

SPECIAL ISSUE ARTICLE

Money talks? How different incentive schemes influence prosocial behavior in economic games

Lina Koppel¹ , Amanda M. Lindkvist¹  and Gustav Tinghög^{1,2}

¹Department of Management and Engineering, Division of Economics, Linköping University, Linköping, Sweden

²The National Center for Priority Setting in Health Care, Department of Health, Medicine and Caring Sciences (HMC), Linköping University, Linköping, Sweden

Corresponding author: Lina Koppel; Email: lina.koppel@liu.se

(Received 20 January 2024; revised 9 June 2025; accepted 22 February 2026)

Abstract

The use of real monetary incentives has long defined experimental economics, setting it apart from disciplines like psychology, where hypothetical choices are common. While full-payment and hypothetical designs represent two clear extremes, random-payment incentive schemes – where one of several decisions is randomly selected for payment – have become a prominent approach in economics. Yet the behavioral validity of random-payment schemes remains underexamined: Do they more closely resemble fully incentivized tasks or hypothetical ones? We compare full-payment (participants paid for every decision), no-payment (participants not paid for any decision), and random-payment (participants paid for one randomly selected decision) incentive schemes, using five standard economic games to measure social preferences (Dictator Game, Ultimatum Game, Trust Game, Public Goods Game, and Prisoner's Dilemma). Results from Experiment 1 ($n = 1,501$), with £1 stakes, indicate no significant differences between incentive schemes in any of the games included or a composite measure of prosocial behavior. In Experiment 2 ($n = 750$), with £10 stakes, results were largely in line with Experiment 1, suggesting no consistent behavioral impact of incentive schemes at increased stakes. The one notable exception was responder behavior in the Ultimatum Game, where participants in the full-payment condition were less likely to reject offers compared to those in the no-payment condition, with random-payment falling in between. Our results challenge the rigid disciplinary norm that real stakes are essential for valid measurement and invite a more nuanced consideration of how and when different incentive schemes are necessary or appropriate in behavioral research.

Keywords: Dictator game; prisoner's dilemma; public goods game; trust game; ultimatum game

JEL codes: B40; C70; C90; D90

1. Introduction

Incentives are a core aspect of what defines the 'economics' part in experimental economics. The basis for incentivization lies in the idea that individuals only show their true preferences if the tasks in the experiment have clear financial implications (Harrison, 1994; Smith, 1976). This idea influences multiple stages of the research process, from experimental design to selection of papers to be published, cited, and read. For example, most economic journals do not publish papers that are based on non-incentivized choices (see, e.g., Azrieli et al., 2018; Camerer & Hogarth, 1999). In contrast, non-incentivized, hypothetical choices are common in other social sciences such as psychology (see,

e.g., Erlandsson, 2015, p. 44–46). However, while incentivization is typically seen as essential in economics, the specific implementation of incentive schemes varies widely. In this paper, we report the results of two preregistered experiments designed to compare the behavioral effects of three commonly used incentive schemes – full-payment, random-payment, and no-payment (hypothetical) – across five standard economic games used to assess prosocial behavior. Experiment 1 used low stakes (£1) and Experiment 2 used higher stakes (£10).

The use of incentivized choice tasks is a convention that separates economics from other social sciences, perhaps especially psychology. The difference in methodological convention implies that there is also a difference in the underlying belief about what monetary incentives *do*. As stated by Camerer and Hogarth (1999): ‘Economists presume that experimental subjects do not work for free and work harder, more persistently, and more effectively, if they earn more money for better performance. Psychologists believe that intrinsic motivation is usually high enough to produce steady effort even in the absence of financial rewards; and while more money might induce more effort, the effort does not always improve performance’ (p. 7). Thus, incentivized-choice tasks remain the gold standard in experimental economics, whereas no such standard exists in experimental psychology.

Whether and how monetary incentives influence behavior in experiments is an empirical question, and the answer may vary across research designs and the types of tasks and outcome measures involved (for reviews, see, e.g., Camerer & Hogarth, 1999; Hertwig & Ortmann, 2001; Smith & Walker, 1993). Experiments often involve multiple decisions, meaning that researchers choose not just between whether or not to incentivize, but between different forms of incentive schemes. Researchers may choose to pay all participants for all decisions (a full-payment incentive scheme), all participants for some decisions (a random-payment incentive scheme), or some participants for some or all decisions (sometimes referred to as a between-subject random incentivized system, or BRIS; see, e.g., Thielmann et al., 2020). Theoretically, each of these may have benefits and drawbacks. For example, if the incentive scheme involves some form of randomization, as in the random-payment incentive scheme and BRIS, researchers may be concerned that participants treat a given choice not in terms of its stated payoff but in terms of some fraction of that payoff. On the other hand, if all participants are paid for all decisions, researchers may be concerned about wealth effects and hedging, where participants’ expectations of payment in one decision task influences their decisions in other tasks.

The choice of incentive scheme may be especially important for research on prosocial behavior, because the reputational benefits of acting prosocially may lead to inflated estimates of prosociality if there are no personal costs associated with prosocial behavior. For example, lack of incentives in the Dictator Game or Ultimatum Game makes it cheap to signal generosity or fairness concerns. Still, empirical studies on the effect of incentives in these economic games and other social dilemmas have yielded somewhat mixed results, with a growing number of studies finding that monetary incentives exert limited influence on behavior in standard economic games at typical stake levels. Some researchers have found that monetary incentives (versus no incentives) reduce generosity in the Dictator Game (Amir et al., 2012; Bühren & Kundt, 2015; Clot et al., 2018); others have found no effect (Ben-Ner et al., 2008). In the Ultimatum Game, research has found no effect of incentives (versus no incentives) on average offers from the proposer (the first player) or on average minimum acceptable offers for responders (the second player), at least not when stake sizes are small (Amir et al., 2012; Cameron, 1999), although effects may emerge at higher stakes (Andersen et al., 2011; Cameron, 1999). Finally, research has found no effect of incentives (versus no incentives) on decisions in the Trust Game (Amir et al., 2012; Thielmann et al., 2016) or on average contributions in the Public Goods Game (Amir et al., 2012).

Importantly, much of the previous literature compares incentivized versus hypothetical choices, without distinguishing between different forms of incentivized designs. In particular, it remains unclear whether behavior under random-payment incentive schemes – in which only one of several

decisions is randomly selected for payment – resembles that under full-payment (where all choices are incentivized) or hypothetical conditions (where none are). Both full-payment and random-payment schemes are commonly used in economics (Azrieli et al., 2018), yet their behavioral equivalence is often assumed rather than empirically tested. Charness et al. (2016) suggested that full-payment schemes may introduce hedging and distort preferences, advocating random-payment (whether it means paying a subset of participants or for a subset of decisions) as a potentially more incentive-compatible design, a rationale that has supported their widespread adoption. A few studies have tested these claims directly. Clot et al. (2018), for instance, examined a between-subjects random incentivized system (BRIS) and found behavior in a Dictator Game largely comparable to full-payment under standard stakes, although some differences emerged at higher stakes (note that this study involved a single decision, rather than multiple decisions). Similarly, Bolle (1990) found no difference in Ultimatum Game behavior between pay-all and pay-some incentive schemes. A recent meta-analysis of Dictator Game experiments implementing either a sure-payment incentive scheme (i.e., all participants are paid for all decisions; including studies in which participants only make a single decision) or an incentive scheme involving randomization (either between subjects [BRIS] or between decisions) found no difference in average offers between sure-payment and randomized protocols, suggesting that random-payment protocols and BRIS are comparable to full-payment protocols (Umer, 2023). However, this meta-analysis did not include a comparison with hypothetical decisions.

In the end, the choice of incentive scheme may come down to logistical considerations. Even if paying all participants for all decisions were the most optimal option from a theoretical standpoint, doing so is rarely practically feasible. Furthermore, any form of incentive scheme brings with it an administrative burden; sometimes, paying participants is not possible at all for bureaucratic reasons. Thus, comparing the effect of various incentive schemes on behavior in experimental tasks has real, practical implications for researchers.

Additionally, the most consistent effect of incentives – when there is one – may be that incentives reduce the variance of the data around the predicted outcome, that is, reduce noise (Camerer & Hogarth, 1999; Hertwig & Ortmann, 2001; Smith & Walker, 1993). Incentives may also affect the underlying model of behavior (Berg et al., 2010). This in itself could be a rationale for the use of incentives, even if there is no effect on average behavior.

In this paper, we report two experiments in which we compare a full-payment incentive scheme and a no-payment (hypothetical choice) incentive scheme to a random-payment incentive scheme, and investigate their effect on social decision-making in five standard economic games. Participants in both experiments were randomly assigned to one of three conditions: a *full-payment* incentive condition, in which they were paid for their decisions (and the decisions of other participants) in all games; a *random-payment* incentive condition, in which they were paid for their decisions (and the decision of other participants) in one randomly selected game; or a *no-payment* (hypothetical) condition, in which they were not paid for their decisions in any of the games. All participants played (in random order) a series of five economic games commonly used to measure social preferences: the Dictator Game, Ultimatum Game, Trust Game, Public Goods Game, and Prisoner's Dilemma. We investigate the effect of incentive schemes on behavior in each of these games, as well as a composite measure of prosocial behavior across all games. In Experiment 1, monetary incentives were at the typical stake size for online studies (around £1 per game). In Experiment 2, we increased the stake size by a tenfold (to around £10 per game). The methods and data analysis plans were preregistered on the Open Science Framework (Experiment 1: <https://osf.io/eg2j4>; Experiment 2: <https://osf.io/2q3da/>). We report any deviations from the preregistrations in the text and label non-preregistered analyses as exploratory. Data, materials, and code are openly accessible on <https://doi.org/10.17605/OSF.IO/PA54Z>. We report all measures, conditions, data exclusions, and how we determined sample size.

