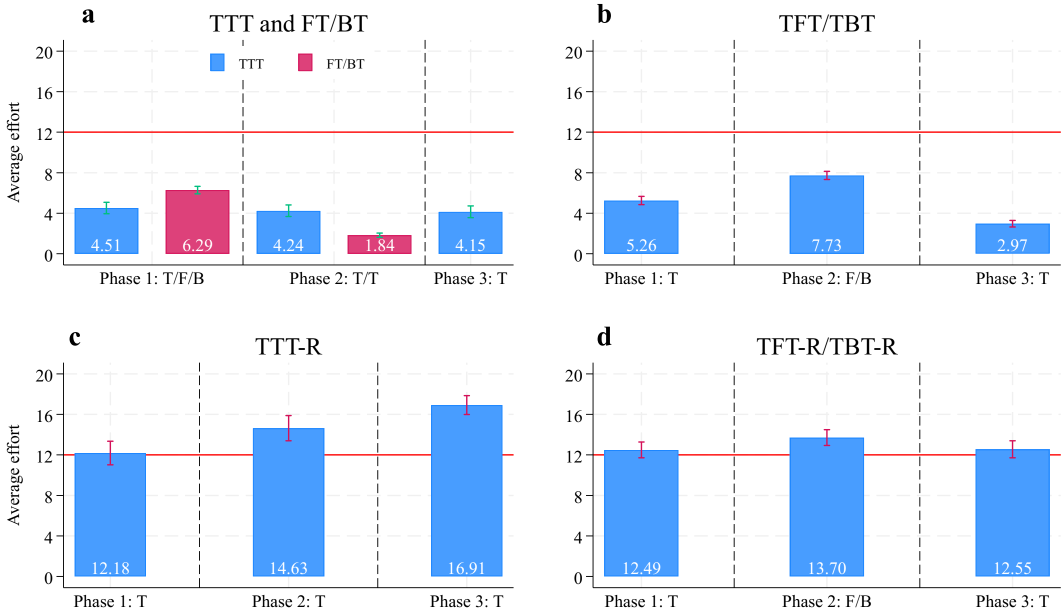

Fig. 1 *Average effort across main experimental conditions*. Panels *a* and *b* are the results of one-shot ('Stranger') matching and panels *c* and *d* of finitely repeated ('Partner') matching (indicated by the suffix -R). The horizontal line at effort equaling 12 indicates the benchmark of the theoretically maximal effort implementable by incentive contracts. the numbers in the bars are average effort levels. error bars are 95% confidence intervals. see Online Appendix Figure B1 for the average effort by period and type of incentive (Fine or Bonus contract)

levels are substantially higher in Partner than Stranger. Interestingly, the profit share captured by principals is particularly low in Stranger Trust: It is only 5.5% (= 4.7/85.8), whereas the profit share of principals is 43% under Partner Trust and even higher with incentive contracts (Stranger and Partner). Thus, our gift-exchange games set a particularly hard task for the principal to achieve beneficial cooperation under Stranger conditions.

5.1.2 *Effort by treatments and phases*

Figure 1 displays the mean effort levels of accepted contracts for each phase and main treatment. It serves to provide initial insights regarding our research questions, as outlined in Table 2. In Sections 5.2 and 5.3, we report detailed statistical tests in terms of the behavioral mechanisms that have produced the results documented in Fig. 1. As we will see, from the agents' perspective, the behavioral mechanisms are positive reciprocity in the form of higher effort for higher wages, negative reciprocity in the form of choosing minimal effort (and rejecting contracts), and reacting to explicit and implicit incentives. For the principal, the main behavioral mechanism is the design of the offered compensation (OC) package.

In this subsection, we focus on effort levels and how they change across the various treatments. Here, we only report simple tests of whether changes in effort levels are significant (all are based on robust ordinary least squares (OLS) regressions clustered on independent matching groups of effort on relevant phase and treatment dummies); we will present more detailed regressions related to behavioral mechanisms in Sections 5.2 and 5.3.

Panels *a* and *b* of Fig. 1 depict the results from the Strangers conditions and panels *c* and *d* from the Partner conditions (indicated by the suffix -R). For the purposes of this overview, we pool the Fine and Bonus conditions because effort levels are not significantly different from each other in any condition (Table B1 in the Online Appendix, all $p > .202$).

https://doi.org/10.1017/eec.2024.14 Published online by Cambridge University Press

Figure 1a reveals that average effort levels in the benchmark TTT condition are around 4.3, with a slight (but insignificant, $p > .678$) decline from 4.51 in phase 1 to 4.13 in phase 3. Incentives (available in phase 1 of conditions FT/BT) increase effort levels to 6.29, but this increase of 1.77 is rather modest and far off the theoretically predicted effort level of 12 (indicated by the horizontal lines in Fig. 1). Average effort in phase 2 of FT/BT, where no incentives are available any longer, is 1.84 whereas in phase 2 of TTT, average effort is 4.24. This decrease of 2.4 is significant ($p = .009$) and, as we will show in detail in Section 5.2, evidence for a crowding-out effect after experiencing incentive contracts.

In the conditions of TFT/TBT, illustrated in Fig. 1b, incentives are introduced in phase 2 after participants had the experience of Trust contracts in phase 1. Incentive contracts in phase 2 of TFT/TBT increase effort significantly after experiencing Trust contracts in phase 1 (from 5.26 in phase 1 to 7.73 in phase 2, $p = .000$). The average phase 2 effort of 7.73 in TFT/TBT is also significantly higher ($p = .000$) than the average effort of 6.29 in phase 1 of FT/BT – that is, without the prior experience of Trust contracts.[6] In phase 3 of TFT/TBT, the average effort is 2.98, which is lower than in phase 1 and phase 2. To gauge the change in effort that might be due to crowding out of effort after having experienced Trust and Fine/Bonus contracts, we compare the average effort in phase 3 of TFT/TBT (2.98) with the average effort in phase 3 of TTT (4.15): The average drop in effort is – 1.18. Compared to the effort reduction of 2.31 in the FT/BT experiments, the drop in effort is reduced almost by half and is insignificant ($p = .200$).

Figures 1c and 1d illustrate the power of implicit (that is, strategic) incentives in the form of sequential reciprocity, available in the repeated games of the Partner conditions TTT-R (panel *c*) and TFT-R, TBT-R (panel *d*), to significantly ($p = .000$) increase effort levels compared to the Stranger conditions (panels *a* and *b*). This result is consistent with previous evidence that shows the cooperation-enhancing effect of implicit (strategic) incentives available in repeated games compared to one-shot games where strategic incentives are absent (e.g., Falk et al., 1999; Gächter & Falk, 2002). Unlike in the Stranger conditions, effort levels exceed 12 in all phases and treatments of the Partner conditions. In contrast to TTT, in TTT-R, effort significantly increases across the three phases of TTT-R ($p < .015$ for phase 2-and phase-3 dummies).

The introduction of Fine or Bonus contracts in phase 2 of TFT-R/TBT-R (panel *d*) increases effort only insignificantly compared to phase 1 ($p = .103$), and effort in phase 3 of TFT-R/TBT-R is the same as effort in phase 1 ($p = .925$). Comparing effort in phase 3 of panel *d* (average effort of 12.55) with effort in phase 3 of panel *c* (average effort of 16.91) suggests a strong (and highly significant, $p = .001$) possible crowding-out effect in terms of foregone implicit cooperation of 4.36 after having experienced incentive contracts in phase 2.

In summary, regarding effort choices, this overview suggests three important results, which we aim to explain in terms of behavioral mechanisms in the remainder of this paper:

1. The experience of incentive contracts reduces effort in subsequent Trust contracts compared to the relevant experience with Trust contracts only. This drop in effort exists in FT/BT, TFT/TBT and TFT-R/TBT-R, but its effect size varies across settings.
2. Implicit incentives, available only under Partner matching, strongly increase effort compared to Stranger conditions.
3. Compared to Trust contracts, explicit incentives in the form of Fine or Bonus contracts increase effort under Stranger conditions but far less than theoretically predicted. In the presence of implicit incentives in the Partner conditions, explicit incentives do not increase effort compared to Trust contracts.

[6] This increase might to some extent be explained by the offered contracts. Recall from equation (1) that $e^* = e^d$ if the offered contract is incentive compatible. The fraction of incentive-compatible contracts in phase 1 of FT/BT is 67.6% and in phase 2 of TFT/TBT is 74.6%. Desired effort levels e^d are 8.9 (phase 1 of FT/BT) and 9.5 (phase 2 of TFT/TBT).

https://doi.org/10.1017/eec.2024.14 Published online by Cambridge University Press

In the Section 5.2, we examine the details of the behavioral mechanisms behind these results. Two important mechanisms, shown in previous research on gift-exchange experiments, are *trust* and *positive reciprocity*: Principals offer wages above Nash-equilibrium wages (under money maximization; see Section 2), and agents respond with effort levels that increase in the wage offered ('gift exchange' e.g., Fehr et al., 1993, 1997; Hannan, Kagel & Moser, 2002; see also Cooper & Kagel, 2016). Agents may also be motivated by *negative reciprocity* – that is, a willingness to punish principals if the OC is unfavorable for the agent (see Fehr & Gächter, 2000) and also the large literature on rejections of unfair offers in ultimatum games (e.g., Güth & Kocher, 2014; Lin et al., 2020). Negative reciprocity may result in rejecting the contract or in choosing minimal effort after the contract has been accepted. We expect the mechanisms of trust and positive and negative reciprocity to be operative in our experiments as well. We focus on *accepted* contracts (1,265 out of 6,710 = 81.2% across all experiments) and study trust and positive and negative reciprocity in them.[7]

5.2. Effort under trust contracts after experiencing incentive contracts

5.2.1. Trust by principals and agents' reciprocal reward

In this subsection, we first provide graphical evidence of how the various treatments affect effort levels as a function of offered wages (Fig. 2; shown are all individual wage–effort pairs of accepted contracts), followed by regression analyses that quantify the treatment effects (Table 4). Figure 2 provides (jittered) scatterplots of individual effort choice against OC (resp. fixed wage) for each phase of the Stranger matching experiments of FT/BT (panels *a* and *b*), TTT (panels *c* to *e*), and TFT/TBT (panels *f* to *h*). The lower set of panels shows the experiments in Partner matching of TTT-R (panels *i* to *k*) and TFT-R/TBT-R (panels *l* to *n*). Panels *a, g*, and *m* (colored in red) show the relationship between OC and effort under Fine or Bonus contracts for NIC contracts, which are incentive-equivalent to the Trust contracts shown in the other panels (in blue). We discuss them in Section 5.3; here we focus on the Trust contracts displayed in the blue-colored panels. The vertical dotted lines are the average wages offered in the respective treatment.

