Journal of Economic Psychology 114 (2026) 102901

Contents lists available at ScienceDirect

Journal of Economic Psychology

journal homepage: www.elsevier.com/locate/joep

Validating a survey measure of conditional cooperation☆

Francesco Fallucchi [a,*], Andrea Marietta Leina [b], Simone Quercia [c]

[a] University of Bergamo, Department of Economics, Via dei Caniana 2, Bergamo, Italy
[b] University of Bologna, Department of Economics, Piazza Scaravilli 2, Bologna, Italy
[c] University of Verona, Department of Economics, Via Cantarane 24, Verona, Italy

ARTICLE INFO

Keywords:
Survey question
Experiment validation
Conditional cooperation
Prisoner's dilemma

ABSTRACT

We study the ability of survey-based measures to predict conditional cooperation in an incentivized Prisoner's Dilemma (PD) game. We assess whether (i) hypothetical game play, (ii) unincentivized social norms, (iii) survey measures of economic preferences, and (iv) personality traits predict conditionally cooperative behavior when monetary stakes are introduced. Our findings reveal that hypothetical PD responses are the strongest predictors of incentivized behavior, with limited evidence of hypothetical bias. Notably, patience is negatively correlated with conditional cooperation, contrary to expectations. Surprisingly, other economic and social preference measures, including reciprocity and normative evaluations, exhibit weak or no predictive power. These results contribute to the debate on the feasibility of survey-based proxies in behavioral research. They suggest that well-designed hypothetical games can reliably substitute incentivized experiments, but that deviations from accurately mirroring the task may weaken the predictive power of survey measures.

1. Introduction

Incentivization is a cornerstone of experimental economics (Hertwig & Ortmann, 2001; Smith, 1982). Although incentives guarantee tight control over preference revelation, they also make experimentation costly, often limiting economic experiments to smaller samples compared to large-scale representative surveys. In recent years, however, due to the availability of large online panels, scholars have started to conduct non-incentivized large-scale experiments and surveys (Stantcheva, 2023). Yet, it is an open question whether — and how — the preference measures that have been operationalized in the lab can be reliably transferred to non-incentivized surveys.

In this paper, we investigate this issue in the context of cooperation in social dilemmas. One of the most common findings in social dilemma experiments — both in the lab and in the field — is that while some individuals behave as purely self-interested, many are conditional cooperators: they choose to cooperate only if others do the same (Fallucchi et al., 2019; Fischbacher et al., 2001, 2012; Frey & Meier, 2004; Thöni & Volk, 2018). Conditional cooperation has been recognized as a crucial element in explaining the successes and failures of cooperation in real-world dilemmas (Andre et al., 2024; Fehr-Duda & Fehr, 2016; Rustagi et al., 2010), making it important to have clean measures of conditionally cooperative attitudes (Fischbacher & Gächter, 2010; Kölle et al., 2025). Insights along these lines can also guide the design of targeted interventions to promote cooperation (Allcott, 2011).

Building on Falk et al. (2023), our study examines how well a range of purely hypothetical, survey-based constructs can predict who will act as a conditional cooperator in a real-stakes Prisoner's Dilemma (PD). Although conditional cooperation has mainly been

☆ This article is part of a Special issue entitled: 'Survey Design' published in Journal of Economic Psychology.

* Corresponding author.
E-mail addresses: francesco.fallucchi@unibg.it (F. Fallucchi), andrea.mariettaleina@gmail.com (A. Marietta Leina), simone.quercia@univr.it (S. Quercia).

https://doi.org/10.1016/j.joep.2026.102901
Received 22 May 2025; Received in revised form 24 February 2026; Accepted 15 March 2026
Available online 20 March 2026

elicited in the public goods game, recent evidence has shown that types transfer consistently between the public goods game and the prisoner's dilemma (Eichenseer & Moser, 2020). Hence, we combine an online survey with an incentivized, sequential PD game to investigate whether (i) hypothetical game play, (ii) unincentivized social norms, (iii) survey measures of economic preferences, and (iv) a wide range of personality traits and attitudes align with observed behavior once monetary payoffs enter the picture. Specifically, we ask: which are the best measures to predict conditional cooperation?