2. Experiment 1

2.1. Methods

2.1.1. Participants

We requested a sample of 1,500 respondents from Prolific (Palan & Schitter, 2018) in April 2023. To be eligible for the study, participants had to be fluent in English and have completed at least 50 previous works on Prolific with an acceptance rate of 95% or higher. Participants were only allowed to complete the study on a computer (not a mobile phone or tablet). These pre-screening criteria were selected in Prolific before the study was launched. All participants gave their informed consent before continuing to the study. Participants were paid £1.50 for completing the study (based on the hourly rate suggested by Prolific for a study with estimated completion time of about 10 minutes). In addition, participants assigned to the random-payment incentive scheme and full-payment incentive scheme received a bonus payment based on their decisions (and the decisions of other participants) in one or all games, respectively (see [section 2.1.2](#) Experimental design).

We received 1,502 responses. Of these, we discovered one response that had the same participant ID as a previous submission. We kept the first submission and excluded the response that was submitted later (based on the EndDate column in the Qualtrics output file). The final sample included 1,501 responses. Demographic information is provided in [Table 1](#).

2.1.2. Experimental design

At the start of the study, participants were randomly assigned to one of three conditions, determining their payment from the study. All participants were informed that they would complete a series of tasks in which they would be asked to distribute money between themselves and other participants, and that they would receive a payment of £1.50 for completing the study. In addition, participants assigned to the full-payment incentive scheme were informed that they would receive a bonus payment based on their decisions in each of the tasks. Participants assigned to the random-payment incentive scheme were informed that they would receive a bonus payment based on their decisions in one randomly selected task. Participants assigned to the no-payment (hypothetical) incentive scheme received no additional bonus payment based on their decisions in the tasks, but instead were instructed to 'Imagine that you influence the payment of money to yourself and other participants in the experiment'. This phrasing was based on instructions from a previous study comparing hypothetical and incentivized social preferences (Bühren & Kundt, 2015). Participants were informed of the payment scheme at the start of the study and again as a reminder before each game. For complete instructions, see Supplementary material.

All participants played all five games (see below). The presentation order of the games was randomized. Instructions for the games were based on a set of standardized instructions developed by Thielmann et al. (2021), who based their standardized instructions on instructions from studies published in high-ranking journals and/or provided by other researchers in the field. One attention check question was included in the study and was presented randomly among the games. The attention check asked participants to select 'milk' as their favorite drink among a list of six options. Finally, participants answered a set of demographic questions, including age, gender, income, and education.

2.1.3. Games

Dictator Game. In the Dictator Game, participants were randomly paired with one other participant and were asked to distribute £1.00 between themselves and the other participant. Participants were asked to enter the amount in pence and could enter any amount (whole numbers only) between 0 and 100. The outcome measure for data analysis was participants' contribution, as a percentage of the starting endowment.

Public Goods Game. In the Public Goods Game, participants were randomly assigned to a group with three other participants and were asked how much of an endowment of £1.00 to contribute to a group account. Any amount contributed to the group account was doubled and equally distributed

Table 1 Sample characteristics (Experiment 1)

	Full sample	By condition			Balance
	<i>N</i> = 1,501	Hypothetical (<i>n</i> = 501)	Random (<i>n</i> = 500)	Full (<i>n</i> = 500)	<i>p</i>
Gender					.422
Female	681 (45%)	247 (49%)	219 (44%)	215 (43%)	
Male	790 (53%)	245 (49%)	270 (54%)	275 (55%)	
Non-binary/third gender	24 (1.6%)	8 (1.6%)	9 (1.8%)	7 (1.4%)	
Prefer not to say	6 (0.4%)	1 (0.2%)	2 (0.4%)	3 (0.6%)	
Age					.286
Mean (SD)	30 (10)	29 (9)	30 (11)	31 (11)	
Median (Range)	26 (18, 76)	26 (18, 73)	26 (18, 76)	27 (18, 73)	
Annual family income^a					.567
Less than £10,000	293 (20%)	98 (20%)	96 (19%)	99 (20%)	
£10,000 – £19,999	342 (23%)	115 (23%)	122 (24%)	105 (21%)	
£20,000 – £29,999	256 (17%)	81 (16%)	79 (16%)	96 (19%)	
£30,000 – £39,000	154 (10%)	51 (10%)	46 (9.2%)	57 (11%)	
£40,000 – £49,999	100 (6.7%)	39 (7.8%)	38 (7.6%)	23 (4.6%)	
£50,000 – £100,000	206 (14%)	68 (14%)	66 (13%)	72 (14%)	
More than £100,000	61 (4.1%)	22 (4.4%)	17 (3.4%)	22 (4.4%)	
Prefer not to say	89 (5.9%)	27 (5.4%)	36 (7.2%)	26 (5.2%)	
Highest completed education level					.845
No High school	10 (0.7%)	4 (0.8%)	5 (1.0%)	1 (0.2%)	
High school graduate	321 (21%)	105 (21%)	107 (21%)	109 (22%)	
Some college	268 (18%)	89 (18%)	97 (19%)	82 (16%)	
2-year college	131 (8.7%)	47 (9.4%)	38 (7.6%)	46 (9.2%)	
4-year college	479 (32%)	163 (33%)	157 (31%)	159 (32%)	
Post-grad	292 (19%)	93 (19%)	96 (19%)	103 (21%)	

^aDue to a typo in the script, the middle category had an upper bound of £39,000 rather than £39,999.

Notes: *p*-values refer to tests of balance across conditions. Gender (*Male* and *Female*) and education level were compared using Fisher's exact tests; for education level, the *p*-value was estimated via Monte Carlo simulation with 5,000 replications to account for the complexity of the contingency table. Age was compared using the Kruskal–Wallis test, $H(2) = 2.50$. Income levels were compared using Pearson's chi-squared test, $\chi^2(14) = 12.49$.

among all four group members. Participants were asked to enter the amount in pence. The outcome measure was participants' contribution, as a percentage of the starting endowment.

Prisoner's Dilemma. In the Prisoner's Dilemma, participants were randomly paired with one other participant. Both participants simultaneously chose between option A (cooperate) and option B (defect). If both players chose option A, each player received £1.00. If both players chose option B, each player received £0.50. If one player chose option A and the other chose option B, the participant who chose A received £0 and the participant who chose B received £1.50. The outcome measure for data analysis was the percentage of participants choosing option A (cooperate).

Ultimatum Game. In the Ultimatum Game, participants were randomly assigned to one of two roles: proposer or responder. Proposers were endowed with £1.00 and were asked to propose how much of this endowment to offer to the responder. They were asked to enter the amount in pence. Responders were asked to accept or reject the proposer's offer, using the strategy method to elicit their

responses (i.e., responders were asked to accept or reject each possible offer from the proposer using the following intervals: 0–9 pence, 10–19 pence, 20–29 pence, 30–39 pence, 40–49 pence, 50 pence, 51–60 pence, 61–70 pence, 71–80 pence, 81–90 pence, 91–100 pence).

For proposers, the outcome measure was the amount offered to the responder, as a percentage of the starting endowment. For responders, the outcome measure was the minimum offer from the proposer for which they chose *accept* (i.e., the switching point). Responders with multiple switching points or who switched from *accept* to *reject* as offers increased ($n = 106$) were excluded from analyses of Ultimatum Game behavior. This decision was specified in the preregistration.

Trust Game. In the Trust Game, participants were randomly assigned to one of two roles: sender or receiver. Both players were endowed with £0.90. The sender was asked to choose how much to transfer to the receiver: £0, £0.30, £0.60, or £0.90. The transferred amount was tripled in the receiver's account. The receiver then chose how much of the tripled transfer to return to the sender, using the strategy method (i.e., they decided how much to return for each of the possible non-zero transfers from the sender).

For senders, the outcome measure was the amount transferred to the receiver, as a percentage of the starting endowment. Although senders could enter any amount (in whole pence) between 0 and 90, the actual transfer amounts possible were fixed at 0, 30, 60, or 90. If the sender entered any other amount, it was recoded during data analysis to the closest of the four fixed amounts. For example, a transfer of 40p was interpreted as 30p and a transfer of 50p was interpreted as 60p. If the sender entered an amount that was exactly in between two of the fixed amounts, it was interpreted as the closest lower value. For example, 75p was interpreted as 60p. These decisions were specified in the preregistration.

For receivers, the outcome measure was the fraction (in %) sent back to the sender out of the total possible amount across the three items. That is, we calculate the sum total of the receiver's back-transfers and divided this number by the sum total of the maximum possible back-transfers.¹ We also conduct exploratory (non-preregistered) analyses where the outcome measure is the fraction sent back to the sender when the sender transfers the maximum amount possible (90p); results from these analyses are consistent with the main analyses (see Supplementary Table S2).

2.1.4. Data analysis

We report descriptive statistics for each game and condition, as well as aggregated across games (representing an index of overall prosocial behavior). For the composite measure of prosocial behavior, we standardized each outcome measure into a z-score and calculated the mean of these z-scores (for Ultimatum Game responders, we used the reversed z-score). For Ultimatum Game responders with multiple switching points or who switched from *accept* to *reject* as offers increased ($n = 106$), the composite measure was created from all other games except the Ultimatum Game. For correlations among the measures, see Supplementary Table S1.

To investigate differences in prosocial behavior between the three incentive schemes, we performed OLS regressions with prosocial behavior (the composite measure) as dependent variable and condition as predictor variable. The reference category for condition was the random-payment incentive scheme in the main analyses (as preregistered); a comparison between the no-payment and full-payment schemes was added on reviewer request (and thus was not preregistered). A second model also included age and gender as control variables (excluding participants who answered 'non-binary/third gender' [$n = 24$] or 'prefer not to say' [$n = 6$] on the question about gender). We then ran separate regressions for each of the games (logistic regression for the Prisoner's Dilemma;

¹ Although receivers were asked how much of the *tripled transfer* to return to the sender, due to an error the amount that they could enter in the response box ranged from 0 to the tripled transfer *plus* the starting endowment (and this range was displayed to participants). Therefore, we decided to include the starting endowment when calculating the denominator for the outcome measure (i.e., we divided the sum total of the senders' back-transfers by 810 rather than 540). This represents an unanticipated but necessary deviation from the preregistration.

OLS regressions for all other games). We ran these regressions first without and then with the control variables age and gender (again excluding participants who answered 'non-binary/third gender' or 'prefer not to say' on the question about gender). As a robustness test, we repeated all regression analyses while excluding participants ($n = 96$) who failed the attention check.