To varying degrees, the panels of Fig. 2 display three clusters. Cluster (i) shows a positive correlation between effort and fixed wage, indicated by a positively sloped OLS regression line conditional on effort greater than one. This cluster indicates that a substantial fraction of effort choices can be interpreted as reciprocal reward: Agents voluntarily cooperate and respond to higher fixed wage offers by higher effort.

A second cluster, comprised of effort equal to 1, exhibits minimal effort for all levels of fixed wage greater than 35. This either represents selfish exploitation by the agent or reciprocal punishment for a low offered wage. It implies that the agent earns a positive fixed wage at no cost, whereas the principal earns a negative payoff if wage is greater than 35.

A third cluster consists of wages less than 35 and effort equal to 1. Choosing a minimal effort when the offered wage is low can be due to reciprocity or selfishness of the agent. However, it also indicates that the principal shows little trust in these cases.

The three clusters are most pronounced under the Strangers matching protocol (upper set of panels *a* to *h*). Under Partner matching (lower set of panels *i* to *n*), where implicit incentives (sequential reciprocity across rounds) are available, the clusters (effort ≈ 1, wages > 35) and (effort ≈ 1,

[7] Out of the 1,264 rejected contracts, 386 contracts (30.5%) violated the participation constraint (i.e., the OC was negative); agents rejected 95.5% of those contracts. In Online Appendix B, we provide (i) further details across experimental conditions and (ii) an analysis of how contract rejections are related to OC that does not violate the participation constraint (see Table B2). We show that contract rejections when the participation constraint is satisfied is strongly related to the OC, which is evidence of negative reciprocity. We see this at the aggregate level, but we also provide evidence for negative reciprocity at the individual level: for instance, under Stranger conditions, up to 70% of people rejected Trust contracts with low wage offers at least once. See Table B2 for the details.

https://doi.org/10.1017/eec.2024.14 Published online by Cambridge University Press

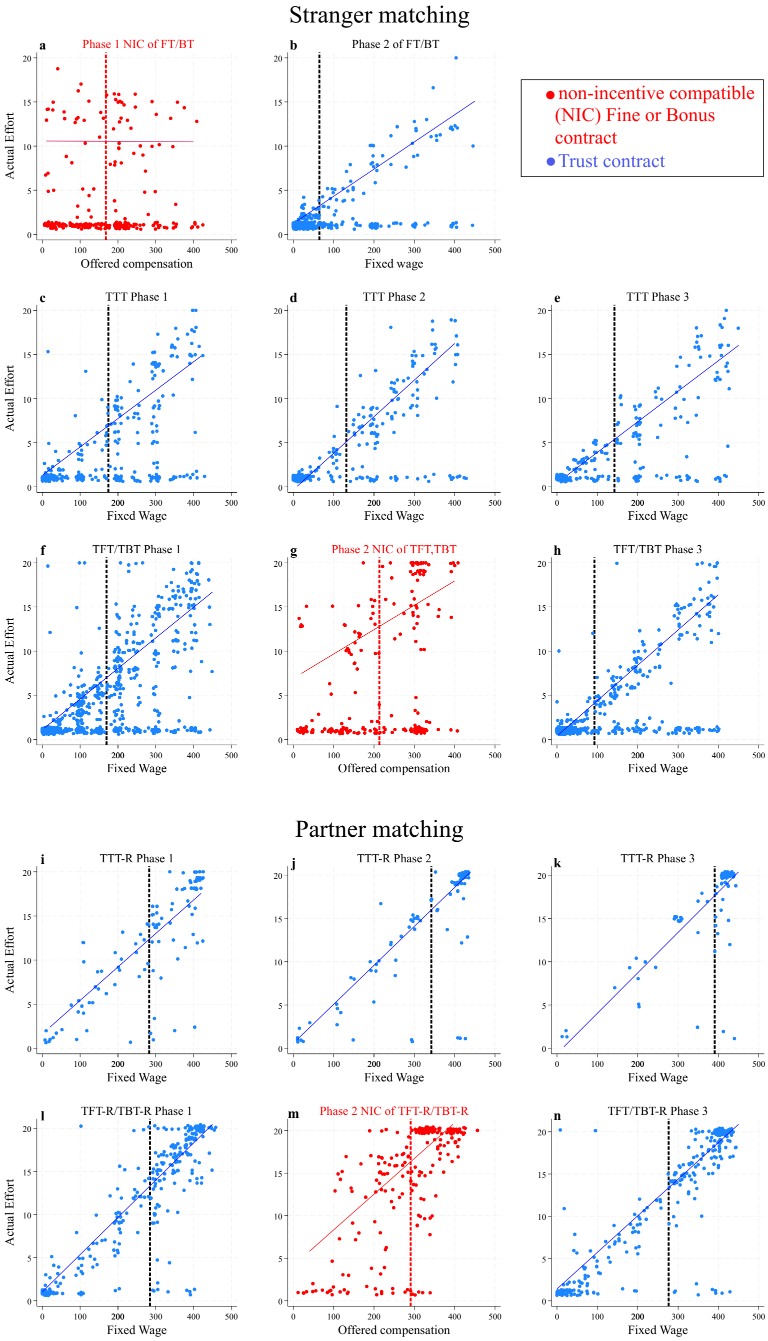

Fig. 2 *The wage–effort relationships in Strangers and Partners and across phases*. Panels a to h illustrate the Stranger matching experiments and panels i to n illustrate the Partner experiments. Dots (jittered) are all individual wage–effort pairs of accepted contracts (up to ten per pair). Phase 1 of FT/BT (panel *a*), phase 2 of TFT/TBT (panel *g*), and phase 2 of TFT-R/TBT-R (panel *m*) show effort under NIC Fine/Bonus contracts that are incentive equivalent to the trust contracts of the respective matching protocol. We discuss these results in Section 5.3. dashed vertical lines are the average accepted wages in a phase and treatment. Solid lines are simple linear regressions of effort >1 on wage (OC) for the respective phase and treatment. A breakdown of Figure 2 by contract type and accompanying regressions are in Online Appendix B, Figure B2.

https://doi.org/10.1017/eec.2024.14 Published online by Cambridge University Press

Table 4 Regression analysis to explain effort choices after the experience of incentive contracts

	Comparing ...			
Treatment	F**T**/B**T** with T**T**T	T**F**T/T**B**T with T**T**T	F**T**/B**T** with TF**T**/TB**T**	TF**T**-R/TB**T**-R with TT**T**-R
Table 4.1: Probit; dependent variable: *effort* = 1				
Model	(1a)	(1b)	(1c)	(1d)
Wage	−.540*** (.052)	−.569*** (.055)	−.562*** (.049)	−.647*** (.076)
Treatment	.429** (.171)	.257 (.205)	.219 (.189)	.438 (.358)
Constant	1.005*** (.152)	.959*** (.154)	1.233*** (.136)	−.045 (.371)
Obs.	876	929	1,240	426
Pseudo R2	.241	.246	.215	.440
Table 4.2: OLS; dependent variable: *effort* > 1				
Model	(2a)	(2b)	(2c)	(2d)
Wage	3.449*** (.224)	3.746*** (.202)	3.558*** (.224)	4.157*** (.258)
Treatment	−.593 (.443)	1.074** (.490)	−.783 (.473)	1.033* (.581)
Constant	1.231** (.583)	−.184 (.458)	1.307** (.497)	.962 (1.089)
Obs.	215	300	276	365
R-squared	.817	.748	.782	.763
Table 4.3: OLS; dependent variable: *wage*				
Model	(3a)	(3b)	(3c)	(3d)
Treatment	−67.043** (25.355)	−49.757 (29.529)	−28.581 (17.429)	−113.970*** (26.514)
Constant	135.344*** (23.936)	145.315*** (24.871)	91.317*** (15.07)	399.799*** (15.62)
Obs.	876	929	1,240	426
R-squared	.078	.033	.025	.119

The compared phases of respective Trust contracts are in bold. In all regressions, the dataset comprises all accepted Trust contracts. All estimations include dummies for periods 1–3 and periods 8–10 to control for (noisy) initial and end behavior; the omitted benchmark category is the central periods 4–7. The full estimation results are in Online Appendix B, Table B3. Table 4.1: The dependent variable is coded as 1 if minimal effort (i.e., effort = 1) is chosen and as 0 otherwise. Table 4.2: Regressions are conditional on effort > 1. Table 4.3: The dependent variable is the offered (and accepted) wage. In Tables 4.1 and 4.2, wage is measured in units of 100. Treatment is a dummy variable that changes between models (but is the same in the column): Models *a*: FT/BT = 1; Models *b*: TFT/TBT = 1; Models *c*: FT/BT = 1; Models *d*: TFT-R/TBT-R = 1. All regressions are robust and clustered on independent matching groups. *** p < .01; ** p < .05; * p < .1.

wages < 35) become thinner, and a new cluster of wages around 400 and efforts between 18 and 20 appears.

Before we turn to a detailed econometric analysis of our results, we point out a couple of noteworthy features of the data shown in Fig. 2. First, the clusters of effort greater than one have remarkably similar slopes across the phases with Trust contracts, which implies that learning does not diminish the reciprocal wage–effort relationship. To our knowledge, this stability is a novel result because most previous gift-exchange experiments were only run for up to ten periods. Second, comparing the upper set of panels with the lower set of panels

https://doi.org/10.1017/eec.2024.14 Published online by Cambridge University Press

(i.e., Strangers with Partners) illustrates the power of implicit incentives to increase wages and effort levels.

In this section, we employ robust regression analyses (clustered on independent matching groups) to answer our main research question: How are trust and reciprocal effort choice under Trust contracts affected *after* the experience of Fine or Bonus incentive contracts? We separate out treatment effects by comparing effort under Trust contracts after participants experienced Fine or Bonus contracts with effort under Trust contracts where the prior experience is with Trust contracts. We make this comparison for three dependent variables (Table 4): We analyze the frequency of minimal effort choices (Table 4.1), effort conditional on above-minimal effort (Table 4.2), and the principals' fixed wage choices (Table 4.3). All three effects together are responsible for the overall change of mean effort between treatments (see Fig. 1), and they correspond to our discussion of psychological mechanisms in the hypothesis section above.

For each of the three dependent variables (Tables 4.1 to 4.3), we report four treatment comparisons in the four columns of Table 4 (models *a* to *d*): We compare phase 2 of FT/BT with phase 2 of TTT, phase 3 of TFT/TBT with phase 3 of TTT, phase 2 of FT/BT with phase 3 of TFT/TBT, and phase 3 of TFT-R/TBT-R with phase 3 of TTT-R. These four comparisons (indicated by bolded letters) correspond to Hypotheses 1a to 1d, which predict that experiencing incentive contracts crowds out voluntary cooperation (that is, reduces effort) under subsequent Trust contracts compared to how effort has evolved under Trust contracts in phase 2 or in phase 3 of the respective experiments. Each model has two main explanatory variables: Wage (measured in units of 100 to display coefficients with fewer decimals) and Treatment (where Treatment is a dummy that corresponds to one of the comparison treatments depending on the model, as indicated in the top row of Table 4). We also control for initial and end-period effects (dummies for periods 1 to 3 and 8 to 10) that potentially may have influenced effort choices differently across rounds (see Fig. B1 in the Online Appendix). Because there are no systematic period effects and to keep the exposition simple, we only report the main variables here and relegate the full estimation results to Online Appendix B (Table B3).