Our results show that eliciting cooperation types in a hypothetical PD strongly predicts how people behave in an incentivized PD game. Although some individuals do switch from conditional cooperation in the hypothetical game to free-riding under real monetary stakes, over 70% remain in the same classification. Across all our specifications, the single most robust predictor of incentivized conditional cooperation is being identified as a conditional cooperator in the hypothetical PD. Interestingly, the second significant predictor is patience, which is negatively related to conditional cooperation. Strikingly, no other variable systematically predicts conditional cooperation across all our specifications. This finding aligns with recent evidence that the predictive relationships of commonly used survey measures — often presumed to tap into core preferences — may fail to replicate across different contexts (Chapman et al., 2025; Kosfeld et al., 2025).

As a second step in our analysis, we use a machine-learning technique (LASSO) and find that certain altruism and reciprocity items in the Global Preference Survey (GPS) module (Falk et al., 2023) are significant in some specifications—but none prove as consistently predictive as the hypothetical PD measure itself.

Overall, our study contributes to the existing literature in three main ways. First, we contribute to the long-standing debate on hypothetical bias (Baillon et al., 2022; Ehmke et al., 2008; Harrison & Rutström, 2008; Matousek et al., 2022; Vranka & Houdek, 2024). We demonstrate that, at least in our sequential PD design, hypothetical choices can reliably mirror incentivized choices with little distortion. Second, we contribute to the growing body of research developing survey-based proxies for preference measures traditionally studied in laboratory settings. Some studies suggest that carefully designed surveys can predict behavior, indicating that self-reported measures may serve as viable substitutes for incentivized tasks in certain contexts (Buser et al., 2024; Fallucchi et al., 2020; Holm & Nystedt, 2008). For example, large-scale surveys correlate with macro-level differences across countries and within countries (Falk et al., 2018). Third, we complement ongoing robustness efforts in behavioral science (e.g., Chapman et al., 2025; Kosfeld et al., 2025) by illustrating how established survey measures — such as reciprocity, altruism, and patience — can vary in their predictive power across settings. Moreover, other studies have tested the stability of measures in predicting individual traits (e.g., Dasgupta et al., 2017; Golsteyn & Schildberg-Hörisch, 2017; Van Landeghem, 2019). Hence, we validate a new survey measure that can be used to examine behavioral correlates of conditional cooperation and highlight both the potential and the limitations of relying on such measures in non-incentivized studies.

The remainder of the paper proceeds as follows. First, we describe our experimental design. Then we present our empirical strategy. Finally, we provide our empirical findings and discuss their implications for the broader literature on cooperation and incentivization.

2. The experiment

The aim of our experiment is twofold. First, we assess when and how unincentivized survey measures can serve as proxies for incentivized measures of conditional cooperation. Second, we compare hypothetical and incentivized conditional cooperation choices and investigate the effects of incentives on behavior. Although a common approach to investigate the latter is using a between-subjects design (e.g., Camerer & Hogarth, 1999; List & Gallet, 2001; List & Shogren, 1998; for a recent multi-context comparison of payment regimes, Brañas-Garza et al., 2023), this would render our prediction exercise with respect to all other survey measures impossible. Hence, in our study we use a within-subjects design that directly links each participant's survey responses to their subsequent incentivized decisions and thereby leverages individual-level variation to assess predictive validity.

We recruited 317 students from the University of Verona via ORSEE (Greiner, 2015) to participate in a two-part study consisting of an online survey followed by a laboratory experiment.[1] Both the online survey and the lab experiment were conducted in Italian and programmed in oTree (Chen et al., 2016). We required all participants to sign up for one of the 15 lab sessions at least 14 days in advance. At that time, we sent them a secure link to the online survey and instructed them to complete it within one week so that each participant completed the survey at least one week before their scheduled lab session. We matched online responses to laboratory decisions via a personal code generated by three identical questions posed at the start of both the online survey and the lab session. In particular, we asked (i) the day of the month in which participants were born, (ii) the first two letters of their mother's name, and (iii) the house number where they live. This allows us to create a 6- or 7-digit unique code to match participants across the two sessions while still preserving full anonymity. We ran all sessions at the Verona Experimental Lab in Economics (VELE) at the University of Verona in November 2024.