2.2. Results

2.2.1. Does the average level of prosocial behavior differ between incentive schemes?

Figure 1 shows the average level of prosocial behavior in each game and condition, as well as aggregated across all games (composite measure). Table 2 shows the regression results. Relative to the random-payment incentive scheme, neither the full-payment incentive scheme nor the no-payment (hypothetical) incentive scheme significantly predicted behavior in any of the games or the composite measure, neither before nor after adjusting for age and gender (full-payment incentive scheme: B range -2.836 to 3.058 ; no-payment [hypothetical] incentive scheme: B range -2.767 to 1.510 ; all $ps > .05$). There also was no statistically significant difference in prosocial behavior between the full-payment incentive scheme and hypothetical decisions (exploratory analysis; B range -2.8 to 2.2 ; all $ps > .05$). Results did not change after excluding participants ($n = 96$) who failed the attention check (see Supplementary Table S3).

To further examine the absence of effects, we conducted exploratory (i.e., not preregistered) equivalence tests using the two one-sided t tests (TOST) procedure. Power analyses (using the TOSTER R package; Caldwell, 2022; Lakens, 2017) showed that to achieve 80% power, 429 respondents per group are needed to test an equivalence bound of 0.2, 191 respondents per group to test an equivalence bound of 0.3, and 108 respondents per group to test an equivalence bound of 0.4. For each comparison in the main regressions (Table 2), we ran equivalence tests using the smallest of these three equivalence bounds that would be detectable given the sample sizes. Tests were bootstrapped with 5000 replications. Results are presented in Supplementary Table S4. All but one test were significant ($ps < .05$), indicating evidence for no meaningful true effects, in line with the regression analyses. The exception was the comparison between random-payment and full-payment in the Dictator Game, where the equivalence test was not significant and thus results are inconclusive.

2.2.2. Does the distribution of responses differ between incentive schemes? (exploratory analysis)

Previous research has suggested that increasing the stake size reduces hyper-altruistic behavior (e.g., Brañas-Garza, Jorrot, et al., 2021). To investigate this possibility, we plotted the distribution of responses for each game and incentive scheme and compared the distributions across incentive schemes (see Supplementary Figure S1; for cumulative distributions, see Supplementary Figure S2). These analyses were not preregistered. Kolmogorov-Smirnov tests (for continuous variables) and Kruskal-Wallis tests (for non-continuous variables; i.e., Trust Game senders and Ultimatum Game responders) revealed no statistically significant differences in any of the games between the random-payment incentive scheme and the no-payment (hypothetical) incentive scheme (Dictator Game: $D = 0.033$, $p = .948$; Public Goods Game: $D = 0.02$, $p > .999$; Ultimatum Game proposers: $D = 0.069$, $p = .594$; Ultimatum Game responders: $H(1) = 0.052$, $p = 0.819$; Trust Game senders: $H(1) = 0.049$, $p = 0.825$; Trust Game receivers: $D = 0.092$, $p = .265$), between the random-payment incentive scheme and the full-payment incentive scheme (Dictator Game: $D = 0.054$, $p = .460$; Public Goods Game: $D = 0.07$, $p = .172$; Ultimatum Game proposers: $D = 0.038$, $p = .995$; Ultimatum Game responders: $H(1) = 0.786$, $p = 0.375$; Trust Game senders: $H(1) = 0.594$, $p = .440$; Trust Game receivers: $D = 0.055$, $p = .853$), or between the no-payment and full-payment incentive schemes (Dictator Game: $D = 0.046$, $p = .655$; Public Goods Game: $D = 0.073$, $p = .137$; Ultimatum Game proposers: $D = 0.059$, $p = .762$; Ultimatum Game responders: $H(1) = 1.321$, $p = .250$; Trust Game senders: $H(1) = 0.293$, $p = .588$; Trust Game receivers: $D = 0.07$, $p = .853$). Exploratory (non-preregistered) tests for first-order stochastic dominance (Barrett & Donald, 2003; Schaub, 2024)

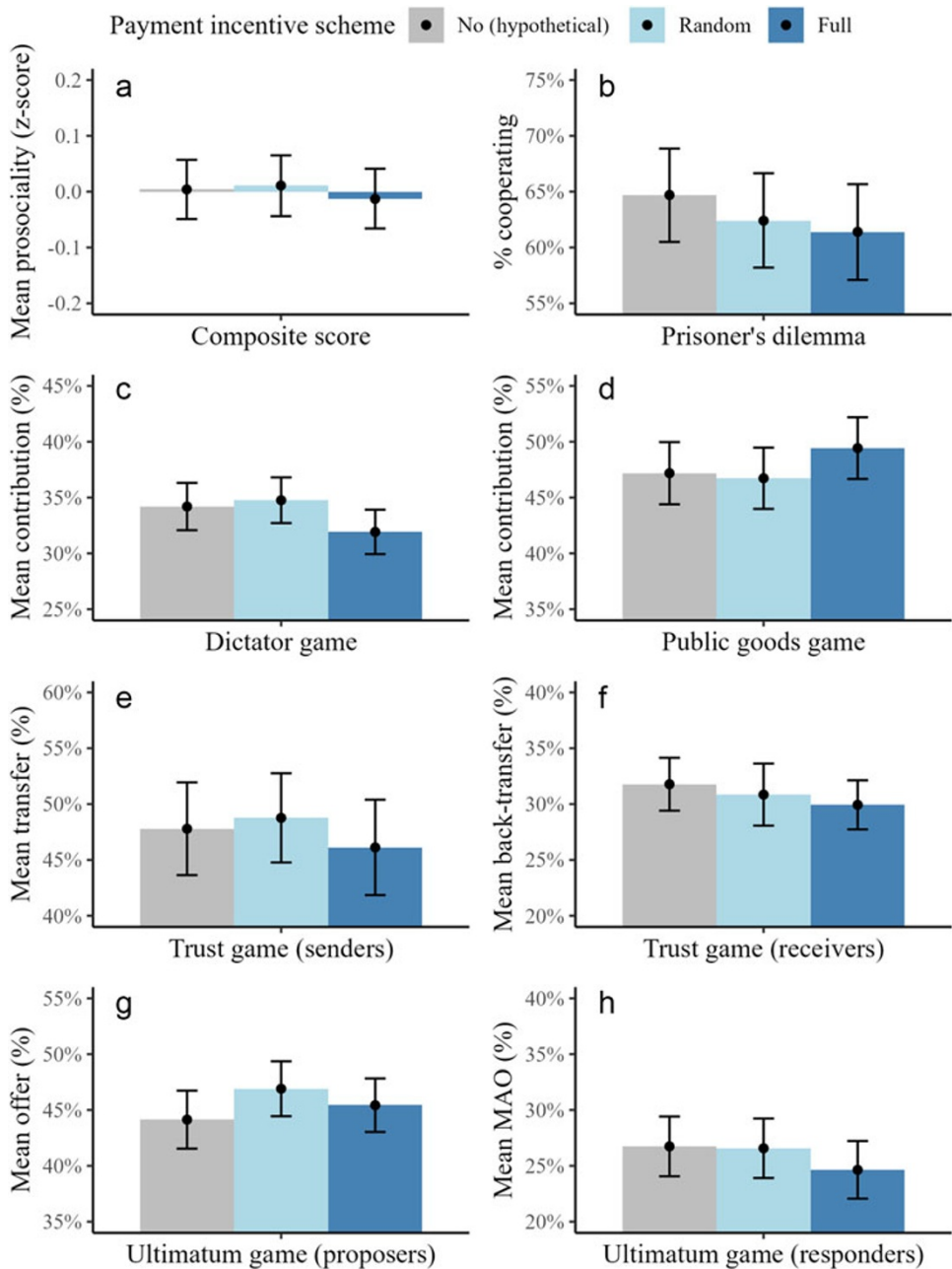


Fig. 1 Level of prosocial behavior in each game and condition (Experiment 1). Error bars represent 95% confidence intervals. (a) Mean prosociality across all games (composite measure; $n = 1501$). (b) Percentage of participants choosing option A (cooperate) in the Prisoner's Dilemma ($n = 1501$). (c) Mean contribution (% of the endowment) in the Dictator Game ($n = 1501$). (d) Mean contribution (% of the endowment) in the Public Goods Game ($n = 1501$). (e) Mean transfer (% of the endowment) in the Trust Game (for participants assigned to the role of sender; $n = 751$). (f) Mean back-transfers (% of total possible back-transfers across all possible non-zero transfers from the sender) in the Trust Game (for participants assigned to the role of receiver; $n = 750$). (g) Mean offer (% of the endowment) in the Ultimatum Game (for participants assigned to the role of proposer; $n = 753$). (h) Mean minimum acceptable offer (MAO; % of the proposer's endowment) in the Ultimatum Game (for participants assigned to the role of responder; $n = 642$)

Table 2 Results from regression analyses of behavior in the games (Experiment 1)

	Composite score		Dictator Game		Public Goods Game		Trust Game (sender)		Trust Game (receiver)		Ultimatum Game (proposer)		Ultimatum Game (responder)		Prisoner's Dilemma	
	Beta (SE)	Beta (SE)	Beta (SE)	Beta (SE)	Beta (SE)	Beta (SE)	Beta (SE)	Beta (SE)	Beta (SE)	Beta (SE)	Beta (SE)	Beta (SE)	Beta (SE)	Beta (SE)	OR (SE)	OR (SE)
Condition																
Hypothetical (ref: Random)	−0.01 (0.039)	0.01 (0.039)	−0.57 (1.50)	−0.36 (1.52)	0.46 (1.99)	1.5 (1.99)	−0.97 (2.93)	0.61 (2.87)	0.93 (1.86)	0.97 (1.88)	−2.8 (1.82)	−2.4 (1.84)	0.17 (1.92)	0.53 (1.90)	1.10 (0.132)	1.11 (0.134)
Full (ref: Random)	−0.02 (0.039)	−0.02 (0.039)	−2.8 (1.45)	−2.6 (1.46)	2.7 (1.98)	3.1 (1.99)	−2.6 (2.98)	−2.2 (2.99)	−0.92 (1.80)	−0.62 (1.82)	−1.5 (1.75)	−1.4 (1.76)	−1.9 (1.89)	−1.8 (1.89)	0.96 (0.131)	0.98 (0.133)
Full (ref: Hypothetical)	−0.02 (0.038)	−0.02 (0.039)	−2.3 (1.48)	−2.2 (1.51)	2.2 (2.00)	1.5 (2.01)	−1.7 (3.03)	−2.8 (3.05)	−1.8 (1.64)	−1.6 (1.66)	1.3 (1.80)	0.96 (1.82)	−2.1 (1.89)	−2.3 (1.88)	0.87 (0.131)	0.88 (0.134)
Age		0.00 (0.001)		0.18 (0.055)**		0.07 (0.074)		0.01 (0.120)		−0.02 (0.064)		0.00 (0.059)		0.27 (0.072)***		1.02 (0.005)***
Gender																
Female (ref: Male)		−0.04 (0.032)		1.9 (1.22)		−8.0 (1.62)***		−14 (2.40)***		−2.1 (1.45)		−2.1 (1.47)		−6.5 (1.54)***		1.36 (0.110)**
<i>N</i>	1,501	1,471	1,501	1,471	1,501	1,471	751	733	750	738	753	740	642	627	1,501	1,471
<i>R</i> ²	0.000	0.003	0.003	0.010	0.001	0.018	0.001	0.045	0.002	0.004	0.003	0.005	0.002	0.042		
AIC															1,986	1,929