Table 4.1 shows Probit regressions for the frequency of minimal effort (coded as 1) or above-minimal effort (coded as 0). The estimated coefficient of *Treatment* is positive and significant for model 1a – that is, the probability of agents choosing minimal effort is higher in phase 2 of FT/BT than in phase 2 of TTT. The size and significance of this effect (models 1b and 1c) is reduced in TFT/TBT, where agents have experience with Trust contracts from phase 1 *before* being exposed to incentive contracts in phase 2. In the Partner matching protocol, the comparison of phase 3 of TFT/TBT-R with phase 3 of TTT-R shows that the coefficient of Treatment is high but noisy and insignificant, which is likely due to very few observations at effort equal to 1 (see Fig. 2, panels *k* and *n*).

The estimated coefficient of *Wage* is negative and highly significant in all models – that is, the probability of an agent choosing minimal effort decreases in wage. This supports the interpretation of minimal effort as reciprocal punishment for low-wage offers (Hypothesis 2b) and rejects the interpretation of minimal effort as selfish rationality, according to which agents choose minimal effort independent of wage – namely, the wage coefficient should be insignificant.

Table 4.2 shows OLS regressions of effort conditional on above-minimal effort (effort > 1) on Wage and Treatment. The model structure is the same as in Table 4.1. Consistent with reciprocal gift exchange, Wage has a positive and highly significant influence across all models 2a to 2d. For instance, the coefficient 3.449 in model 2a means that increasing wage by 100 units increases effort by 3.449 units. This supports Hypothesis 2a, which predicts a positive wage–effort relationship.

Treatment plays a minor role here. The only significant effect (at $p < .05$) of Treatment is an increase of effort in phase 3 of TFT/TBT compared to phase 3 of TTT (model 2b); in model 2d (phase 3 of TFT-R/TBT-R compared to phase 3 of TTT-R), the effect of Treatment is marginally significant ($p = .083$). However, as we will show in Section 5.4, this is overcompensated by the countervailing

https://doi.org/10.1017/eec.2024.14 Published online by Cambridge University Press

effects of a higher probability of minimal effort and reduced trust. We conclude that, for a substantial fraction of participants who choose above-minimal effort ($e > 1$), a positive wage–effort relationship is intact even after experiencing incentives in the previous phase.[8]

Finally, we investigate whether the level of trust shown by the principal as expressed by their wage offer is lower after experiencing agents' behavior under incentive contracts. Visual inspection of Fig. 2 suggests that average wages (indicated by the dashed lines) in the Trust phase after Fine/Bonus contracting are lower than after the respective phase of Trust contracting in TTT or TTT-R (compare Fig. 2b with 2d, Fig. 2h with 2e, and Fig. 2n with 2k). To assess the treatment-specific size of reduced wage offers, Table 4.3 shows OLS regressions of fixed wage based on Treatment (analogous to Tables 4.1 and 4.2). All estimated coefficients of Treatment are negative (and significant in models 3a and 3d), which implies that principals are more cautious (less trusting) in their wage offers after experiencing their agents' effort behavior under Fine/Bonus contracts compared to Trust contracts in the previous phase.

Although our focus is mostly on agents' effort choice as a function of the contract they receive, it is worth noticing that the offered contracts may also react to the agents' previous effort choices. In a regression analysis that complements the results of Table 4.3, we find that average wage offers in a phase after incentives (that is, in phase 2 of FT/BT, phase 3 of TFT/TBT, and phase 3 of TFT-R/TBT-R) are higher the higher average effort was in the preceding phase. This increase per unit of effort is weak and insignificant in FT/BT (6.9 per unit, $p = .180$) but substantial and significant in TFT/TBT (12.7 per unit, $p = .019$) and in particular in TFT-R/TBT-R (19.9 per unit, $p = .000$). Average wage offers also react positively and significantly to average effort choices in preceding phases in the TTT experiments (offered wages increased on average by 21.3 per effort unit in both phase 2 and phase 3 as a reaction to phase 1 (phase 2). In TTT-R, the offered wage in phase 2 also increased by 21.3 per unit of effort in phase 1, but in phase 3 the average wage only increased by 4.9 per unit of effort ($p = .081$). This is likely due to a ceiling effect because wages and effort are already very high in phase 2 (see Fig. 1c and Fig. 2i, j, k). In sum, trust by principals is also a reaction to the principals' experience with agents' trustworthiness as revealed by their effort choices.

Taken together, these detailed analyses support the impressions from Fig. 1: Prior experience of incentive contracting reduces mean effort under Trust contracting (Hypothesis 1). This 'crowding-out effect' is mainly driven by two behavioral responses: A reduction in the level of trust by the principal as expressed by lower wage offers (which are a reaction to the received effort choices) and an increased probability of minimal effort by the agent conditional on wage (more frequent reciprocal punishment of a low wage offer). Interestingly, there remains a substantial fraction of effort choices in line with a positive wage–effort correlation (effort as a reciprocal reward) that is also similar between phase 1 and phase 3 of TFT/TBT and TFT-R/TBT-R, respectively (compare Fig. 2 panels *f* and *h*, and *l* and *n*).

[8] We should add two comments to these results. First, we acknowledge that there might be a selection effect going on here on who ends up in the effort equal to 1 or the effort greater than 1 cluster, as shown in Fig. 2. For lack of an instrumental variable, there is no obvious remedy for this problem. The econometric analysis reported here is therefore descriptive, aimed at quantifying the treatment comparisons in the four models we study here. Second, in the models reported in Table 4.2, we kept it simple and did not include interaction effects of Wage with Treatment. Including them does not change the main result of a positive wage–effort relationship after experiencing incentives in the preceding phase. Including an interaction effect (Wage*Treatment) changes the estimated coefficients of Wage slightly, but all four coefficients remain significant at p less than .001. In light of Fig. 2, this result is not surprising. However, the interaction variable Wage*Treatment is significantly negative in models 2a and 2c (coefficients (se) are − 1.173 (.191) and − .984 (.294)) and insignificant in models 2b and 2d. Including the interaction effect renders Treatment significant in model 2a (coeff (se): 1.73 (.463), $p < .01$); weakly significantly positive in model 2c; and insignificant in models 2b and 2d. These results are also consistent with OLS and Tobit regressions of Effort on Wage run for each of the eight experiments separately. See Online Appendix, Fig. B2, and the accompanying OLS and Tobit regressions.

https://doi.org/10.1017/eec.2024.14 Published online by Cambridge University Press

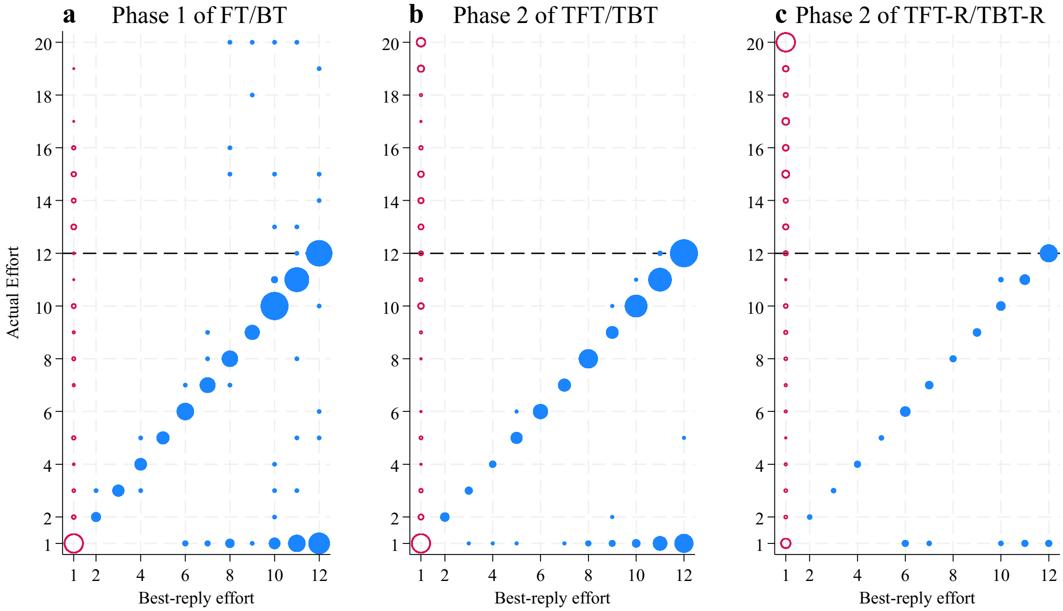

Fig. 3 *Actual effort and best-reply effort in phases with Fine/Bonus contracts.* panel *a*: phase 1 of FT/BT (panel *a*), panel *b*: phase 2 of TFT/TBT, and panel *c*: phase 3 of TFT-R/TBT-R. The size of the dots is proportional to the number of underlying observations. Blue dots: effort choices if $e^* > 1$; red dots: effort choices 1. The horizontal line at 12 indicates the maximally enforceable effort level under incentive-compatible contracts. see Figure B3 in the Online Appendix for a breakdown by contract type

5.3. *Effort under fine and bonus contracts*

5.3.1 *Overview of effort under fine and bonus contracts*

We have seen in Fig. 1 that explicit monetary incentives increase mean effort with Stranger matching but not with Partner matching, where implicit incentives in the form of strategic sequential reciprocity across rounds are feasible. We will now examine these and other questions regarding the effectiveness of incentive contracts in detail.

Figure 3 illustrates the behavior of agents under explicit monetary incentives. Across all experiments with accepted incentive contracts, 59.2% (988 out of 1,668 contracts) are incentive compatible and 40.8% are not incentive compatible. Figure 3 displays scatterplots of observed effort choices against the theoretical best-reply effort e^* (assuming rational money maximization). Figure 3 is remarkable because distributions are highly structured, and there is little noise. There are three clusters that correspond to different behavioral modes. The clustering looks much stronger than the one we described in Fig. 2, which was strong already. It provides some answers to our hypotheses even without statistical testing.

The *first cluster* – 746 out of 969, or 75.5%, of effort choices of accepted incentive-compatible contracts with e^* is greater than 1 – are on the 45-degree line, that is, effort choices that are exactly equal to best-reply effort e^* is greater than 1 (fractions are 72.5%, 80.3%, and 83% in panels *a* to *c*, respectively). This result is a conceptual replication of Anderhub et al. (2002) – who found that two-thirds of their agents chose best-reply effort – and clear evidence for Hypothesis 3 (higher incentives induce higher effort). If best-reply calls for effort greater than 1, there is (almost) no voluntary cooperation beyond the best-reply effort level at all. This fact supports Hypothesis 5c (reduction of voluntary cooperation under incentives) and is shown most clearly in panels *b* and *c* – that is, in phase 2 data.