We pre-registered the experimental design, hypotheses and analysis on Open Science Framework (OSF) (https://doi.org/10.17605/OSF.IO/59RZX). We received ethical approval for the experiment from the Research Ethics Committee of the Department of Economics at the University of Verona (approval number 0456852 — repository 2549/2024 — approved on 22/10/2024).

Below, we report the experimental measures relevant to our pre-registered research question, that is, identifying the best survey predictors of conditional cooperation. Additional measures included for exploratory purposes, as specified in our pre-registration, are documented in the experimental instructions in Appendix D and analyzed in Appendix C.

[1] We first presented non-incentivized measures to reduce potential bias, as incentivized responses are more likely to influence subsequent non-incentivized ones than vice versa. Moreover, prior studies have found no evidence of order effects in similar settings (Falk et al., 2023).

2.1. Online survey

We designed the online survey to gather a range of hypothetical and unincentivized measures that could predict conditional cooperation in a subsequent incentivized laboratory experiment based on theoretical constructs and previous evidence. In total, the survey contained 89 questions. Three additional questions served to match participants between the two parts of the study, and 17 questions for exploratory purposes elicited environmental beliefs, attitudes and environmental policy preferences (Andre et al., 2024; Dechezleprêtre et al., 2025; Pace et al., 2025). For completeness, the full set of questions is provided in Appendix D. On average, participants spent about 20 min answering the full set of questions. In what follows, we describe the core of the survey consisting of 72 questions, organized into four broad domains.

1. Hypothetical Sequential Prisoner's Dilemma (PD). The first part of the survey introduced participants to a hypothetical sequential PD framed as a fictional scenario rather than a standard payoff matrix as we do in the laboratory experiment. We described two persons who are endowed with €10. Each of them can decide to send the entire sum to the other participant. If €10 are sent to the other participant, these are multiplied by 2.5, resulting in €25 for the recipient. Hence, if both players chose to send their €10, each ended up with €25. Conversely, if one player sent their €10 while the other did not, the sender received €0 and the non-sender €35, and if both kept their €10, they remained with €10. After reading this description, participants answered three questions reflecting first- and second-mover decisions. First, they stated whether they would send their €10 or keep it without observing the other person's choice, thus capturing an unconditional (first-mover) decision. Next, they imagined seeing the other person's choice in advance — whether that person sent or kept their €10 — and indicated how they would respond in each contingency, thereby simulating a second-mover response elicited using the strategy method (Selten, 1967). Although this hypothetical PD retained the same strategic structure as the lab-based game, we introduced it in a distinct framing to reduce potential social desirability bias or "taste for consistency" (Falk & Zimmermann, 2017). Moreover, as the survey was always before the lab experiment and participants did not know the content of the experiment while completing the survey, any of the above effects should be alleviated by the incentives in the lab experiment, making it less likely that participants deliberately replicate their hypothetical decisions when faced with real monetary stakes.

2. Hypothetical social norms. Building on the PD just described, we asked participants to rate the social appropriateness of sending or keeping their €10 as second mover (SM) in the conditional scenarios (d'Adda et al., 2016; Krupka & Weber, 2013). This norm-elicitation component allows us to capture whether perceived appropriateness aligns with participants' choices in the lab, potentially offering initial insights into how social norms might shape conditional cooperation. Models with norm-dependent utility predict that perceived appropriateness enters payoffs and guides behavior; thus, higher injunctive appropriateness of reciprocating (cooperating when the other cooperates, defecting when the other defects) should predict conditional cooperation. Consistent with this, norm ratings are predictive of choices in social dilemmas (Krupka & Weber, 2013) and closely related norm-sensitivity and rule-following measures correlate with prosocial behavior (Kimbrough & Vostroknutov, 2018). In particular, Kölle and Quercia (2021) show that injunctive norms predict conditionally cooperative behavior in a public goods game.