Notes: This table shows results from regression analyses of behavior in the games (logistic regression for the Prisoner's Dilemma; OLS regressions for all other games and the composite measure; robust standard errors in parentheses). The first two rows show estimates from models with the Random incentive scheme as reference condition and the third row shows estimates from a model with Hypothetical as the reference condition. OR = Odds ratio. In the Dictator Game, the dependent variable (DV) is the participant's contribution to the other player (% of the endowment). In the Public Goods Game, the DV is the participant's contribution to the group account (% of the endowment). For Trust Game senders, the DV is an ordinal variable indicating the % of the endowment transferred to the receiver (with possible values 0%, 33%, 67%, or 100%). For Trust Game receivers, the DV is the % of the total possible back-transfer that the participant transferred to the sender (i.e., across all possible non-zero transfers from the sender). For Ultimatum Game proposers, the DV is the amount offered to the responder (% of the endowment). For Ultimatum Game responders, the DV is the minimum acceptable offer from the proposer (i.e., the switching point; % of the proposer's endowment). In the Prisoner's Dilemma, the DV is a binary variable, coded as 1 if the participant chose A (cooperate) and 0 if they chose B (defect). For the composite measure of prosocial behavior, each DV was standardized into a z-score and averaged. *Hypothetical* is the no-payment (hypothetical) incentive scheme; *Full* is the full-payment incentive scheme; both variables are coded as binary variables with the random-payment incentive scheme as the reference category. *Age* is the participant's age in years. *Female* is a gender dummy.

** *p* < .01, *** *p* < .001.

indicated that the full-payment incentive scheme dominated the random-payment and no-payment incentive schemes in the Public Goods Game (see Supplementary Figure S2; for full output from these tests, see Supplementary Table S5).

2.2.3. *Do incentive schemes affect the variance of the data? (exploratory analysis)*

Previous research has suggested that financial incentives reduce the variance of the data around the predicted outcome (Camerer & Hogarth, 1999; Hertwig & Ortmann, 2001; Smith & Walker, 1993). To investigate whether there was a difference in the variance of the data between incentive schemes, we performed a Levene's test of homogeneity of variances for each outcome measure. This test was not preregistered. Results revealed no statistically significant differences in any of the games (Dictator Game: $F(2, 1498) = 0.623, p = .536$; Public Goods Game: $F(2, 1498) = 0.489, p = .613$; Trust Game sender: $F(2, 748) = 0.112, p = .894$; Trust Game receiver: $F(2, 747) = 1.632, p = .196$; Ultimatum Game proposer: $F(2, 750) = 1.351, p = .260$; Ultimatum Game responder: $F(2, 639) = 0.495, p = .610$).

3. Experiment 2

In Experiment 1, we found no statistically significant differences in prosocial behavior between a no-payment incentive scheme and a full-payment incentive scheme, relative to a random-payment incentive scheme. A question that begs to be asked in light of these findings is what would happen if the stake size were increased. Could the absence of a difference in behavior between the incentive schemes be attributed to £1 being too trivial for people to care about? Although £1 is at the typical level of stakes for online studies, making our conclusions applicable to existing experimental literature on prosocial behavior, it is still possible that incentive schemes affect decisions with higher stakes. For example, rejection rates of unfair offers in the Ultimatum Game have been shown to decline when stakes increase to levels far exceeding those typically used in lab settings (Andersen et al., 2011; Cameron, 1999), suggesting that the influence of monetary incentives on responder behavior may become more pronounced when stake sizes are high. Furthermore, a recent meta-analysis of Dictator Game giving and Ultimatum Game offers found a small but significant negative effect of stake size in the Dictator Game (i.e., contributions decreased with increasing stake size), although there was no effect in the Ultimatum Game (Larney et al., 2019). All studies in the meta-analysis involved real monetary stakes, although the analysis did not differentiate between different forms of incentive schemes. Thus, it remains an open question whether incentive schemes affect social decision making involving higher stakes. Therefore, in Experiment 2, we repeated Experiment 1 but increased the stake size tenfold.

3.1. Methods

3.1.1. Participants

We requested a sample of 750 respondents from Prolific (Palan & Schitter, 2018) in April 2024. The tenfold increase in stakes compared to Experiment 1 introduced a necessary tradeoff between sample size and incentive level to keep the study financially viable. A power calculation in G*Power 3.1 indicated that a sample size of 250 participants per group provides statistical power of 80% to detect an effect size of $d = 0.25$ in a two-sided between-subjects t-test with $\alpha = .05$. As in Experiment 1, to be eligible for the study, participants had to be fluent in English and have completed at least 50 previous works on Prolific with an acceptance rate of 95% or higher. Individuals who participated in the previous study were also excluded. Participants were only allowed to complete the study on a computer (not a mobile phone or tablet). These pre-screening criteria were selected in Prolific before the study was launched. All participants gave their informed consent before continuing to the study. Participants were paid £1.50 for completing the study (based on the hourly rate suggested by Prolific for a study with estimated completion time of about 10 minutes). In addition, participants assigned to

Table 3 Sample characteristics (Experiment 2)

	Full sample	By condition			Balance
	<i>N</i> = 750	Hypothetical (<i>n</i> = 250)	Random (<i>n</i> = 250)	Full (<i>n</i> = 250)	<i>p</i>
Gender					.502
Female	388 (52%)	130 (52%)	135 (54%)	123 (49%)	
Male	351 (47%)	117 (47%)	110 (44%)	124 (50%)	
Non-binary/third gender	9 (1.2%)	2 (0.8%)	4 (1.6%)	3 (1.2%)	
Prefer not to say	2 (0.3%)	1 (0.4%)	1 (0.4%)	0 (0%)	
Age					.485
Mean (SD)	32 (10)	33 (10)	32 (11)	32 (10)	
Median (Range)	29 (18, 75)	30 (18, 75)	29 (20, 71)	29 (18, 69)	
Annual family income					.075
Less than £10,000	154 (21%)	49 (20%)	57 (23%)	48 (19%)	
£10,000 – £19,999	151 (20%)	53 (21%)	51 (20%)	47 (19%)	
£20,000 – £29,999	122 (16%)	32 (13%)	44 (18%)	46 (18%)	
£30,000 – £39,999	89 (12%)	27 (11%)	40 (16%)	22 (8.8%)	
£40,000 – £49,999	54 (7.2%)	22 (8.8%)	11 (4.4%)	21 (8.4%)	
£50,000 – £100,000	113 (15%)	41 (16%)	33 (13%)	39 (16%)	
More than £100,000	23 (3.1%)	11 (4.4%)	2 (0.8%)	10 (4.0%)	
Prefer not to say	44 (5.9%)	15 (6.0%)	12 (4.8%)	17 (6.8%)	
Highest completed education level					.945
No High school	9 (1.2%)	4 (1.6%)	2 (0.8%)	3 (1.2%)	
High school graduate	103 (14%)	31 (12%)	40 (16%)	32 (13%)	
Some college	113 (15%)	42 (17%)	38 (15%)	33 (13%)	
2-year college	68 (9.1%)	21 (8.4%)	21 (8.4%)	26 (10%)	
4-year college	289 (39%)	96 (38%)	95 (38%)	98 (39%)	
Post-grad	168 (22%)	56 (22%)	54 (22%)	58 (23%)	

Notes: *p*-values refer to tests of balance across conditions. Gender (*Male* and *Female*) and education level were compared using Fisher's exact tests; for education level, the *p*-value was estimated via Monte Carlo simulation with 5,000 replications to account for the complexity of the contingency table. Age was compared using the Kruskal–Wallis test, $H(2) = 1.45$. Income levels were compared using Pearson's chi-squared test, $\chi^2(14) = 22.20$.

the random-payment incentive scheme and full-payment incentive scheme received a bonus payment based on their decisions (and the decisions of other participants) in one or all games, respectively (see [section 2.1.2](#) Experimental design).

We received 751 responses. Of these, we discovered one response that had the same participant ID as a previous submission. We kept the first submission and excluded the response that was submitted later (based on the *EndDate* column in the Qualtrics output file). The final sample included 750 responses. Demographic information is provided in [Table 3](#).

3.1.2. Experimental design

The experimental design followed Experiment 1, except for the following changes: (1) endowments in all games were multiplied by 10; (2) participants were asked to enter amounts in pounds (rather than pence as in the previous study); and (3) we corrected an error in the Trust Game where the amount that receivers could enter into the response box ranged from zero to the tripled transfer *plus*

the starting endowment, so that it only ranged from zero to the tripled transfer. We also corrected a typo in the question about income. For complete instructions, see Supplementary Materials.

3.1.3. Data analysis

Data analysis followed Experiment 1, except some of the analyses that were previously exploratory (non-preregistered; see sections 2.2.2 and 2.2.3) were preregistered for this study. For correlations among the measures that were included in the composite score of prosocial behavior, see Supplementary Table S6. As in Experiment 1, for Ultimatum Game responders with multiple switching points or who switched from *accept* to *reject* as offers increased ($n = 59$), the composite measure was created from all other games except the Ultimatum Game.