In the *second cluster*, the best-reply effort is greater than 1, but in a substantial fraction of cases (182/969 = 18.8%), agents deviate to the minimal effort equal to 1 (fractions are 19.9%, 18.2%, and 15.1% in panels *a* to *c*, respectively). Across all three settings of Fig. 3, at the individual level, 47.7% of

https://doi.org/10.1017/eec.2024.14 Published online by Cambridge University Press

agents chose minimal effort at least once. Similar to our discussion of behavior under Trust contracts, this might be an expression of negative reciprocity due to the dissatisfaction of the agent with the OC.

The *third cluster* is the distribution at best-reply effort equal to 1 (242, 214, and 243 cases in panels *a* to *c*, respectively). A best-reply effort of 1 can occur for three reasons: First, either no fine or bonus has been specified ($f = 0$, $b = 0$; 8.3%, 2.8%, and 10.7% of cases in panels *a* to *c*, respectively), which implies that the principal has effectively designed the contract as a Trust contract.[9] Second, the contract specifies a positive fine or bonus ($f > 0$, $b > 0$), but the desired effort is set at 1 – that is, the principal does not ask for the maximal effort that is implementable under selfish rationality (suboptimal desired effort). Overall, this happened in only 19 out of 699 – that is, 2.7% of cases. Or third, a fine (bonus) has been specified ($f > 0$, $b > 0$), but the desired effort is set too large, so the contract is a NIC contract. The latter reason is the most frequent cause of best-reply effort equal to 1 (it comprises 91.7%, 97.2%, and 89.3% of all $e^* = 1$ contracts in panels *a* to *c*, respectively), which leads to our separate analysis of NIC contracts below.

In Subsections 5.3.2 and 5.3.3, we analyze the three clusters identified here in more detail and provide statistical tests.

5.3.2 Effort under incentive-compatible contracts

Figure 3 is based on accepted Fine and Bonus contracts. With Stranger matching, these are 1,372 observations, of which 935 contracts (68.1%) are incentive compatible and 437 (31.9%) are not incentive compatible (as defined in Section 2.2, equation (1)). With Partner matching, there are 296 accepted contracts (53 incentive compatible (17.9%), 243 NIC (82.1%)).[10] The relative frequencies of incentive-compatible contracts of 68.1% in Stranger versus 17.9% in Partner indicate that short-run monetary incentives are more important under Strangers matching than Partner matching.

Most incentive-compatible contracts exhibit a fine or bonus of 80 (869/935 = 92.9% in Stranger, 45/53 = 84.9% in Partner). However, many of these contracts do not specify a desired effort of 12 but a smaller one. A desired effort of 12 has a relative frequency of only 27.6% (258/935) in Stranger and 35.9% (19/53) in Partner. That is, more than half of the contracts exhibit a suboptimal desired effort if one takes the perspective of money-maximizing rationality, according to which 12 is the maximally implementable effort with a fine or bonus of 80.

Table B5 in the Online Appendix reports, separately for Fine and Bonus conditions, the detailed distributions of best-reply effort, minimal effort and other effort choices.[11] Across treatments, between 69.2% and 89.7% of effort choices are best-reply choices. Between 10.3% and 23.8% are minimal (effort = 1) choices, and any other effort has a negligible frequency between 0% and 8.1%. The latter is clear support for Hypothesis 5c (no voluntary cooperation above best-reply effort if the contract is incentive compatible). Regarding the framing of monetary incentives (Fine vs. Bonus contracts, which are incentive equivalent), there is no systematic pattern. The relative frequency of

[9] Fehr and Rockenbach (2003) and Fehr and List (2004) provide evidence in single-shot trust game experiments that principals who deliberately design trust contracts by setting incentives to zero ($f = 0$, $b = 0$) might trigger more reciprocal reward than incentive contracts. In our dataset, as the results quoted here show, this is not a frequent motivation. In terms of effort, for contracts with fine equal to 0 and bonus equal to 0, the average effort in FT/BT was 1.75; and in TFT/TBT average effort was 1. By contrast, in the repeated games of TFT-R/TBT-R average effort for contracts with fine equal to 0 and bonus equal to 0 was 18.4, whereas average effort in TTT-R was 14.6. Thus, deliberately selecting no incentives in an incentive contracting environment was unsuccessful in our Stranger one-shot environments of FT/BT and TFT/TBT but very successful in our repeated game environments of TFT-R/TBT-R compared to TTT-R.

[10] With Partner matching incentive compatibility is a more complex issue than with Strangers matching. Specifically, with Partner matching incentive compatibility does not need to hold in each period. Nevertheless, for statistical comparison it is instructive to apply the same criterion as with Strangers matching. We think furthermore that granting incentive compatibility in each period of a repeated game is a reasonable and natural way for experimental participants to approach the problem.

[11] In rare cases best-reply predicted a choice of 1. Thus, an effort choice of 1 represents a best-reply choice and a minimal effort choice at the same time. We counted these observations as best-reply since best-reply is by far more frequent than minimal effort.

https://doi.org/10.1017/eec.2024.14 Published online by Cambridge University Press

Table 5 Effort choice under incentive-compatible contracts

	FT/**B**T	T**F**T/T**B**T	T**F**T-R/T**B**T-R
Table 5.1: Probit; dependent variable: $effort = 1$			
Model	(1a)	(1b)	(1c)
OC	−.232 (.146)	−.121 (.136)	−.053 (.345)
Treatment	.122 (.128)	−.500*** (.161)	.452 (.405)
Constant	−.39* (.224)	−.525*** (.201)	−1.196** (.488)
Observations	446	489	53
Pseudo R-squared	.021	.0312	.0254
Table 5.2: OLS; dependent variable: $effort > 1$			
Model	(2a)	(2b)	(2c)
Best-reply effort	.953*** (.027)	.995*** (.005)	.992*** (.007)
OC	.217 (.237)	−.013 (.042)	.056 (.049)
Treatment	.185 (.191)	−.006 (.05)	.042 (.038)
Constant	.044 (.211)	.019 (.071)	.045 (.06)
Observations	360	401	45
R-squared	.754	.967	0.998

Bolded letters indicate the phase under consideration. Dataset: accepted and incentive-compatible Fine or Bonus contracts. All estimations include dummies for periods 1–3 and periods 8–10 to control for (noisy) initial and end behavior; the omitted benchmark category is the central periods 4–7. The full estimation results are in Online Appendix B, Table B4. In Table 5.1, the dependent variable is coded as 1 if minimal effort is chosen and coded as zero if effort > 1 is chosen. In Table 5.2, the dependent variable is effort > 1. Offered Compensation (OC) is measured in units of 100 and is *wage* under Fine contracts and *wage* + *bonus* under Bonus contracts. Treatment is a dummy for FT in models 1a and 2a, TFT in models 1b and 2b, and TFT-R in models 1c and 2c. The best-reply effort is calculated according to equation (1) in the main text (Section 2.2). Results are robust for controlling for initial and end-period effects, which are all insignificant at $p < .05$. * $p < .10$; ** $p < .05$; *** $p < .01$.

best-reply choices is smaller in FT (69.2%) than in BT (76.9%) but larger in TFT (86.1%) than in TBT (74%).

To investigate these issues in more detail, Table 5 reports Probit and OLS regressions for different data subsets. In Table 5.1, we report Probit regressions that estimate the probability of minimal effort, and in Table 5.2 we document OLS regressions that estimate effort conditional on above-minimal effort choice. This partitioning of the data analyses follows from our identification of data clusters in Fig. 3 and is analogous to our analyses of Trust contracts. As explanatory variables, we use a treatment dummy (a dummy for either FT, TFT, or TFT-R, depending on the dataset; this dummy identifies the difference between Fine and Bonus contracts), best-reply effort (only in Table 5.2), and Offered Compensation (OC). For an incentive-compatible Fine contract, the OC is equal to a fixed wage, since the fine is not paid if the agent chooses the best-reply effort (which maximizes the agent's payoff). For an incentive-compatible Bonus contract, OC is calculated as fixed wage plus bonus, since best-reply effort means that the bonus is received. If the contract is not incentive compatible, a money-maximizing agent should choose minimal effort under both Fine and Bonus contracts. Hence, NIC Fine or Bonus contracts render them incentive equivalent to Trust contracts (see also Section 2.2 for a detailed discussion).

https://doi.org/10.1017/eec.2024.14 Published online by Cambridge University Press

Table 5.1 shows that all estimated coefficients of OC have the expected negative sign (higher OC reduces the probability of minimal effort). Under Stranger matching, the effect is marginally significant in the two-phase FT/BT experiments (two-tailed $p = .093$), insignificant in TFT/TBT (two-tailed $p = .373$), and also insignificant under the Partner matching of TFT-R/TBT-R ($p = .879$). The latter is intuitive: In a repeated game, it may be better not to immediately punish a low OC by choosing minimal effort.

The treatment dummies (dummies for FT, TFT, and TFT-R, respectively), which measure whether the framing of incentives (Fine vs. Bonus) matters, are insignificant in FT/BT and TFT-R/TBT-R (two-sided, both $p > .263$) but significantly negative in TFT/TBT (two-side $p = .002$), indicating that minimal effort choices under incentive-compatible contracts are less likely under Fine contracts than Bonus contracts. In sum, the influence of framing of incentives is ambiguous.

Regarding non-minimal effort choices ($e > 1$), which we analyze in Table 5.2, OC is insignificant in all three models (two-sided, all $p > .243$). This means that there is no positive reciprocity above e^*, and this is not due to a ceiling effect. The variable Treatment is insignificant as well – that is, there are no framing effects (all $p > .256$). The only, and highly significant ($p < .0001$), regressor is the best-reply effort. The estimated coefficients are very close to 1. Judging by the displayed R^2-values, best-reply effort explains very large fractions of variance (all $R^2 > .753$), and this holds with Stranger and Partner matching. These regressions reflect the strong clustering at best-reply effort visible in Fig. 3.

From both analyses, Probit and OLS regressions, we conclude that, given an accepted and incentive-compatible Fine or Bonus contract, the agent either chooses best-reply effort or minimal effort, and there is very little noise. Incentives have a very strong influence on effort (confirming Hypothesis 3), and if agents choose an above-minimal effort level, they choose the best-reply effort almost perfectly (supporting Hypothesis 5b). This also means that it is clearly disadvantageous for the principal to specify a desired effort below the maximally implementable level. For instance, if the specified fine is 80 and the desired effort is 11 instead of 12, the effort provided (conditional on $e > 1$) will be about 11 instead of 12 – regardless of the size of the OC. Thus, by offering higher compensation, the principal cannot increase effort beyond the best-reply level (Hypothesis 5c). The influence of framing is ambiguous, which contradicts Hypothesis 4. There is some evidence of minimal effort as reciprocal punishment (Hypothesis 2b), but it is weaker than under Trust contracts.