3. Survey measures of economic preferences. To elicit fundamental economics preferences we also added the Global Preference Survey (GPS) module (Falk et al., 2018, 2023), covering time preferences, trust, altruism, and both positive and negative reciprocity. To avoid losing participants' attention due to the length of the survey (Stantcheva, 2023), we decided to have only one measure of risk preferences instead of two as in the original GPS. We included the general risk question, which has been shown to be highly reliable and predictive of risk across many domains (Dohmen et al., 2011). GPS constructs summarize preferences (time, risk, prosociality, reciprocity) that, in standard models of social preferences and belief-dependent cooperation, shape beliefs about others and willingness to bear strategic risk; hence they are expected to covary with conditional cooperation. We also included two survey questions measuring preferences for competition validated by Fallucchi et al. (2020). A stronger taste for competition prioritizes relative payoffs and status, which can crowd out cooperative motives; experiments often find a negative link between competitiveness and cooperation (Buser & Dreber, 2016).

4. Personality traits and attitudes. The final set of variables includes personality traits and additional questions that have been shown to be correlated with conditional cooperation. First, among personality questionnaires, we included the Ten-Item Personality Inventory (I-TIPI) (Gosling et al., 2003) to elicit Big5 personality traits. Agreeableness is theoretically and empirically linked to prosociality (Volk et al., 2011); conscientiousness can increase norm compliance; evidence for extraversion, openness, and neuroticism is mixed but sometimes positive for prosocial behavior (e.g., meta-analytic evidence in Thielmann et al., 2020). We also included 30 items targeting cooperative tendencies of the Behavioral, Emotional, and Social Skills Inventory (BESSI) (Feraco et al., 2024; Soto et al., 2021) as these items directly tap cooperative tendencies and should align with conditional cooperation by design.

Second, we incorporated two trust-related questions from the World Values Survey as reported in Thöni et al. (2012), who find that these two variables correlate with conditional cooperation in the context of a public goods game.

Third, we also added four self-developed items. Three of these concern teamwork and conditional effort. These questions ask whether people in a team context would agree with the following three statements: (i) *I try to do my best regardless of what others do*, (ii) *I try to put in the minimum effort to achieve the maximum result*, and (iii) *I try to do my part only if the others do.* These items extend the notion of contingent cooperation beyond monetary contexts to collaborative work settings, allowing us to assess whether a more general conditional-collaboration mindset correlates with PD behavior. We aggregate these three items into a teamwork index by averaging the responses.

Finally, we elicited a self-reported level of rule following, as rule following has been shown to correlate with cooperative behavior in public goods games (Kimbrough & Vostroknutov, 2018; Kölle et al., 2025).

2.2. Laboratory experiment

Upon arrival at the laboratory, we seated participants at visually isolated computer stations and provided detailed instructions about the experiment. To ensure comprehension, we read the instructions aloud and administered a short comprehension quiz.

Each laboratory session was composed of three parts. In the first part, participants played the sequential one-shot PD that we will use as the main incentivized measure for our pre-registered research question. In Parts 2 and 3, participants completed a norm-elicitation task (Krupka & Weber, 2013), a norm-following task (Kimbrough & Vostroknutov, 2018), and a real effort task with environmental externalities (Pace et al., 2025). These additional measures were pre-registered only to be used as exploratory analyses and hence we do not include them in the main analysis of the paper. However, full details, analysis and instructions can be found in Appendix C and E. Importantly, participants did not know the content of subsequent parts while completing each part.

Each participant was presented with the payoff matrix summarized in Table 1. We designed it such that mutual cooperation yields $R = €15$, unilateral defection gives the defector $T = €19$ and leaves the cooperator with $S = €0$, and mutual defection produces $P = €6$ each. These amounts are different from the ones used in the hypothetical PD in the online survey to minimize the possibility that participants merely replicate their behavior online. However, we kept constant the $EFF = \frac{R-P}{R} = 0.6$ parameter. The latter has been recently shown to be a very good predictor of average levels of cooperation (Gächter et al., 2024).