3.2. Results

3.2.1. Does the average level of prosocial behavior differ between incentive schemes?

Figure 2 shows the average level of prosocial behavior in each game and condition, as well as aggregated across all games (composite measure). Table 4 shows the regression results. Consistent with Experiment 1, there was no statistically significant difference between incentive schemes in the Dictator Game, Prisoner's dilemma, or Trust Game (for either senders or receivers, although there was a significant effect for senders after controlling for age and gender [see Table 4]). In all cases, p -values exceeded .05, suggesting that even with a tenfold increase in stakes from £1 to £10, prosocial behavior in these tasks was not affected by the form of incentive scheme. However, unlike in Experiment 1, we did find statistically significant differences in the Ultimatum Game, specifically for responders: Compared to participants in the random-payment incentive scheme, participants in the hypothetical incentive scheme had higher minimum acceptable offers (MAOs; $B = 0.95$, $p < .001$), and participants in the full-payment scheme had lower MAOs ($B = -0.48$, $p = .044$; although this effect was no longer statistically significant after controlling for age and gender). Participants in the full-payment incentive scheme also had lower MAOs than participants in the hypothetical incentive scheme (exploratory analysis; $B = -1.4$, $p < .001$). In line with Experiment 1, there were no statistically significant differences between incentive schemes for participants playing in the role of proposer.

Furthermore, unlike Experiment 1, we also found effects in the Public Goods Game: participants who made hypothetical decisions and those who were paid for all decisions contributed more to the public good compared to participants in the random payment condition, and these effects remained after controlling for age and gender. However, these findings are difficult to interpret, as there is no compelling rationale for why both full and hypothetical incentives would promote more cooperation than random payment. Furthermore, there was no statistically significant difference between the full-payment and hypothetical conditions.

For the composite measure of prosocial behavior, as in Experiment 1, there was no significant difference between the hypothetical condition and the random payment scheme. However, unlike Experiment 1, the full-payment incentive scheme increased prosocial behavior relative to both the random-payment scheme ($B = 0.17$, $p < .001$) and the no-payment (hypothetical) scheme (exploratory analysis; $B = 0.14$, $p = .012$), and these associations remained while controlling for age and gender. However, given the heterogeneous results across games – especially the influence of responder behavior in the Ultimatum Game – the interpretation of the composite score warrants caution, as it may reflect an aggregation of inconsistent and potentially unrelated treatment effects. Excluding Ultimatum Game responder behavior from the composite measure in exploratory (non-preregistered) analyses made the comparison between the full-payment and no-payment (hypothetical) incentive schemes non-significant, but the full-payment incentive scheme still increased prosocial behavior compared to the random-payment incentive scheme (significant at $p < .01$ instead of $p < .001$; see Supplementary Table S12).

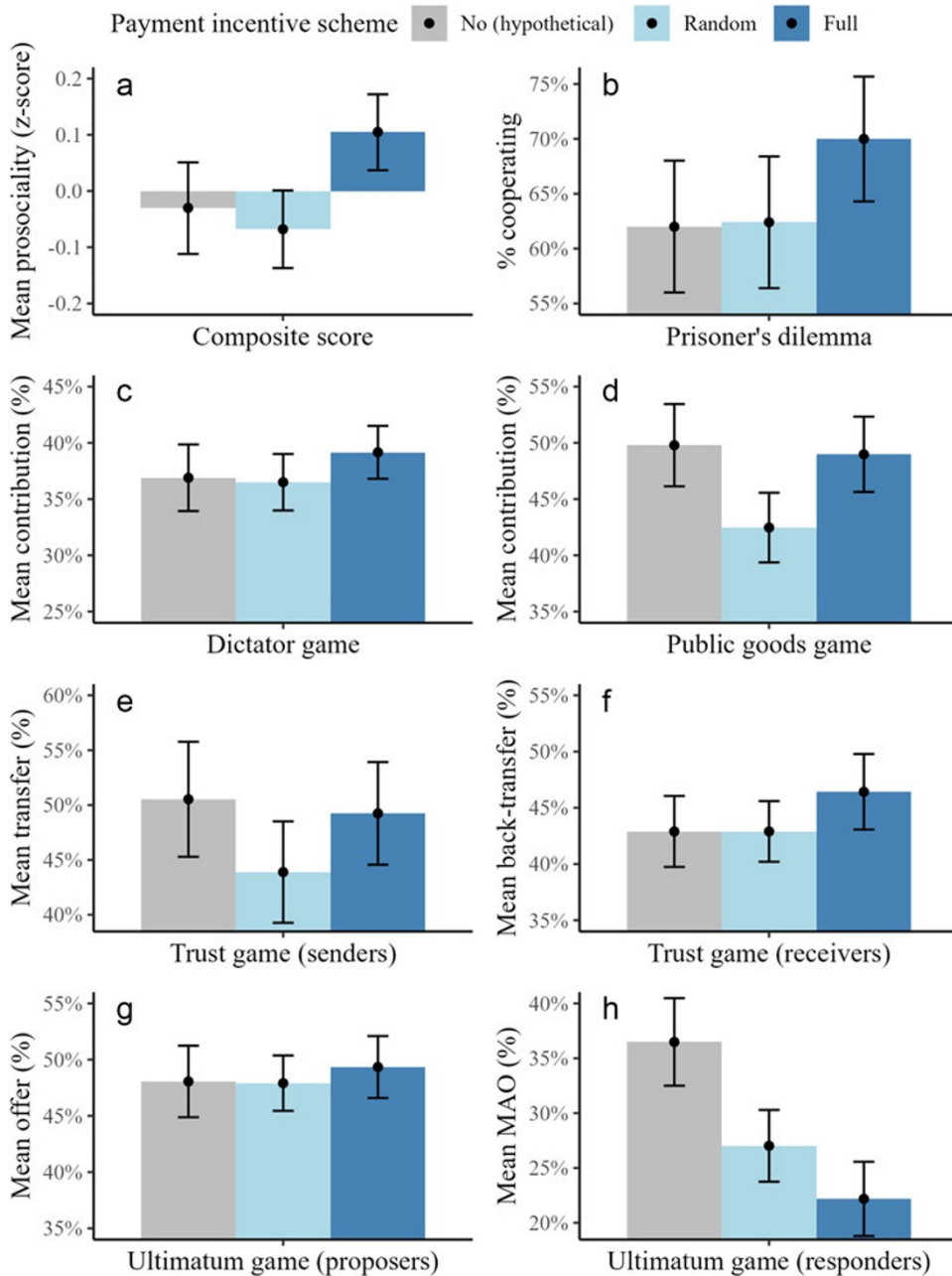


Fig. 2 Level of prosocial behavior in each game and condition (Experiment 2). Error bars represent 95% confidence intervals. (a) Mean prosociality across all games (composite measure; $n = 750$). (b) Percentage of participants choosing option A (cooperate) in the Prisoner's Dilemma ($n = 750$). (c) Mean contribution (% of the endowment) in the Dictator Game ($n = 750$). (d) Mean contribution (% of the endowment) in the Public Goods Game ($n = 750$). (e) Mean transfer (% of the endowment) in the Trust Game (for participants assigned to the role of sender; $n = 378$). (f) Mean back-transfers (% of total possible back-transfers across all possible non-zero transfers from the sender) in the Trust Game (for participants assigned to the role of receiver; $n = 372$). (g) Mean offer (% of the endowment) in the Ultimatum Game (for participants assigned to the role of proposer; $n = 377$). (h) Mean minimum acceptable offer (MAO; % of the proposer's endowment) in the Ultimatum Game (for participants assigned to the role of responder; $n = 314$)

Table 4 Results from regression analyses of behavior in the games (Experiment 2)

	Composite score		Dictator Game		Public Goods Game		Trust Game (sender)		Trust Game (receiver)		Ultimatum Game (proposer)		Ultimatum Game (responder)		Prisoner's Dilemma	
	Beta (SE)	Beta (SE)	Beta (SE)	Beta (SE)	Beta (SE)	Beta (SE)	Beta (SE)	Beta (SE)	Beta (SE)	Beta (SE)	Beta (SE)	Beta (SE)	Beta (SE)	Beta (SE)	OR (SE)	OR (SE)
Condition																
Hypothetical (ref: Random)	0.04 (0.054)	0.04 (0.055)	0.39 (1.98)	0.47 (2.00)	7.3 (2.44)**	7.1 (2.46)**	6.6 (3.54)	7.2 (3.56)*	0.00 (2.11)	−0.02 (2.12)	0.15 (2.04)	−0.04 (2.15)	0.95 (0.261)***	1.0 (0.268)***	0.98 (0.185)	1.02 (0.188)
Full (ref: Random)	0.17 (0.049)***	0.17 (0.050)***	2.7 (1.75)	2.7 (1.78)	6.5 (2.32)**	6.0 (2.32)**	5.4 (3.33)	5.0 (3.33)	3.5 (2.19)	3.4 (2.22)	1.4 (1.88)	1.2 (2.00)	−0.48 (0.239)*	−0.47 (0.240)	1.41 (0.191)	1.44 (0.194)
Full (ref: Hypothetical)	0.14 (0.054)*	0.13 (0.054)*	2.3 (1.92)	2.3 (1.95)	−0.81 (2.52)	−1.1 (2.54)	−1.3 (3.56)	−2.2 (3.61)	3.5 (2.34)	3.4 (2.36)	1.3 (2.14)	1.3 (2.16)	−1.4 (0.265)***	−1.5 (0.268)***	1.43 (0.191)	1.41 (0.193)
Age		0.00 (0.002)		0.06 (0.074)		0.04 (0.094)		−0.24 (0.136)		0.10 (0.086)		0.10 (0.078)		0.00 (0.010)		1.00 (0.008)
Gender																
Female (ref: Male)		−0.06 (0.044)		−0.48 (1.58)		−7.0 (2.01)***		−7.7 (2.88)**		−2.0 (1.86)		−2.2 (1.75)		−0.28 (0.214)		1.49 (0.156)*
N	750	739	750	739	750	739	378	372	372	367	377	368	314	312	750	739
R ²	0.016	0.019	0.003	0.004	0.015	0.031	0.011	0.037	0.009	0.015	0.002	0.010	0.093	0.100		
AIC															975	959