5.3.3 *Effort under NIC contracts*

We now analyze the third cluster shown in Fig. 3 – that is, contracts that are accepted but are NIC. This cluster comprises 437 out of 1,372 observations (31.9%) in the Strangers treatments, of which 61.6% (269/437) exhibit maximal fine (bonus). In Partner treatments 244 out of 296 observations (82.1%) are NIC contracts, of which 56.2% (137/244) contain the maximum fine or bonus.

Recall from Section 2 that, in our setting, NIC incentive contracts are incentive-equivalent to Trust contracts. Thus, here we investigate to what extent NIC contracts impact on reciprocal gift exchange that we observe under all Trust contracts (see Fig. 2). Figure 2, panels *a, g* and *m*, illustrate the relationship of OC and effort that is behind the third cluster – that is the distribution of effort choices at best-reply effort equal to in Fig. 3. Under NIC contracts, OC amounts to *wage less fine* under Fine contracts and to *wage* under Bonus contracts because the agent has the incentive to choose minimal effort which triggers the specified fine or forfeits the bonus.

Comparing Fig. 2, panel *a*, which depicts NIC Fine/Bonus contracts, with panels *g* and *m* (representing NIC contracts in TFT/TBT and TFT-R/TBT-R, respectively) suggests two patterns. First, without experience of Trust contracting (as is the case in FT/BT), being exposed to incentive contracts, albeit NIC ones, leads agents to choose effort levels that are all over the place and are unrelated

https://doi.org/10.1017/eec.2024.14 Published online by Cambridge University Press

Table 6 Effort choice under NIC contracts

	FT/**B**T	T**F**T/T**B**T	T**F**T-R/T**B**T-R
Table 6.1: Probit; dependent variable: *effort* $= 1$			
Model	(1a)	(1b)	(1c)
OC	−.045 (.087)	−.341*** (.067)	−.792*** (.11)
Treatment	.189 (.257)	.109 (.42)	−.094 (.287)
Constant	.615*** (0.22)	.53 (.432)	.827** (.347)
Observations	229	208	243
Pseudo R-squared	.0435	.0697	.295
Table 6.2: Tobit; dependent variable: *effort* > 1			
Model	(2a)	(2b)	(2c)
OC	.442 (.474)	3.855*** (.932)	6.24*** (.868)
Treatment	−.142 (.851)	2.278 (2.033)	−.223 (1.548)
Constant	8.807*** (1.232)	4.946** (2.38)	−.214 (2.469)
Observations	77	108	214
Pseudo R-squared	.00681	.0512	.15

Bolded letters indicate the phase under consideration. The dataset is accepted and NIC Fine and Bonus contracts. All estimations include dummies for periods 1–3 and periods 8–10 to control for (noisy) initial and end behavior; the omitted benchmark category is the central periods 4–7. The full estimation results are in Online Appendix B, Table B6. The dependent variable in Table 6.1 is a dummy variable (1 if effort $= 1$, 0 otherwise) and in Table 6.2, effort > 1. Offered compensation (OC) is measured in units of 100. Treatment is a dummy for FT (in models 1a and 2a), for TFT (in models 1b and 2b), and for TFT-R (in models 1c and 2c). Standard errors (in parentheses) are adjusted for clustering on matching groups. p $<$ .05. * p $<$ .10; ** p $<$ .05; *** p $<$ .01

to OC. Second, the experience of Trust contracting in phase 1 of TFT/TBT and TFT-R/TBT-R returns a positive relationship between effort and OC. In the following, we investigate these patterns econometrically.

In Table 6, we estimate the probability of minimal effort by a Probit regression (Table 6.1) and effort conditional on effort greater than 1 by a Tobit regression (Table 6.2). Unlike our analysis of incentive-compatible contracts, we apply Tobit with an upper bound of 20 rather than OLS because Fig. 3 (and Fig. 2) reveal a high frequency of boundary choices at effort equal to 20 (in particular in TFT-R/TBT-R). As explanatory variables, both regressions use OC and a treatment dummy that measures the difference between Fine and Bonus contracts (dummy for FT in models 1a and 2a; TFT in models 1b and 2b; and TFT-R in models 1c and 2c). Like before, we include dummies for initial periods 1–3 and end periods 8–10. The full set of results is in Online Appendix B, Table B6.

As can be seen in Table 6.1, in TFT/TBT (model 1b) and in TFT-R/TBT-R (model 1c), a higher OC significantly decreases the probability of minimal effort, which suggests that minimal effort choices are an expression of negative reciprocity. Consistent with the gift-exchange hypothesis (positive reciprocity), OC significantly increases effort conditional on effort greater than 1, as shown in Table 6.2, models 2b and 2c.

In stark contrast, in FT/BT (models 1a and 2a), the coefficients are insignificant – that is, effort choice is unrelated to OC (see also Fig. 2, panel *a*, and compare to panels *g* and *m*). Thus, when participants had experienced Trust contracting (and hence positive and negative reciprocity) in phase 1

https://doi.org/10.1017/eec.2024.14 Published online by Cambridge University Press

before being exposed to NIC contracts in phase 2, agents behaved reciprocally like under Trust contracting: They were less likely to choose minimal effort the higher the offered wage was and to choose higher effort the higher the wage was. Without such experience (treatments FT/BT), the presence of incentive contracts seems to have removed reciprocity. Or could the lack of reciprocity also be due to chance?

Here we briefly describe two placebo tests to see whether the lack of a wage–effort relationship illustrated in Fig. 2a and estimated in Table 6, models 1a and 2a, are due to chance, rather than a 'crowding-out-of-reciprocity effect.' See Section B5 in the Online Appendix for further details and an illustration (Fig. B4).

For the placebo tests, we use bootstrap to draw 500 random samples of 77 contracts (the number of accepted NIC contracts with effort > 1, see Table 6) from the data of Trust contracts in phase 1 of the TTT and TFT and TBT experiments, where incentive contracts cannot have influenced the wage–effort relationship. We ran 500 Probit and 500 Tobit regressions (following models 1a and 2a in Table 6 with wage instead of OC because incentive contracts are not available in the placebo data of phase 1 TTT/TFT/TBT).

The bootstrap Probit regressions are on a dummy of minimal effort. The mean of the estimated coefficients of wage is $-.483$, and the 99% confidence interval is $[-.501, -.465]$, which does not include the estimated coefficient of $-.045$ in Table 6.1. 91.8% of all p-values are less than .05. The bootstrap Tobit regressions are on effort greater than 1. The mean of the estimated coefficients of wage is 3.15, and the 99% confidence interval is [3.09, 3.21] – far away from the estimated coefficient of .442 in Table 6.2; 98.4% of all p-values less than .05. We conclude that the absence of a reciprocal wage–effort relationship in phase 1 of FT/BT is not due to chance but reflects a true 'crowding-out-of-reciprocity effect.'

Is it possible that agents, when thinking about their effort choice, pay attention to the elements of the offered contract despite the contract not being incentive compatible? From a theoretical point of view, the desired effort level and the stipulated fine or bonus should not matter, but agents may nevertheless be influenced by them. To investigate this possibility, we repeated our analysis of Table 6 by also including the stipulated desired effort and fine or bonus. We report those results in the Online Appendix Table B7. We find that in FT/BT, the elements of the NIC contract matter to some extent: The higher the desired effort, the higher the likelihood of minimal effort ($p = .002$), and the higher the wage, the lower the likelihood of minimal effort ($p = .025$); fine or bonus is insignificant ($p = .163$). For effort greater than 1, we find that wage is now positive but only weakly significant ($p = .05$); effort also increases significantly with a fine or bonus ($p = .000$) but not with desired effort ($p = .498$). These observations may explain the noisy effort choices in phase 1 of FT/BT (see Fig. 2a). Interestingly, under NIC contracts in TFT/TBT and TFT-R/TBT-R desired effort and fine/bonus have no significant impact (all $p > .114$, except for desired effort on minimal effort in TFT/TBT where $p = .077$).

Finally, we briefly investigate the *role of framing* of incentive contracts as either Fine or Bonus. Visual inspections of Figs. B2 and B3 in the Online Appendix, which provide a breakdown of Figs. 2 and 3 by contract type, suggest very limited framing effects. The econometric analyses of Table 6 support this impression. The dummy variable Treatment, which measures the difference between the respective Fine contracts and Bonus contracts (FT vs. BT, TFT vs. TBT, TFT-R vs. TBT-R), has an insignificant effect throughout. We conclude that framing incentives as fine or bonus does not matter in our dataset, which is evidence against our Hypothesis 4.

5.4. Comparison of predicted effort across treatments

We now collect the results of the previous sections and compare them along predicted values, most importantly expected effort $E(e)$, derived from the regression analyses. Table 7, panel A, shows the predicted values for Trust contracts, panel B for incentive-compatible Fine and Bonus contracts, and panel C for NIC Fine and Bonus contracts. To determine $E(e)$, we first calculate the predicted

https://doi.org/10.1017/eec.2024.14 Published online by Cambridge University Press

Table 7 Predicted values based on regression models

	Scenario 1					*Scenario 2*				
	OC	$Pr(e = 1)$	$e \vert e > 1$	$E(e)$	$E(\pi_P)$	*OC*	$Pr(e = 1)$	$e \vert e > 1$	$E(e)$	$E(\pi_P)$
Panel A: *Trust contracts (and after Fine or Bonus contracts)*										
F**T**/B**T**	64.6	.861	3.1	1.29	−19.4	200	.636	7.85	3.49	−77.7
T**T**T	131.6	.616	5.45	2.71	−36.8	200	.47	7.85	4.63	−37.9
TF**T**/TB**T**	92.8	.764	4.37	1.8	−30	200	.544	8.38	4.36	−47.5
TT**T**	142.5	.574	5.14	2.76	−45.8	200	.444	7.29	4.5	−42.6
F**T**/B**T**	64.6	.836	3.21	1.36	−16.9	200	.577	8.09	4	−60
TF**T**/TB**T**	92.8	.792	4.23	1.67	−34.3	200	.577	8.09	4	−60
TF(B)**T**-R	276.5	.1	13.1	11.89	139.7	200	.213	9.9	8	80.2
TT**T**-R	390.8	.022	17.8	17.43	219.3	200	.213	9.9	8	80.2
Panel B: *Incentive-compatible Fine or Bonus contracts*										
FT	144.5	.193	9.28	7.68	124.4	200	.193	9.28	7.68	68.9
BT	124.2	.193	9.28	7.68	144.7	200	.193	9.28	7.68	68.9
T**F**T	120.8	.18	9.59	8.04	160.7	200	.19	9.59	7.96	78.5
T**B**T	120.8	.18	9.59	8.04	160.7	200	.19	9.59	7.96	78.5
T**F**T-R	127.1	.151	9.38	8.11	156.9	200	.145	9.38	8.16	85.8
T**B**T-R	127.1	.151	9.38	8.11	156.9	200	.145	9.38	8.16	85.8
Best-reply effort $e^* = 9.16$ in FT/BT; $e^* = 9.61$ in TFT/TBT; $e^* = 9.36$ in TFT-R/TBT-R.										
Panel C: *NIC Fine or Bonus contracts*										
FT	162.8	.664	4.21	2.08	−90.1	200	.664	4.21	2.08	−127.3
BT	173	.664	4.21	2.08	−100.3	200	.664	4.21	2.08	−127.3
T**F**T	178.8	.53	6.92	3.78	−46.4	200	.5	7.58	4.29	−49.9
T**B**T	241	.443	8.87	5.38	−52.6	200	.5	7.58	4.29	−49.9
T**F**T-R	290.8	.07	17.91	16.73	294.6	200	.222	11.33	9.04	116.3
T**B**T-R	290.8	.07	17.91	16.73	294.6	200	.222	11.33	9.04	116.3