Table 1
Monetary payoffs in the Prisoner's Dilemma game.

		Player 2	
		Cooperate	Defect
Player 1	*Cooperate*	€15, €15	€ 0, €19
	Defect	€19, €0	€6, €6

Each participant initially made one unconditional decision, i.e., the first mover (FM) choice between cooperate and defect, and then reported their unincentivized belief (0%–100%) about the likelihood that the other player would cooperate. Next, each participant made two conditional decisions, specifying how they would respond if the other player defected and if the other player cooperated (see Kölle et al., 2025, for a similar approach).

Roles were randomly assigned at the end of the session within each pair: one participant's unconditional decision was implemented as the FM choice, while their partner's SM action depended on the corresponding conditional decision tied to the FM's actual move. This random role assignment ensured incentive compatibility, as any participant's unconditional or conditional choice might determine the actual monetary earnings.

On average, sessions lasted approximately 1 h and 10 min, and participants earned approximately 14.50 euros.

2.3. Sample

We determined our sample size by looking at effect sizes in Falk et al. (2023) and Fallucchi et al. (2020), who conduct studies methodologically similar to ours, albeit for different types of preferences. For each type of preference, we correlated each non-incentivized measure and the corresponding incentivized measure and then we averaged across the absolute value(s) of these correlations (see Table I in Appendix A). Among the two studies, Falk et al. (2023) report the lowest domain-specific averages and hence we relied on this study to be able to detect a lower effect size. Across their six preference measures, Falk et al. (2023) find an overall mean correlation of $r = 0.209$. To minimally detect this correlation with $\alpha = 0.05$ and $(1 - \beta) = 0.90$, we need a sample size of 232 participants. We aimed at around 300 signed-up participants to account for no shows, attrition, and potential code mismatches, and still be able to reach 232 participants.

A total of 317 participants initially signed up for our laboratory experiment; 292 attended and completed the experimental sessions. We excluded 7 subjects who entered mismatched codes across online survey and lab session (presumably typographical errors) and 26 who failed to finish the online survey, yielding a final sample of 259 participants (see Table II in Appendix A for attrition details). Importantly, comparisons between those who dropped out and those who remained in the study revealed no significant differences across any demographic characteristics, indicating that attrition did not introduce systematic bias.

Our final sample is composed of 70.3% female; the average age is 21.8, with 67.8% aged 19–22. Most of the participants are enrolled in Bachelor's degrees (71.4%) and come from various study areas. Prior work often reports behavioral differences for economics students in social dilemma and allocation tasks, while some studies find opposite evidence (e.g., Carter & Irons, 1991; Frank et al., 1993; Gerlach, 2017; Yezer et al., 1996); our composition, where only 10.4% are studying economics, reduces the concern that such differences may drive our results. A complete breakdown of the summary statistics is reported in Table III in Appendix A.

3. Results

As pre-registered, our empirical analysis proceeds in three steps. First, we characterize each participant's behavior in the sequential Prisoner's Dilemma (PD) by assigning them to one of four types — Conditional Cooperator (CC), Free Rider (FR), Full Cooperator (FC), or Other (OT) — based on their conditional decisions in PD. We apply this characterization in both the online survey, where the PD is hypothetical, and the laboratory setting.

Second, we create three binary dependent variables to highlight different possible classifications using these four types (see Section 3.1 below). Each classification captures a distinct perspective on cooperative behavior, from distinguishing conditional cooperators from everyone else to isolating cooperative or defecting strategies.