Notes: This table shows results from regression analyses of behavior in the games (logistic regression for the Prisoner's Dilemma; OLS regressions for all other games and the composite measure; robust standard errors in parentheses). The first two rows show estimates from models with the Random incentive scheme as reference condition and the third row shows estimates from a model with Hypothetical as the reference condition. OR = Odds ratio. In the Dictator Game, the dependent variable (DV) is the participant's contribution to the other player (% of the endowment). In the Public Goods Game, the DV is the participant's contribution to the group account (% of the endowment). For Trust Game senders, the DV is an ordinal variable indicating the % of the endowment transferred to the receiver (with possible values 0%, 33%, 67%, or 100%). For Trust Game receivers, the DV is the % of the total possible back-transfer that the participant transferred to the sender (i.e., across all possible non-zero transfers from the sender). For Ultimatum Game proposers, the DV is the amount offered to the responder (% of the endowment). For Ultimatum Game responders, the DV is the minimum acceptable offer from the proposer (i.e., the switching point; % of the proposer's endowment). In the Prisoner's Dilemma, the DV is a binary variable, coded as 1 if the participant chose A (cooperate) and 0 if they chose B (defect). For the composite measure of prosocial behavior, each DV was standardized into a z-score and averaged. *Hypothetical* is the no-payment (hypothetical) incentive scheme; *Full* is the full-payment incentive scheme; both variables are coded as binary variables with the random-payment incentive scheme as the reference category. *Age* is the participant's age in years. *Female* is a gender dummy.

p* < .05, *p* < .01, ****p* < .001.

Results remained after excluding participants ($n = 48$) who failed the attention check, except the effect of the full-payment (versus random-payment) incentive scheme in the Ultimatum Game (responders) was no longer statistically significant and an effect appeared in the Prisoner's Dilemma where participants in the full-payment incentive condition were more likely to cooperate than participants in the no-payment (hypothetical) incentive condition ($OR = 1.49, p < .05$), although this effect was not robust to adjusting for age and gender (see Supplementary Table S7).

An exploratory (not preregistered) analysis of back-transfers in the Trust Game (receivers) when the sender transferred the maximum amount possible indicated that back-transfers were significantly higher in the full-payment incentive scheme than the no-payment (hypothetical) and random-payment incentive schemes (see Supplementary Table S8), although only the former remained statistically significant at $p < .05$ when adjusting for age and gender.

To further examine the mostly non-significant effects, we conducted exploratory (not preregistered) equivalence tests following the same procedure as in Experiment 1. Results are presented in Supplementary Table S9. Results are largely in line with results from the regression analyses: where we find no statistically significant differences in the regressions, the equivalence tests also indicate evidence for no meaningful effects (and vice versa), except for some comparisons in the Trust Game where results are inconclusive.

3.2.2. Does the distribution of responses differ between incentive schemes?

As in Experiment 1, we plotted the distribution of responses for each game and incentive scheme and compared the distributions across incentive schemes (see Supplementary Figure S3; for cumulative distributions, see Supplementary Figure S4). As in Experiment 1, Kolmogorov-Smirnov tests (for continuous variables) and Kruskal-Wallis tests² (for non-continuous variables; i.e., Trust Game senders and Ultimatum Game responders) revealed no statistically significant difference in the distribution of responses between the random-payment and the full-payment incentive scheme in the Dictator Game ($D = 0.104, p = .134$), Public Goods Game ($D = 0.12, p = .055$), or Trust Game (senders: $H(1) = 3.312, p = .069$; receivers: $D = 0.166, p = .066$), between the no-payment and random-payment incentive scheme in the Dictator Game ($D = 0.048, p = .936$) or Trust Game (senders: $H(1) = 3.034, p = .082$; receivers: $D = 0.075, p = .867$), or between the no-payment and full-payment incentive scheme in the Dictator Game ($D = 0.092, p = .241$) and Public Goods Game ($D = 0.048, p = .936$).³ However, unlike Experiment 1, there was a statistically significant difference in the distribution of responses between all incentive schemes for Ultimatum Game responders (no-payment versus random-payment: $H(1) = 11.044, p < .001$; full-payment versus random-payment: $H(1) = 4.21, p = .040$; no-payment versus full-payment: $H(1) = 25.319, p < .001$) but not for proposers (no-payment versus random-payment: $D = 0.088, p = .730$; full-payment versus random-payment: $D = 0.062, p = .968$; no-payment versus full-payment: $D = 0.075, p = .868$). These differences appear to be driven by fewer participants accepting low offers (especially offers of £0–0.99) and more participants requiring an equal split (50%) in the no-payment condition compared to the random-payment and full-payment conditions (see Figure S3). Exploratory (non-preregistered) tests for first-order stochastic dominance (Barrett & Donald, 2003; Schaub, 2024) also indicated that the no-payment incentive scheme dominated the random-payment and full-payment incentive schemes and that the random-payment incentive scheme dominated the full-payment incentive scheme for Ultimatum Game responder behavior (see Supplementary Figure S4; for full output from these tests, see Supplementary Table S10).

There was also a difference between distributions in the no-payment and random-payment incentive scheme in the Public Goods Game ($D = 0.144, p = .011$). This difference appears to be

²In the preregistration, we stated that we would perform Kolmogorov-Smirnov tests for these analyses; however, a reviewer pointed out that another test would be more appropriate for non-continuous variables (i.e., Trust Game senders). Thus, we deviate from the preregistration by using Kruskal-Wallis tests for the analyses of Trust Game senders.

³The comparison between no-payment and full-payment was not preregistered.

driven by more participants contributing 100% of their endowment in the no-payment (hypothetical) incentive condition compared to the random-payment condition (see Figure S3); in other words, the no-payment incentive scheme increased ‘hyper-cooperative’ behavior. This finding is in line with previous literature that has shown that increasing stakes reduce hyper-altruistic behavior (Brañas-Garza, Jorrat, et al., 2021). On the other hand, the proportion of participants that contributed 100% (or near 100%) was also larger in the full-payment incentive condition compared to the random-payment condition, but this difference was not statistically significant. Exploratory (non-preregistered) tests for first-order stochastic dominance also indicated that the no-payment and full-payment incentive schemes dominated the random-payment incentive scheme in the Public Goods Game (see Supplementary Figure S4).

There was also a difference in the distribution of responses between the no-payment and full-payment incentive schemes for Trust Game receivers ($D = 0.188$, $p = .028$; but not senders ($H(1) = 0.006$, $p = .939$). This difference appears to be driven by fewer participants in the full-payment condition transferring 31–40% of the possible amount and more participants transferring 51–80% (see Supplementary Figure S3). Exploratory (non-preregistered) tests for first-order stochastic dominance also indicated that the full-payment incentive scheme dominated the no-payment incentive scheme (see Supplementary Figure S4 and Supplementary Table S10). Additionally, the full-payment scheme dominated the random-payment incentive scheme.

In addition to the results reported above, the exploratory tests for first-order stochastic dominance also indicated that the full-payment incentive scheme dominated the no-payment and random-payment incentive schemes in the Dictator Game and that the no-payment and full-payment incentive schemes dominated the random-payment incentive scheme for Trust Game senders (see Supplementary Figure S4 and Table S10).

Finally, we conducted exploratory (non-preregistered) analyses investigating specifically whether there was a difference in the proportion of participants giving 0%, 50%, and 100% in the Dictator Game; these analyses indicated a significant difference in the proportion of participants giving 0% ($\chi^2 = 6.53$, $p = .038$, $n = 102$), with more participants giving 0 in the no-payment (hypothetical) incentive condition, followed by the random-payment incentive condition and lastly the full-payment condition (see Supplementary Table S11). There was also a significant difference in the proportion of participants giving 100%; however, this analysis includes only 15 participants (2% of our sample).

3.2.3. Do incentive schemes affect the variance of the data?

To investigate whether there was a difference in the variance of the data between incentive schemes, we performed a Levene’s test of homogeneity of variances for each outcome measure (except the Prisoner’s Dilemma). As in Experiment 1, there was no statistically significant difference in the variance of the data in the Public Goods Game ($F(2, 747) = 2.101$, $p = .123$), Trust Game (senders: $F(2, 375) = 2.526$, $p = .081$; receivers: $F(2, 369) = 1.005$, $p = .367$), or Ultimatum Game (proposers: $F(2, 374) = 1.451$, $p = .236$; responders: $F(2, 311) = 0.398$, $p = .672$). However, unlike Experiment 1, there was a statistically significant difference in the variance of the data in the Dictator Game ($F(2, 747) = 5.284$, $p = .005$). Inspection of box plots indicated that the variance was lower in the full-payment incentive scheme (interquartile range [IQR] = 0.30 – 0.50) than in the other conditions (hypothetical: IQR = 0.20 – 0.50; full-payment: IQR = 0.21 – 0.50).

4. Discussion

We conducted two preregistered behavioral experiments to investigate the effect of different incentive schemes on prosocial behavior. We found no statistically significant effect of a full-payment or a no-payment (hypothetical) incentive scheme relative to a random-payment incentive scheme on average behavior in one-shot economic games and social dilemmas with £1 stakes. This finding was consistent across all games and dilemmas included, as well as a composite measure of prosocial behavior. In

a follow-up experiment where stakes were increased to £10, we still found no consistent effect of the incentive schemes, although effects varied somewhat across games and outcome measures. Our results raise questions about the reasonableness of the prevalent norm in experimental economics to exclusively rely on incentivized experiments as the sole legitimate type of experimental input for understanding human behavior.