Notes: Predicted values based on regressions of minimal effort and effort conditional on above-minimal effort. Data: All accepted Trust contracts, subsets as defined in panel headers and the first column; bolded letters indicate respective phase under consideration. Compared to the Probit and OLS-regressions reported in the previous section, the models were re-estimated after eliminating explanatory variables that were insignificant for a one-tailed test. Scenario 1 determines predictions for means of offered compensation (OC) of the respective data subset (shown in column 2). Scenario 2 determines predictions for an offered compensation of 200. $Pr(e = 1)$ is the estimated probability for minimal effort; $e|e>1$ is the estimated effort conditional on aboveminimal effort; E(e) is the expected effort combining the partial effects, i.e., $E(e) = Pr(e = 1) \cdot 1 + (1 - Pr(e = 1)) \cdot (e|e>1)$ and $E(\pi P) = Pr(e = 1) \cdot 1 \cdot 35 + (1 - Pr(e = 1)) \cdot (e|e>1) \cdot 35 - OC$ is expected profit of the principal. For panel B also the mean values of best-reply effort e* are provided for which the predicted value calculations are done (see note to Panel B).

probability of minimal effort ($Pr\,(e = 1)$) based on Probit regressions and then the predicted effort conditional on above-minimal effort ($e\,|e\rangle\,1$) based on OLS regressions and assuming values of OC as shown in column OC. $E\,(e)$ is then calculated as $E\,(e) = Pr\,(e = 1) \cdot 1 + (1 - Pr\,(e = 1) \cdot (e\,|e\rangle\,1)$. We also report the expected profit of the principal calculated as $E\,(\pi_P) = Pr\,(e = 1) \cdot 1 \cdot 35 + (1 - Pr\,(e = 1) \cdot (e\,|e\rangle\,1) \cdot 35 - OC$.

We study two scenarios. In *Scenario 1*, we use the mean values of *OC*, calculated separately for each treatment, which capture changes in the principal's trust across two treatments under comparison. Thus, in Scenario 1, $E\,(e)$ combines three partial effects of experiencing monetary incentives before Trust contracting: a change in trust, a change in the probability of minimal effort, and a change in effort conditional on above-minimal effort.

In *Scenario 2*, we calculate the same variables as in Scenario 1 but assume a fixed OC of 200, which implies that in our treatment comparisons, $E\,(e)$ displays changes in expected effort that are not associated with changes in trust by the principal. In both scenarios we only use significant effects: If a regression analysis returned an insignificant (at $p > .1$) influence of treatment or OC or both, we recalculated the regression after eliminating all insignificant explanatory variables. For this reason, some predictions and OC values reported in Table 7 do not differ between treatments.

Table 7A records $E\,(e)$ under Trust contracts after Fine or Bonus contracts. All differences in $E\,(e)$ in treatment comparisons reported for Scenario 1 are as predicted by Hypotheses 1a to 1d, which suggest reduced voluntary cooperation (effort) in Trust contracts after having experienced incentive contracts (see Section 4). For instance, for Trust contracts in phase 2 after Fine/Bonus contracts in phase 1 of the FT/BT experiments, the expected effort $E\,(e)$ is 1.29, which results from the fact that principals on average offered a wage of 64.6 and, in response, agents chose the minimum effort ($e = 1$) in 86.1% of the cases and on average put in an effort of 3.1 for their non-minimal effort choices. These effort choices yield an expected payoff for the principal of $E(\pi_P) = -19.4$ money units. In contrast, in phase 2 of TTT – that is, after a Trust contract in phase 1, $E\,(e)$ is 2.71. Expected effort is lower after incentive contracts than after Trust contracts in all comparisons recorded in Scenario 1 of panel A.

In Scenario 2, which fixes OC at 200, $E\,(e)$ equals 3.49 in phase 2 of FT/BT and $E\,(e)$ equals 4.63 in phase 2 of TTT. Here, this difference in $E\,(e)$ is not due to differences in OC (and the reciprocal reaction to it in $e\,|e\rangle\,1$) but due to $Pr\,(e = 1)$, which is 63.6% in phase 2 of FT/BT and 47% in phase 2 of TTT. Scenario 2 also shows that effort conditional on effort greater than varies very little between treatments if it varies at all. This shows that the overall differences in $E\,(e)$ displayed in Scenario 1 are mainly driven by the two partial effects of a reduction in trust by the principal and an increasing probability of minimal effort, but not by changes in effort conditional on effort greater than 1.

The main conclusion from Table 7A is that the data supports Hypothesis 1: Trust contracting after a phase of incentives leads to lower effort than experiencing Trust contracts throughout. This negative influence of experiencing incentives is stronger with Partner matching than with Stranger matching. Under Partner matching, expected effort $E\,(e)$ is reduced by 5.54 units (17.43 – 11.89). This crowding out of voluntary cooperation comes through two channels: A reduction in the principal's trust level and an increased willingness to provide only minimal effort. The wage–effort relationship conditional on effort greater than 1 remains intact.

Column $E\,(\pi_P)$ reports that an average Trust contract induces a positive expected profit for the principal only under Partner matching. Figure 4a illustrates that our experimental game sets a hard task for the principal to achieve profitable cooperation under Stranger matching. The variance in the principal's profit increases substantially in the offered wage. Offering a high wage is very risky, frequently leading to negative profit due to choices of minimal effort. This is different under Partner matching (Fig. 4b), according to which negative profits are much less likely.

Table 7B shows similar calculations for incentive-compatible Fine and Bonus contracts. Three effects are striking: First, under Stranger matching, $E\,(e)$ is higher than for Trust contracts shown in panel A – that is, monetary incentives are highly effective, although they fall short of the theoretically predicted level of 12. Second, there is no difference in $E\,(e)$ between Stranger and Partner matching. Thus, even with Partner matching, an incentive-compatible contract induces agents to just focus on incentives and nothing else. Third, $E\,(e)$ and $E\,(\pi_P)$ are substantially lower compared to Trust contracts with Partner matching. Thus, when incentive-compatible contracts are used in a repeated relationship, explicit incentives dominate, and implicit incentives – that is, the sequential reciprocity mechanism across periods – do not work as effectively as under Trust contracting.

Table 7C displays predicted values for NIC Fine and Bonus contracts. Looking at Scenario 2, it is apparent that, under Stranger conditions, NIC contracts perform worse than incentive-compatible contracts (Hypothesis 5d). Under Partner conditions, however, NIC contracts induce similarly high

https://doi.org/10.1017/eec.2024.14 Published online by Cambridge University Press

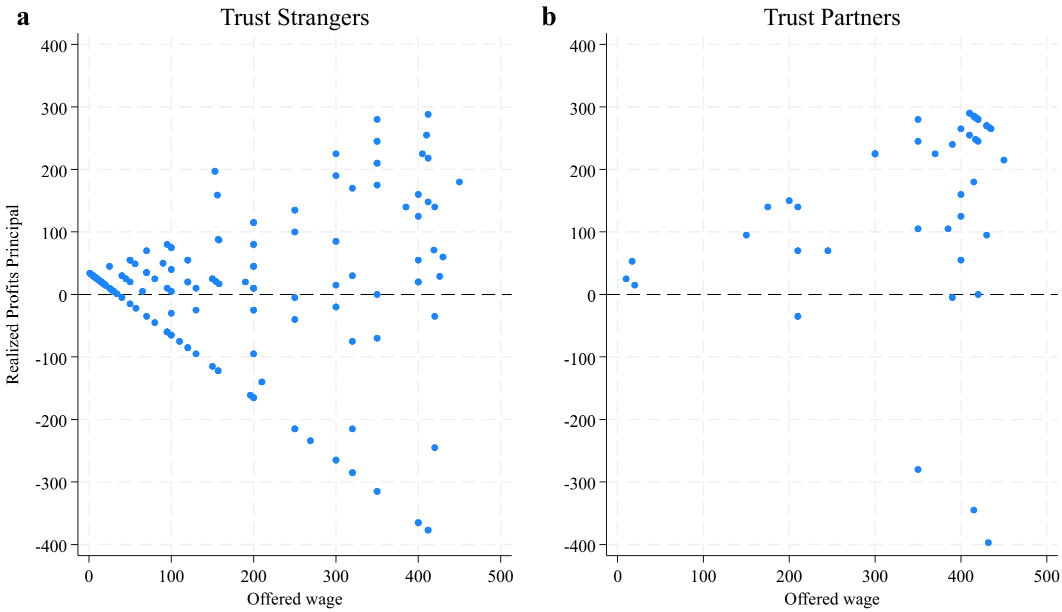

Fig. 4 Principal's profits under phase 3 Trust contracts in TTT strangers and TTT-R partners

effort levels as Trust contracts. Many NIC contracts in TFT-R/TBT-R stipulate a desired effort of 20 (132 out of 243 cases; 54.3%), similar to phase 2 TTT-R contracts (64 out of 109 cases, 58.7%) and slightly lower than in phase 3 TTT-R contracts (78 out of 114; 68.4%). Furthermore, in case the desired effort is 20, most participants provide an effort of 20 (95/132 = 72% with phase 2 NIC contracts in TFT-R/TBT-R; 39/64 = 60.9% with phase 2 Trust contracts in TTT-R; and 59/78 = 75.6% under phase 3 TTT-R contracts). Together, the principal's choice of high desired effort and the agent's choice to provide this high effort are responsible for the effectiveness of NIC contracts in Partner matching. Comparing columns $E(\pi_P)$ in panel C with panel B confirms that under Partner matching, the expected profit of the principal is higher with an NIC contract than with an incentive-compatible contract.