Third, we investigate which survey responses best predict our conditional cooperation classification found in the second stage using OLS regressions. To reduce the dimensionality of the survey data, we first aggregate items according to existing questionnaire sub-scales, producing a set of composite variables. We then supplement this approach with an item-level examination, employing the Least Absolute Shrinkage and Selection Operator (LASSO) (Tibshirani, 2018) for selecting the best predictors of conditional cooperation and OLS to identify individual survey questions that most strongly predict our classifications of CC. By combining these aggregated and item-level analyses, we balance interpretability against the need to detect predictors of conditional cooperators.

3.1. Behavioral types and hypothetical versus incentivized behavior

As a first step of our analysis, we characterize each participant according to their decisions in a sequential PD. This characterization takes place in both the hypothetical online survey and the incentivized laboratory experiment, enabling us to assess whether stated cooperative intentions align with actual behavior under real monetary stakes. We assign individuals to one of four distinct behavioral types based on their conditional decisions:

- **CC:** Cooperate if the other player cooperates and defects otherwise.
- **FR:** Defect regardless of the other player's choice.
- **FC:** Cooperate regardless of the other player's choice.
- **OT:** Defect when the other player cooperates; cooperate when the other player defects.

Table 2 – panel (a) presents summary statistics on the distribution of types, comparing their prevalence in the online survey versus the laboratory experiment. We observe that, in the hypothetical setting, 82% of participants are characterized as CC, 8% as FR, 8% as FC, and 2% as OT. In contrast, the incentivized lab data reveal a slightly lower prevalence of CC (78%) and a corresponding increase of FR (19%). FC and OT become comparatively rare in the lab (2% and 1%, respectively). While the two distributions differ ($\chi^2(3) = 20.352, p < 0.001$), the raw percentages still indicate substantial stability for most participants—particularly for CC.

The contingency table in Table 2 – panel (b) shows that, out of 259 participants, 187 have a consistent classification in the Survey and the Lab. The most prevalent changes involve moving from Survey–CC to Lab–FR, which occurs in 34 cases, and from Survey–FC to Lab–CC, which takes place in 12 cases. Fewer instances are observed for the reverse directions, with counts of 11 and 3, respectively.

To quantify individual-level alignment, we report percent agreement and Cohen's κ and interpret κ using Landis–Koch benchmarks—slight (0–0.20), fair (0.21–0.40), moderate (0.41–0.60), substantial (0.61–0.80), almost perfect (0.81–1.00) (Landis & Koch, 1977). Alignment is 72.2% (187/259) with $\kappa = 0.18$ (95% CI [0.07, 0.29]), which is "slight" on this scale (CI spanning "slight" to "fair"). For the binary contrast CC vs. non-CC, $\kappa = 0.22$ (fair); other contrasts yield $\kappa = 0.14$ for FR vs. non-FR (slight), $\kappa = 0.12$ for FC vs. non-FC (slight), and $\kappa = 0.28$ for OT vs. non-OT (fair).

Directional switching is not symmetric: Bowker's test (Bowker, 1948) rejects symmetry ($\chi^2 = 22.73$, df = 5, $p < 0.001$). The largest asymmetry is between CC and FR (Survey–CC to Lab–FR = 34 vs. Survey–FR to Lab–CC = 11), with additional imbalances for FC vs. CC (12 vs. 3) and FR vs. FC (6 vs. 1). Two-by-two McNemar tests indicate that FR increases in the lab ($p < 0.001$) and FC decreases ($p = 0.003$), while the change in CC is not statistically significant ($p = 0.17$). Table IV in Appendix A reports the off-diagonal flow matrix with pairwise Bowker contributions; Figure 1 in Appendix B visualizes the main transitions. Overall, the distributional differences and asymmetric switches suggest some hypothetical bias; nevertheless, the 72% exact agreement and the fair CC vs. non-CC κ indicate that the hypothetical PD retains individual-level predictive value.

Table 2
Comparison of lab and survey.

(a) Frequency of types

Types	Survey	Lab
CC	0.826	0.780
FR	0.077	0.185
FC	0.081	0.023
OT	0.016	0.012

(b) Contingency table of types

		Survey			
	Types	CC	FR	FC	OT
Lab	CC	176	11	12	3
	FR	34	8	6	0
	FC	3	1	2	0
	OT	1	0	1	1

After characterizing participants into one of four behavioral types (CC, FR, FC, or OT), in the second step, we define a set of binary dependent variables designed to capture different dimensions of cooperative behavior. In particular, following our pre-registration, we construct three binary classifications.