The use of incentivized choice tasks is a methodological convention that separates economics from other social sciences such as psychology. Yet empirical research has provided mixed results on the impact of incentive schemes on behavior in experimental settings. Although incentives may matter in some domains and for some types of measures, the most consistent finding seems to be that when monetary incentives influence behavior (if there is any effect at all), they reduce the variance of the data around the predicted outcome (Camerer & Hogarth, 1999; Hertwig & Ortmann, 2001; Smith & Walker, 1993). Here, we find no significant difference between incentive schemes either in average prosocial behavior or in the variance of the data when choices involve £1 stakes. The non-significant effects on average behavior were corroborated in exploratory (non-preregistered) equivalence tests indicating no meaningful true effects, although evidence was inconclusive for one of three comparisons in the Dictator Game (between full- versus random-payment). Tests for first-order stochastic dominance also found no significant effects except in the Public Goods Game, where the full-payment incentive scheme dominated the random-payment and no-payment incentive schemes. When choices involved £10 stakes (Experiment 2), the full-payment incentive scheme increased overall prosocial behavior (the composite measure) relative to the random-payment and no-payment (hypothetical) incentive schemes; however, looking separately at each game, the only case where we find consistent and statistically significant differences between incentive schemes is in the Ultimatum Game, where responders in the full-payment condition accepted lower offers from the proposer compared to responders in the hypothetical incentive scheme, with responders in the random-payment incentive scheme falling in between. This pattern aligns with Andersen et al. (2011), who demonstrated that rejection rates decline as stakes increase, particularly for low offers. However, the interpretation of responder behavior in the Ultimatum Game is complex: Low minimum acceptable offers (MAOs) could reflect selfish maximization or altruistic tendencies, as those who accept zero in the Ultimatum Game have been shown to be the most prosocial in the Dictator Game (Staffiero et al., 2013). Likewise, rejecting unfair offers may indicate a form of punishment by individuals who behave fairly, or spite by individuals who behave unfairly (Brañas-Garza et al., 2014). The ambiguity of Ultimatum Game responses affects not only the interpretation of Ultimatum Game responder behavior, but also of the composite measure, as the Ultimatum Game is included in the composite measure, in line with previous research (e.g., Gärtner et al., 2022). Excluding Ultimatum Game responder behavior from the composite measure in exploratory (non-preregistered) analyses made the comparison between the full-payment and no-payment (hypothetical) incentive schemes non-significant, but the full-payment incentive scheme still increased prosocial behavior compared to the random-payment incentive scheme.

In Experiment 2, we also found statistically significant effects in the Public Goods Game; however, these effects appear self-contradictory, as they indicate that prosocial behavior increases both when participants are paid for all decisions (full-payment incentive scheme) and when they are not paid at all (hypothetical incentive scheme), compared to when they are paid for a randomly selected decision (random-payment incentive scheme). The only circumstance under which these seemingly contradictory effects would make sense is if they are driven by two separate mechanisms. For example, it is possible that the full-payment incentive scheme increased contributions in the Public Goods Game because participants were paid a larger total payoff from the experiment, making prosocial behavior in each individual task ‘cheaper’ in relation to the total payoff. In contrast, the no-payment condition may have increased contributions because acting prosocially had no personal cost. However, these explanations are post-hoc rationalizations of the observed data rather than theory-driven predictions; based on existing literature on incentive schemes it is hard to find any reason why both a full-payment

and a no-payment incentive scheme would increase contributions in the Public Goods Game, compared to a random-payment incentive scheme. Therefore, these effects may be due to noise. In either case, our results make it clear that hypothetical choice tasks do not necessarily lead to inflated rates of prosocial behavior, as one might think. If anything, incentives (particularly the full-payment incentive scheme) *increased* prosocial behavior in the present study, as this was the effect that most consistently appeared in the Ultimatum Game, Public Goods Game, and composite measure taken together, although the precise interpretation is somewhat ambiguous and in most tasks there was no effect at all.

Exploratory (non-preregistered) equivalence tests for Experiment 2 also indicated evidence in line with the conclusions above (i.e., no meaningful effects in the Dictator Game or Ultimatum Game proposer behavior, but some effects in the Public Goods Game, Ultimatum Game responder behavior, and the composite measure). These tests also indicated evidence of no meaningful effects in the Trust Game, except for some comparisons where evidence was inconclusive (specifically, for senders, evidence was inconclusive for random- versus no-payment and random- versus full-payment, and for receivers, evidence was inconclusive for random- versus full-payment). Exploratory tests for first-order stochastic dominance also indicated that no-payment and full-payment dominated random-payment for Trust Game senders (which, as in the Public Goods Game, appears self-contradictory) and that full-payment dominated both no-payment and random-payment for Trust Game receivers. Thus, even though there were no significant effects for average behavior in the Trust Game, there may be some differences in the underlying distributions, although the precise interpretation and practical implication is somewhat elusive. Furthermore, in the Trust Game, when first-players transferred the maximum possible amount, second-players in the full-payment incentive scheme back-transferred a larger amount than second-players in the random-payment or hypothetical incentive schemes. However, these analyses were exploratory (not preregistered) and need to be confirmed in future studies.

Finally, we found no significant differences in the variance of the data, except in the Dictator Game (in Experiment 2), where the full-payment incentive scheme reduced variance. This is in line with expectations (Camerer & Hogarth, 1999; Hertwig & Ortmann, 2001; Smith & Walker, 1993), although a recent study found effects in the opposite direction: hypothetical choices were less dispersed than choices that were fully incentivized (Brañas-Garza et al., 2025). Exploratory tests also found that the full-payment incentive scheme stochastically dominated the no-payment and random-payment incentive schemes in the Dictator Game (in Experiment 2).

Much of the previous literature on the role of incentives in prosocial behavior has focused on comparisons between decisions that are either incentivized or not, without distinguishing between different forms of incentive schemes. An exception is a recent meta-analysis of Dictator Game experiments, which found no difference in average offers between sure-payment protocols (i.e., studies where participants are paid for all their decisions) and protocols in which participants are paid for one or some of their decisions, suggesting that random-payment protocols are as effective as full-payment protocols (Umer, 2023). Importantly, we here also included a condition in which decisions were entirely hypothetical; thus, whereas Umer (2023) concludes that random-payment protocols are as effective as full-payment protocols, the findings from the present studies suggest that random-payment protocols are as *ineffective* as full-payment protocols, as we found little difference in average behavior between the full-payment or the no-payment (hypothetical) incentive schemes, relative to the random-payment incentive scheme. This conclusion is in line with recent work suggesting that incentives (versus no incentives) do not affect risk preferences (Brañas-Garza, Estepa-Mohedano, et al., 2021; Hackethal et al., 2022) or time preferences (Brañas-Garza et al., 2023), as well as a previous study from our team where monetary incentives had no effect on payoff comprehension in the same economic games that we included here (Koppel et al., 2025). It is also in line with a recent meta-analysis of effects of incentives on performance across various contexts, which found little evidence that incentives improve performance (Cala et al., 2026). On the other hand, we did find

effects in some games at higher stake sizes (£10), although effects were not consistent across games. Furthermore, tests of differences in distributions and (exploratory) first-order stochastic dominance suggest that in some games, even though there is no effect on average behavior or evidence is inconclusive, there may be differences in the underlying distributions of responses. However, the practical implication of those differences is somewhat unclear.

A question that begs to be asked in light of the current findings is whether our results are good or bad news for experimental research in this domain. On one hand, implementing incentive-compatible choice tasks is often administratively burdensome and time consuming. If hypothetical scenarios yield equally valid responses, it would significantly simplify and reduce the costs of conducting experiments in this field, making life easier for experimental researchers. On the other hand, our findings could be interpreted as indicating that incentive-compatible choice tasks yield equally *unreliable* results as hypothetical choice tasks, which would pose a significant challenge to the future of experimental research on prosocial behavior. Concerns about external validity have persistently plagued this field (see, e.g., Bostyn et al., 2018; Galizzi & Navarro-Martinez, 2019). Our findings could intensify the argument that experimental researchers should shift their focus toward behavior in natural settings rather than conducting experiments in highly stylized environments that offer limited insight into real-life behavior.

A limitation of the present study is that we did not include a between-subjects random incentive scheme (BRIS), where there is a random drawing such that some participants receive payment for some (or all) of their decisions. A benefit of BRIS is that it allows for even higher stake sizes without increasing the costs of the experiment. On the other hand, like any incentive scheme that involves some form of randomization, participants may treat a given choice not in terms of its stated payoff but in terms of some fraction of that payoff. Relatedly, in the present study, we did not inform participants of the exact number of tasks that they would complete, which may add ambiguity about the probabilities of payoffs in the random-payment condition, potentially creating an additional source of variation in this condition.

Additionally, our sample consisted of online adult participants recruited via Prolific rather than a typical university student sample, which is commonly used in experimental economics. While our sampling choice may enhance external validity of the findings, it limits direct comparability with many previous studies and raises questions about generalizability across different populations and experimental contexts. Moreover, our findings may not generalize to other, more complex, decision environments such as markets or auctions. In addition, it is possible that group-level averages obscure meaningful individual-level variation. Some individuals may become more prosocial in hypothetical contexts, while others may behave less prosocially. These opposing tendencies could cancel out in the aggregate, masking underlying heterogeneity. As our study was not designed to investigate such trait-based moderation, future research using richer psychological or personality measures could explore how individual differences shape responses to different incentive schemes (for a comparison of incentive schemes in terms of attention, response time, and previous experience, see Supplementary Table S12; exploratory tests revealed no statistically significant differences between conditions).

These points also raise the broader issue of which incentive scheme, if any, best approximates behavior outside the lab. Full-payment schemes offer clear financial consequences but may induce artificial effects such as hedging across tasks (see, e.g., Charness et al., 2016). Hypothetical choices, while not monetarily consequential, may reflect moral or reputational motivations relevant in many real-world social interactions. On the other hand, experimental instructions about hypothetical payments may be difficult for participants to understand; participants may still believe that there is some probability of receiving payment, despite being instructed otherwise (Brañas-Garza et al., 2025). Random-payment schemes offer some balance between these features, yet their external validity remains largely assumed. Given that many everyday prosocial decisions – such as helping others, cooperating, or donating – occur without explicit financial incentives, hypothetical or low-stakes

designs may sometimes better capture naturally occurring behavior. That said, the most appropriate incentive scheme ultimately depends on the specific behavioral domain and context under investigation. Our findings suggest that in typical online environments, the choice of incentive scheme may be less critical than is often assumed in experimental economics.