6. Summary and concluding remarks

Our paper's main research goal is a comprehensive investigation of how explicit incentives interact with voluntary cooperation in one-shot and repeated gift-exchange environments. Understanding this interaction is important because contractual relationships in real life often have explicit incentives but also rely on trust and voluntary cooperation because contracts are typically incomplete. We focused on two main questions: How do incentive contracts affect voluntary both when contracts are incentive compatible and when they are not? Does experience with incentive contracts spill over to behavior under Trust contracts even after explicit incentives have been abolished? The main behavioral reason for why such interaction effects exist is that explicit incentives focus an agent's attention on their self-interest, which may undermine ('crowd out') voluntary cooperation, whereas voluntary cooperation rests on social preferences. Naturally occurring contractual relationships often have explicit incentives, and people may also have experience with pure trust contracts without explicitly specified incentives. Our experiments aimed at cleanly separating contemporaneous incentive effects from experience effects.

https://doi.org/10.1017/eec.2024.14 Published online by Cambridge University Press

Starting with the question whether experience with incentive contracts affects voluntary cooperation in their absence, one major result is that voluntary cooperation is reduced under Stranger matching, consistent with Hypotheses 1a and 1b, which predicted this crowding-out effect. This finding supports the conjecture that prior experience of Fine or Bonus contracts undermines the development of cooperation in one-shot interactions that come through the trust–reciprocity mechanism. The effect is stronger in phase 2 of FT/BT than in phase 3 of TFT/TBT, which suggests that experiencing Trust contracts in phase 1 before experiencing Fine or Bonus contracts in phase 2 diminishes the crowding-out effect in phase 3 (Hypothesis 1c).

Another novel result, and a twist on Hypothesis 1, is the persistent negative effect of explicit incentives even under Partner matching (Hypothesis 1d). Experiencing Fine or Bonus contracts in phase 2 of TFT-R/TBT-R weakens cooperation under Trust contracting in phase 3 even more than in Stranger one-shot interactions. Thus, implicit incentives provided by sequential reciprocity across rounds that are inherent in repeated game interactions are substantially compromised by previous experience of incentive contracting.

A further important new result is our detailed identification of two underlying channels through which these crowding-out effects occur: A reduction in the trust level exhibited by the principal in the form of their OC and an increase in the willingness of the agent to provide minimal effort. These two effects are responsible for reducing mean observed effort. A third potential channel – a change in the wage–effort relationship – is unimportant in our data (see Fig. 2). Conditional on an above-minimal effort choice, the reciprocal wage–effort relationship remains intact. We deem this an important result because it suggests that experience with incentive contracting might not destroy the possibility of voluntary cooperation: If principals pay well enough (and hence exhibit enough trust), reciprocity still works to produce high effort. This effect is particularly pronounced in Partner relationships.

Our framework explains our data as a function of three fundamental behavioral mechanisms: negative and positive reciprocity and self-interest. Agents' effort choices reflect reciprocal behavior in its negative and positive forms (Hypothesis 2): Agents are more likely to reject contracts if the principal offers low compensation (supporting Hypothesis 2a) or will choose minimal effort (supporting Hypothesis 2b). This negative reciprocity is consistent with many results from ultimatum bargaining, which showed that many people reject unfair offers (e.g., Güth & Kocher, 2014; Lin et al., 2020). On the positive side, as expected from many gift-exchange games (e.g., Cooper & Kagel, 2016; Fehr & Gächter, 2000), agents display a positive wage–effort correlation conditional on an above-minimal effort choice (confirming Hypothesis 2c). The only exception is a lack of experience with trust and reciprocity before being exposed to incentive contracting. In this case, and consistent with Fehr and Gächter (2002), reciprocity (both negative and positive) does not work, and agents choose either minimal effort or seemingly random positive effort.

Consistent with self-interest motivation, we also find that explicit incentive contracts are effective in inducing high effort (e.g., Gächter et al., 2016; supporting Hypothesis 3). More importantly, and in line with quantitative theoretical predictions, if the contract is incentive compatible, agents choose exact best-reply efforts in many cases (consistent with Anderhub et al., 2002, and confirming Hypothesis 5b). However, we also find that there is no voluntary effort beyond incentive-compatible best-reply levels, although, under Trust contracts, agents are willing to provide those levels. This also means that there is less voluntary cooperation than with Trust contracts (as predicted by Hypothesis 5c). This holds under Stranger matching and, maybe more surprising, under Partner matching. With Partner matching, asking for a desired effort that is higher than the incentive-compatible level is beneficial because it can induce effort levels above 12 if the wage is high enough. This results in higher effort and a higher expected profit of the principal than an incentive-compatible contract. On the contrary, with Stranger matching, NIC contracts perform worse than incentive-compatible contracts (Hypothesis 5d).

Contracts that do not satisfy the participation constraint are almost always rejected (confirming Hypothesis 5a). In addition, contracts are likely rejected, and there is a higher probability of minimal

https://doi.org/10.1017/eec.2024.14 Published online by Cambridge University Press

effort if OC is low, replicating evidence in Anderhub et al. (2002). Unlike Fehr and Gächter (2002) and Fehr et al. (2007), but consistent with de Quidt et al. (2017), we do not find a framing effect. Fine and Bonus contracts are equally effective. Under Trust contracting as well as incentive contracting Partner matching induces higher effort than Stranger matching (in line with Hypothesis 6, and replicating evidence by Falk et al., 1999; Gächter & Falk, 2002).

In summary, explicit incentives lead to failures of separability and tend to crowd out voluntary cooperation. Incentives can also create history effects that can have spillover effects that are detrimental to voluntary cooperation. The details of these effects depend on the features of the situation in which explicit incentives are embedded; in our context, these are whether agents have prior experience with trust and reciprocity-based incentives and whether the interaction is repeated or not.

Supplementary material. The supplementary material for this article can be found at https:/doi.org/10.1017/eec.2024.14.

Data availability statement. Data and analysis code (in Stata) are available at https://doi.org/10.17605/OSF.IO/ACH8X.

Acknowledgements. We benefited from helpful comments by David Cooper, two referees, Nick Bardsley, Sam Bowles, Maria Garcia-Vega, Uri Gneezy and his students, Bernd Irlenbusch, Martin Sefton, Maroš Servátka, Eva Poen, Christian Thöni, Till O. Weber, and participants in numerous workshops and seminars. S. G. is grateful for the hospitality of the Institute for Advanced Studies at Hebrew University Jerusalem, the NZEEL at the University of Christchurch, and briq Bonn while working on this paper.

Funding statement. This work was supported by the European Research Council [Grant Numbers ERC-AdG 295707 COOPERATION and ERC-AdG 101020453 PRINCIPLES] and the Economic and Social Research Council [Grant Number ES/K002201/1]. We also acknowledge support from the EU-TMR Research Network ENDEAR (FMRX-CT98-0238) and from the Grundlagenforschungsfonds at the University of St. Gallen.

Competing interests. Not applicable.

References

Akerlof, G. A. (1982). Labor contracts as partial gift exchange. *Quarterly Journal of Economics*, 97(4), 543–569.

Anderhub, V., Gächter, S., & Königstein, M. (2002). Efficient contracting and fair play in a simple principal-agent experiment. *Experimental Economics*, 5(1), 5–27.

Andreoni, J., & Bernheim, D. B. (2009). Social image and the 50–50 norm: A theoretical and experimental analysis of audience effects. *Econometrica*, 77, 1607–1636.

Bandiera, O., Barankay, I., & Rasul, I. (2005). Social preferences and the response to incentives: Evidence from personnel data. *Quarterly Journal of Economics*, 120, 917–962.

Barr, A., & Serneels, P. (2009). Reciprocity in the workplace. *Experimental Economics*, 12(1), 99–112.

Battigalli, P., & Dufwenberg, M. (2007). Guilt in games. *American Economic Review*, 97(2), 170–176.

Bénabou, R., & Tirole, J. (2006). Incentives and prosocial behavior. *American Economic Review*, 96(5), 1652–1678.

Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, reciprocity, and social history. *Games and Economic Behavior*, 10(1), 122–142.

Besley, T., & Ghatak, M. (2018). Prosocial motivation and incentives. *Annual Review of Economics*, 10(1), 411–438.

Bewley, T. (1999). *Why wages don't fall in a recession*. Harvard University Press.

Bohnet, I., Frey, B. S., & Huck, S. (2001). More order with less law: On contract enforcement, trust, and crowding. *American Political Science Review*, 95(1), 131–144.

Bolton, G. E., & Ockenfels, A. (2000). ERC: A theory of equity, reciprocity, and competition. *American Economic Review*, 90(1), 166–193.

Bowles, S. (2003). *Microeconomics: Behavior, institutions, and evolution*. Princeton University Press.

Bowles, S. (2008). Policies designed for self-interested citizens may undermine “the moral sentiments”: Evidence from economic experiments. *Science*, 320(5883), 1605–1609.

Bowles, S. (2014). Niccolò Machiavelli and the origins of mechanism design. *Journal of Economic Issues*, 48(2), 267–278.

Bowles, S. (2016). *The moral economy. Why good incentives are no substitute for good citizens*. Yale University Press.

Bowles, S., & Hwang, S.-H. (2008). Social preferences and public economics: Mechanism design when social preferences depend on incentives. *Journal of Public Economics*, 92(8-9), 1811–1820.

Bowles, S., & Polania-Reyes, S. (2012). Economic incentives and social preferences: Substitutes or complements? *Journal of Economic Literature*, 50(2), 368–425.

Brown, M., Falk, A., & Fehr, E. (2004). Relational contracts and the nature of market interactions. *Econometrica*, 72(3), 747–780.

Burks, S., Carpenter, J., & Goette, L. (2009). Performance pay and worker cooperation: Evidence from an artefactual field experiment. *Journal of Economic Behavior & Organization*, 70(3), 458–469.

Cardenas, J. C., Stranlund, J., & Willis, C. (2000). Local environmental control and institutional crowding-out. *World Development*, 28(10), 1719–1733.

Charness, G. (2004). Attribution and reciprocity in an experimental labor market. *Journal of Labor Economics*, 22(3), 665–688.

Charness, G., Frechette, G. R., & Kagel, J. H. (2004). How robust is laboratory gift exchange? *Experimental Economics*, 7(2), 189–205.

Charness, G., & Kuhn, P. (2011). Lab labor: What can labor economists learn from the lab? In O. Ashenfelter & D. Card (Eds.), *Handbook of labor economics* (229–330). Elsevier.

Chaudhuri, A. (2011). Sustaining cooperation in laboratory public goods experiments: A selective survey of the literature. *Experimental Economics*, 14(1), 47–83.

Cohn, A., Fehr, E., & Goette, L. (2015). Fair wages and effort provision: Combining evidence from a choice experiment and a field experiment. *Management Science*, 61(8), 1777–1794.