The first classification (CC_1) distinguishes CC from the aggregate of all other types (FR, FC, and OT). By isolating CC in this way, we can directly test which factors best predict a willingness to conditionally cooperate, relative to any other strategy.

The second classification (CC_2) compares CC exclusively with FR, thereby sharpening the critical contrast between individuals who reciprocate cooperation and those who always defect. We think that excluding some participants is not particularly problematic as CC and FR are the most numerous groups of types in our characterization and also in the literature (Thöni & Volk, 2018). By applying this classification, we lose only 9 observations (FC and OT), merely 3% of our data.

Finally, as the third classification (CC_3), we group CC and FC and compare them against FR and OT. This broader classification contrasts predominantly cooperative behavior — whether conditional or unconditional — with full defection or reversed-pattern choices.

3.2. Best predictors of conditional cooperation using aggregated survey measures

Table 3 details the variables used in our regression analysis in this section, along with their descriptions. The dependent variables in our analysis are the three different classifications of CC. The independent variables include Hypothetical CC, which classifies cooperative behavior under hypothetical conditions following the classification above, hypothetical social norms regarding the PD, Big5 personality traits (Extraversion, Agreeableness, Conscientiousness, Emotional Stability, and Openness), Cooperation Skills as elicited using BESSI, economic preferences from GPS (Risk, Altruism, Trust, Positive and Negative Reciprocity, Time discounting), Teamwork, and Rule Following.

Table 3
Summary of regression variables and their explanation.

Variable	Explanation
`Hypothetical CC`	Characterization of CC behavior based on hypothetical PD. = 1 if player cooperates if the other cooperates, and defects if the other defects: Q5, Q6.
`Extraversion`	Personality trait reflecting sociability and enthusiasm. Aggregated following Gosling et al. (2003): Q62, Q67.
`Agreeableness`	Personality trait indicating friendliness and cooperation. Aggregated following Gosling et al. (2003): Q63, Q68.
`Conscientiousness`	Personality trait related to organization and diligence. Aggregated following Gosling et al. (2003): Q64, Q69.
`Emotional stability`	Personality trait measuring emotional resilience. Aggregated following Gosling et al. (2003): Q65, Q70.
`Openness`	Personality trait related to creativity and curiosity. Aggregated following Gosling et al. (2003): Q66, Q71.
`Cooperation Skills`	Measure of an individual's ability to cooperate in teams. Aggregated following Soto et al. (2021): from Q32 to Q61.
`Risk`	Individual's willingness to take risks. Single item: Q16.
`Altruism`	Tendency to act in a selfless way to benefit others. Aggregated following Falk et al. (2023): Q20, Q27.
`Positive reciprocity`	Tendency to reciprocate kindness and positive actions. Aggregated following Falk et al. (2023): Q21, Q26.
`Time discounting`	Measure of an individual's patience and time preference. Aggregated following Falk et al. (2023): from Q11 to Q15, Q17.
`Negative reciprocity`	Tendency to retaliate against perceived unfairness. Aggregated following Falk et al. (2023): Q18, Q19, Q22.
`Trust`	Belief in the reliability and honesty of others. Single item: Q23.
`Appropriateness of C on C`	Norm indicating appropriateness of cooperation if the other cooperates. Single item: Q7.
`Appropriateness of D on C`	Norm indicating appropriateness of defection if the other cooperates. Single item: Q8.
`Appropriateness of C on D`	Norm indicating appropriateness of cooperation if the other defects. Single item: Q9.
`Appropriateness of D on D`	Norm indicating appropriateness of defection if the other defects. Single item: Q10.
`Rule Following`	Tendency to follow established rules and regulations. Single item: Q28.
`Teamwork`	Teamwork and conditional-collaboration mindset. We average the answers to three questions: Q29, Q30, Q31.