The idea that only incentivized decisions can be trusted has for decades been a near mantra in the field of experimental economics, and viewed as a defining feature that sets it apart from experimental work in other social sciences. For example, this journal (*Experimental Economics*) explicitly states in their publication guidelines that it only considers studies in which participants are incentivized. We think our results call for a more nuanced approach toward the paradigmatic role of monetary incentives in economics in general, and experimental economics in particular. What this more nuanced approach entails we are not certain of, but it should not be based on ‘check-list thinking’ about the use of monetary incentives for assessing whether a study provides valuable insights about human prosociality. Instead, researchers should carefully consider the specific context, goals, and underlying assumptions of their study and, in case incentives are deemed not necessary, explicitly justify their choice. Referees and editors, in their turn, should not reject papers purely on the grounds that subjects were not paid. This will help to bridge the divide between economics and other social sciences in the study of human behavior.

Supplementary material. The supplementary material for this article can be found at <https://doi.org/10.1017/eec.2026.10044>. The replication material for the study is available at <https://doi.org/10.17605/OSF.IO/PA54Z>

Acknowledgements. We thank Magnus Johannesson and Eirik Strömland for helpful comments on earlier versions of this work. This research was funded by the Swedish Research Council. Funders had no role in study design, data collection, analysis, decision to publish, or preparation of the manuscript.

Competing interests. The authors have no conflicts of interest to declare.

References

- Amir, O., Rand, D. G., & Gal, Y. K. (2012). Economic games on the internet: The effect of \$1 stakes. *PLoS ONE*, 7(2), e31461. <https://doi.org/10.1371/journal.pone.0031461>
- Andersen, S., Ertac, S., Gneezy, U., Hoffman, M., & List, J. A. (2011). Stakes matter in ultimatum games. *American Economic Review*, 101(7), 3427–3439. <https://doi.org/10.1257/aer.101.7.3427>
- Azrieli, Y., Chambers, C. P., & Healy, P. J. (2018). Incentives in experiments: A theoretical analysis. *Journal of Political Economy*, 126(4), 1472–1503. <https://doi.org/10.1086/698136>
- Barrett, G. F. & Donald, S. G. (2003). Consistent tests for stochastic dominance. *Econometrica*, 71(1), 71–104. <https://doi.org/10.1111/1468-0262.00390>
- Ben-Ner, A., Kramer, A., & Levy, O. (2008). Economic and hypothetical dictator game experiments: Incentive effects at the individual level. *The Journal of Socio-Economics*, 37(5), 1775–1784. <https://doi.org/10.1016/j.socec.2007.11.004>
- Berg, J. E., Dickhaut, J. W., & Rietz, T. A. (2010). Preference reversals: The impact of truth-revealing monetary incentives. *Games and Economic Behavior*, 68(2), 443–468. <https://doi.org/10.1016/j.geb.2009.07.008>
- Bolle, F. (1990). High reward experiments without high expenditure for the experimenter? *Journal of Economic Psychology*, 11(2), 157–167. [https://doi.org/10.1016/0167-4870\(90\)90001-P](https://doi.org/10.1016/0167-4870(90)90001-P)
- Bostyn, D. H., Sevenhant, S., & Roets, A. (2018). Of mice, men, and trolleys: Hypothetical judgment versus real-life behavior in trolley-style moral dilemmas. *Psychological Science*, 29(7), 1084–1093. <https://doi.org/10.1177/0956797617752640>
- Brañas-Garza, P., Espín, A. M., Exadaktylos, F., & Herrmann, B. (2014). Fair and unfair punishers coexist in the Ultimatum Game. *Scientific Reports*, 4. Article 6025. <https://doi.org/10.1038/srep06025>
- Brañas-Garza, P., Espín, A. M., & Jorrat, D. (2025). Paying £1(£5) or nothing in dictator games: Unexpected differences. *SSRN Electronic Journal*. <https://dx.doi.org/10.2139/ssrn.4723871>
- Brañas-Garza, P., Estepa-Mohedano, L., Jorrat, D., Orozco, V., & Rascón-Ramírez, E. (2021). To pay or not to pay: Measuring risk preferences in lab and field. *Judgment and Decision Making*, 16(5), 1290–1313. <https://doi.org/10.1017/S1930297500008433>
- Brañas-Garza, P., Jorrat, D., Espín, A. M., & Sánchez, A. (2023). Paid and hypothetical time preferences are the same: Lab, field and online evidence. *Experimental Economics*, 26(2), 412–434. <https://doi.org/10.1007/s10683-022-09776-5>
- Brañas-Garza, P., Jorrat, D., Kovářík, J., & López, M. C. (2021). Hyper-altruistic behavior vanishes with high stakes. *PLOS ONE*, 16(8), e0255668. <https://doi.org/10.1371/journal.pone.0255668>

- Bühren, C. & Kundt, T. C. (2015). Imagine being a nice guy: A note on hypothetical vs. incentivized social preferences. *Judgment and Decision Making*, 10(2), 185–190. <https://doi.org/10.2139/ssrn.2487072>
- Cala, P., Havranek, T., Irsova, Z., Luskova, M., Matousek J., & Novak, J. (2026). Financial incentives and performance: A meta-analysis of experiments in economics. *Journal of Political Economy Microeconomics*.
- Caldwell, A. R. (2022). Exploring equivalence testing with the updated TOSTER R package. PsyArXiv. <https://doi.org/10.31234/osf.io/ty8de>
- Camerer, C. F. & Hogarth, R. M. (1999). The effects of financial incentives in experiments: A review and capital-labor-production framework. *Journal of Risk and Uncertainty*, 19, 7–42. https://doi.org/10.1007/978-94-017-1406-8_2
- Cameron, L. A. (1999). Raising the stakes in the ultimatum game: Experimental evidence from Indonesia. *Economic Inquiry*, 37(1), 47–59. <https://doi.org/10.1111/j.1465-7295.1999.tb01415.x>
- Charness, G., Gneezy, U., & Halladay, B. (2016). Experimental methods: Pay one or pay all. *Journal of Economic Behavior and Organization*, 131, 141–150. <https://doi.org/10.1016/j.jebo.2016.08.010>
- Clot, S., Grolleau, G., & Ibanez, L. (2018). Shall we pay all? An experimental test of random incentivized systems. *Journal of Behavioral and Experimental Economics*, 73, 93–98. <https://doi.org/10.1016/j.socec.2018.01.004>
- Erlandsson, A. (2015). *Underlying psychological mechanisms of helping effects: Examining the when × why of charitable giving* [Thesis/doccomp, Lund University]. <http://lup.lub.lu.se/record/4861638>
- Galizzi, M. M. & Navarro-Martinez, D. (2019). On the external validity of social preference games: A systematic lab-field study. *Management Science*, 65(3), 976–1002. <https://doi.org/10.1287/mnsc.2017.2908>
- Gärtner, M., Andersson, D., Västfjäll, D., & Tinghög, G. (2022). Affect and prosocial behavior: The role of decision mode and individual processing style. *Judgment and Decision Making*, 17(1), 1–13. <https://doi.org/10.1017/S1930297500008998>
- Hackethal, A., Kirchler, M., Laudenbach, C., Razen, M., & Weber, A. (2022). On the role of monetary incentives in risk preference elicitation experiments. *Journal of Risk and Uncertainty*, 66, 189–213. <https://doi.org/10.1007/s11166-022-09377-w>
- Harrison, G. W. (1994). Expected utility theory and the experimentalists. In J. D. Hey (Ed.), *Experimental Economics* (pp. 43–73). Heidelberg: Physica. https://doi.org/10.1007/978-3-642-51179-0_4
- Hertwig, R. & Ortmann, A. (2001). Experimental practices in economics: A methodological challenge for psychologists? *Behavioral and Brain Sciences*, 24(3), 383–403. <https://doi.org/10.1017/S0140525X01004149>
- Koppel, L., Andersson, D., Johannesson, M., Strömblad, E., & Tinghög, G. (2025). Comprehension in economic games. *Journal of Economic Behavior and Organization*, 234, 107039. <https://doi.org/10.1016/j.jebo.2025.107039>
- Lakens, D. (2017). Equivalence tests: A practical primer for t-tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, 8(4), 355–362. <https://doi.org/10.1177/1948550617697177>
- Larney, A., Rotella, A., & Barclay, P. (2019). Stake size effects in ultimatum game and dictator game offers: A meta-analysis. *Organizational Behavior and Human Decision Processes*, 151, 61–72. <https://doi.org/10.1016/j.obhdp.2019.01.002>
- Palan, S. & Schitter, C. (2018). Prolific.ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, 17, 22–27. <https://doi.org/10.1016/j.jbef.2017.12.004>
- Schaub, S. (2024). stodom: Estimating consistent tests for stochastic dominance. R package version 0.0.1. <https://doi.org/10.32614/CRAN.package.stodom>
- Smith, V. L. (1976). Experimental economics: Induced value theory. *The American Economic Review*, 66(2), 274–279.
- Smith, V. L. & Walker, J. M. (1993). Monetary rewards and decision cost in experimental economics. *Economic Inquiry*, 31(2), 245–261. <https://doi.org/10.1111/j.1465-7295.1993.tb00881.x>
- Staffiero, G., Exadaktylos, F., & Espín, A. M. (2013). Accepting zero in the ultimatum game does not reflect selfish preferences. *Economics Letters*, 121(2), 236–238. <https://doi.org/10.1016/j.econlet.2013.08.021>
- Thielmann, I., Böhm, R., Ott, M., & Hilbig, B. E. (2021). Economic games: An introduction and guide for research. *Collabra: Psychology*, 7(1), 19004. <https://doi.org/10.1525/collabra.19004>
- Thielmann, I., Heck, D. W., & Hilbig, B. E. (2016). Anonymity and incentives: An investigation of techniques to reduce socially desirable responding in the Trust Game. *Judgment and Decision Making*, 11(5), 10. <https://doi.org/10.1017/S1930297500004605>
- Thielmann, I., Spadaro, G., & Balliet, D. (2020). Personality and prosocial behavior: A theoretical framework and meta-analysis. *Psychological Bulletin*, 146(1), 30–90. <https://doi.org/10.1037/bul0000217>
- Umer, H. (2023). Effectiveness of random payment in experiments: A meta-analysis of dictator games. *Journal of Economic Psychology*, 96, 102608. <https://doi.org/10.1016/j.joep.2023.102608>