Cooper, D. J., & Kagel, J. H. (2016). Other-regarding preferences: A selective survey of experimental results. In J. H. Kagel & A. E. Roth (Eds.), *Handbook of experimental economics, volume 2* (217–289). Princeton University Press.

Cooper, D. J., & Stockman, C. K. (2011). History dependence and the formation of social preferences: An experimental study. *Economic Inquiry*, 49(2), 540–563.

Cox, J. C., Friedman, D., & Sadiraj, V. (2008). Revealed altruism. *Econometrica*, 76(1), 31–69.

Croson, R., & Gächter, S. (2010). The science of experimental economics. *Journal of Economic Behavior & Organization*, 73(1), 122–131.

Dana, J., Weber, R., & Kuang, J. (2007). Exploiting moral wiggle room: Experiments demonstrating an illusory preference for fairness. *Economic Theory*, 33(1), 67–80.

Deci, E. L., Koestner, R., & Ryan, R. M. (1999). A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological Bulletin*, 125(6), 627–668.

de Quidt, J., Fallucchi, F., Kölle, F., Nosenzo, D., & Quercia, S. (2017). Bonus versus penalty: How robust are the effects of contract framing? *Journal of the Economic Science Association*, 3(2), 174–182.

Dickinson, D. L. (1999). An experimental examination of labor supply and work intensities. *Journal of Labor Economics*, 17(4), 638–670.

Dickinson, D., & Villeval, M. C. (2008). Does monitoring decrease work effort? The complementarity between agency and crowding-out theories. *Games and Economic Behavior*, 63(1), 56–76.

Drouvelis, M. (2021). *Social preferences: An introduction to behavioural economics and experimental research*. Agenda Publishing.

Dufwenberg, M., & Kirchsteiger, G. (2004). A theory of sequential reciprocity. *Games and Economic Behavior*, 47(2), 268–298.

Ellingsen, T. (2024). *Institutional and organizational economics: A behavioral game theory introduction*. Polity Press.

Ellingsen, T., & Johannesson, M. (2008). Pride and prejudice: The human side of incentive theory. *American Economic Review*, 98(3), 990–1008.

Embrey, M., Fréchette, G. R., & Yuksel, S. (2018). Cooperation in the finitely repeated prisoner's dilemma. *The Quarterly Journal of Economics*, 133(1), 509–551.

Englmaier, F., & Leider, S. (2020). Managerial payoff and gift-exchange in the field. *Review of Industrial Organization*, 56(2), 259–280.

Falk, A. (2007). Gift exchange in the field. *Econometrica*, 75(5), 1501–1511.

Falk, A., & Fischbacher, U. (2006). A theory of reciprocity. *Games and Economic Behavior*, 54(2), 293–315.

Falk, A., Gächter, S., & Kovacs, J. (1999). Intrinsic motivation and extrinsic incentives in a repeated game with incomplete contracts. *Journal of Economic Psychology*, 20(3), 251–284.

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

Falk, A., & Kosfeld, M. (2006). The hidden costs of control. *American Economic Review*, 96(5), 1611–1630.

Falkinger, J., Fehr, E., Gächter, S., & Winter-Ebmer, R. (2000). A simple mechanism for the efficient provision of public goods: Experimental evidence. *American Economic Review*, 90(1), 247–264.

Fehr, E., & Charness, G. (2024). Social preferences: Fundamental characteristics and economic consequences. *Journal of Economic Literature* forthcoming.

Fehr, E., & Falk, A. (2002). Psychological foundations of incentives. *European Economic Review*, 46(4-5), 687–724.

Fehr, E., & Fischbacher, U. (2003). The nature of human altruism. *Nature*, 425(6960), 785–791.

Fehr, E., & Gächter, S. (1998). Reciprocity and economics: The economic implications of homo reciprocans. *European Economic Review*, 42(3-5), 845–859.

Fehr, E., & Gächter, S. (2000). Fairness and retaliation: The economics of reciprocity. *Journal of Economic Perspectives*, 14(3), 159–181.

Fehr, E., & Gächter, S. (2002). *Do incentive contracts undermine voluntary cooperation?* (IEW Working Paper No. 34). Unversity of Zurich.

https://doi.org/10.1017/eec.2024.14 Published online by Cambridge University Press

Fehr, E., Gächter, S., & Kirchsteiger, G. (1997). Reciprocity as a contract enforcement device: Experimental evidence. *Econometrica*, 65(4), 833–860.

Fehr, E., Goette, L., & Zehnder, C. (2009). A behavioral account of the labor market: The role of fairness concerns. *Annual Review of Economics*, 1(1), 355–384.

Fehr, E., Kirchler, E., Weichbold, A., & Gächter, S. (1998). When social norms overpower competition: Gift exchange in experimental labor markets. *Journal of Labor Economics*, 16(2), 324–351.

Fehr, E., Kirchsteiger, G., & Riedl, A. (1993). Does fairness prevent market clearing? An experimental investigation. *Quarterly Journal of Economics*, 108(2), 437–459.

Fehr, E., Klein, A., & Schmidt, K. M. (2007). Fairness and contract design. *Econometrica*, 75(1), 121–154.

Fehr, E., & List, J. A. (2004). The hidden costs of incentives – Trust and trustworthiness among CEOs. *Journal of the European Economic Association*, 2(5), 743–727.

Fehr, E., & Rockenbach, B. (2003). Detrimental effects of sanctions on human altruism. *Nature*, 422(6928), 137–140.

Fehr, E., & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics*, 114(3), 817–868.

Fehr, E., & Schurtenberger, I. (2018). Normative foundations of human cooperation. *Nature Human Behaviour*, 2(7), 458–468.

Fischbacher, U. (2007). z-Tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics*, 10(2), 171–178.

Frey, B. S. (1997). *Not just for the money: An economic theory of personal motivation*. Edward Elgar Publishing Ltd.

Frey, B. S., & Jegen, R. (2001). Motivation crowding theory. *Journal of Economic Surveys*, 5(5), 589-611.

Gächter, S., & Falk, A. (2002). Reputation and reciprocity: Consequences for the labour relation. *The Scandinavian Journal of Economics*, 104(1), 1–26.

Gächter, S., Huang, L., & Sefton, M. (2016). Combining "real effort" with induced effort costs: The ball-catching task. *Experimental Economics*, 19(4), 687–712.

Gintis, H., Bowles, S., Boyd, R., & Fehr, E. (2005). *Moral sentiments and material interests: The foundations of cooperation in economic life*. MIT Press.

Gneezy, U. (2004). *Does high wage lead to high profits? An experimental study of reciprocity using real effort*. The University of Chicago GSB.

Gneezy, U., & List, J. A. (2006). Putting behavioral economics to work: Testing for gift exchange in labor markets using field experiments. *Econometrica*, 74(5), 1364–1985.

Gneezy, U., & Rustichini, A. (2000). A fine is a price. *The Journal of Legal Studies*, 29(1), 1–17.

Güth, W., & Kocher, M. G. (2014). More than thirty years of ultimatum bargaining experiments: Motives, variations, and a survey of the recent literature. *Journal of Economic Behavior & Organization*, 108, 396–409.

Hannan, R. L., Kagel, J. H., & Moser, D. V. (2002). Partial gift exchange in an experimental labor market: Impact of subject population differences, productivity differences, and effort requests on behavior. *Journal of Labor Economics*, 20(4), 923–951.

Kirchler, M., & Palan, S. (2018). Immaterial and monetary gifts in economic transactions: Evidence from the field. *Experimental Economics*, 21(1), 205–230.

Kranton, R. (2019). The devil is in the details: Implications of Samuel Bowles's *The moral economy* for economics and policy research. *Journal of Economic Literature*, 57(1), 147–160.

Kreps, D., Milgrom, P., Roberts, J., & Wilson, R. (1982). Rational cooperation in the finitely repeated prisoners' dilemma. *Journal of Economic Theory*, 27(2), 245–252.

Kube, S., Maréchal, M. A., & Puppe, C. (2012). The currency of reciprocity: Gift exchange in the workplace. *American Economic Review*, 102(4), 1644–1662.

Kujansuu, E., & Schram, A. (2021). Shocking gift exchange. *Journal of Economic Behavior & Organization*, 188, 783–810.

Lazear, E. P. (2000). Performance pay and productivity. *American Economic Review*, 90(5), 1346–1361.

Lin, P.-H., Brown, A. L., Imai, T., Wang, J. T. Y., Wang, S. W., & Camerer, C. F. (2020). Evidence of general economic principles of bargaining and trade from 2,000 classroom experiments. *Nature Human Behaviour*, 4(9), 917–927.

Rabin, M. (1993). Incorporating fairness into game-theory and economics. *American Economic Review*, 83, 1281–1302.

Rand, D. G., & Peysakhovich, A. (2016). Habits of virtue: Creating norms of cooperation and defection in the laboratory. *Management Science*, 62(3), 631–647.

Reuben, E., & Suetens, S. (2012). Revisiting strategic versus non-strategic cooperation. *Experimental Economics*, 15(1), 24–43.

Sandel, M. (2012). *What money can't buy: The moral limits of markets*. Allen Lane.

Schmelz, K., & Bowles, S. (2021). Overcoming COVID-19 vaccination resistance when alternative policies affect the dynamics of conformism, social norms, and crowding out. *Proceedings of the National Academy of Sciences*, 118(25), e2104912118.

Schmelz, K., & Ziegelmeyer, A. (2020). Reactions to (the absence of) control and workplace arrangements: Experimental evidence from the internet and the laboratory. *Experimental Economics*, 23(4), 933–960.

Selten, R., & Stoecker, R. (1986). End behavior in sequences of finite prisoners-dilemma supergames: A learning-theory approach. *Journal of Economic Behavior & Organization*, 7(1), 47–70.

Shearer, B. S. (2004). Piece rates, fixed wages and incentives: Evidence from a field experiment. *Review of Economic Studies*, 71(2), 513–534.

https://doi.org/10.1017/eec.2024.14 Published online by Cambridge University Press

Simon, H. (1991). Organizations and markets. *Journal of Economic Perspectives*, 5(2), 25–44.

Sliwka, D. (2007). Trust as a signal of a social norm and the hidden costs of incentive schemes. *American Economic Review*, 97(3), 999–1012.

Williamson, O. (1985). *The economic institutions of capitalism*. Free Press.

Ziegelmeyer, A., Schmelz, K., & Ploner, M. (2012). Hidden costs of control: Four repetitions and an extension. *Experimental Economics*, 15(2), 323–340.

Cite this article: Gächter, S., Kaiser, E., & Königstein, M. (2025). Incentives crowd out voluntary cooperation: evidence from gift-exchange experiments. *Experimental Economics*, 28(1), 75–106. https://doi.org/10.1017/eec.2024.14

https://doi.org/10.1017/eec.2024.14 Published online by Cambridge University Press