Note: See Table V in Appendix A for the questions' numbering.

Table 4
OLS regression for three CC classifications.

	CC_1	CC_2	CC_3
Constant	0.489	0.613	0.755
	(0.422)	(0.417)	(0.410)
Hypothetical CC	0.209**	0.151*	0.167*
	(0.073)	(0.074)	(0.071)
Extraversion	0.002	0.001	−0.003
	(0.020)	(0.020)	(0.020)
Agreeableness	0.011	0.014	0.016
	(0.028)	(0.028)	(0.027)
Conscientiousness	−0.003	−0.006	−0.012
	(0.026)	(0.026)	(0.026)
Emotional stability	0.013	0.003	0.005
	(0.022)	(0.021)	(0.021)
Openness	0.004	0.008	0.009
	(0.028)	(0.028)	(0.027)
Cooperation skills	−0.055	−0.085	−0.063
	(0.080)	(0.080)	(0.078)
Risk	0.072	0.071	0.065
	(0.063)	(0.062)	(0.061)
Altruism	−0.001	−0.0003	−0.0005
	(0.001)	(0.001)	(0.001)
Positive reciprocity	0.033	0.027	0.019
	(0.053)	(0.052)	(0.052)
Time discounting	−0.179*	−0.164*	−0.163*
	(0.071)	(0.069)	(0.068)
Negative reciprocity	−0.037	−0.009	−0.011
	(0.059)	(0.058)	(0.057)
Trust	0.052	0.052	0.053
	(0.043)	(0.043)	(0.042)
Appropriateness of C on C	0.013	0.032	0.027
	(0.053)	(0.052)	(0.051)
Appropriateness of D on C	−0.046	−0.028	−0.031
	(0.035)	(0.035)	(0.034)
Appropriateness of C on D	0.004	0.021	0.011
	(0.037)	(0.037)	(0.036)
Appropriateness of D on D	0.019	0.016	0.009
	(0.040)	(0.040)	(0.039)
Rule following	0.014	0.005	−0.007
	(0.018)	(0.018)	(0.017)
Teamwork	0.019	0.012	0.011
	(0.028)	(0.028)	(0.028)
Observations	259	250	259
R^2	0.073	0.068	0.066
Adjusted R^2	0.023	0.017	0.021
Residual Std. Error	0.394 (df = 239)	0.391 (df = 230)	0.394 (df = 239)
F Statistic	1.473 (df = 19; 239)	1.334 (df = 19; 230)	1.451 (df = 19; 239)

Note: *p < 0.05; **p < 0.01; ***p < 0.001.

In Table 4 we present three OLS models, one for each classification of CC, using the full set of regressors described above. In all our regressions, we will consider a variable to be a consistent predictor if it predicts conditional cooperation across the three classification criteria.

Across columns (1)–(3), Hypothetical CC consistently exhibits a positive and statistically significant effect at the 1% or 5% level. This pattern suggests that individuals reporting being conditional cooperators in the survey are indeed more likely to exhibit corresponding cooperative behavior in the lab. This is not surprising given the very high rate of consistency discussed in Section 3.1. Patience (Time discounting) shows a negative and significant coefficient at the 5% level in each specification, indicating that individuals who place greater value on immediate rewards tend to be more likely conditional cooperators during the lab experiment.

None of the other aggregated constructs, such as personality traits or Cooperation Skills, reach conventional levels of significance in these specifications. Risk, Altruism, Positive Reciprocity, Negative Reciprocity, and Trust likewise display no robust relationships with CC in this particular model setup. Overall, the adjusted R^2 values (ranging from 0.017 to 0.023) remain modest, underscoring that, although Hypothetical CC and Time Discounting exhibit meaningful associations, the combined set of aggregated variables explains a very small share of the variation in lab-based conditional cooperation outcomes.

A plausible reason why only a few aggregate constructs reach significance in the regressions is the high dimensionality of the specification. Several aggregated constructs are conceptually related and potentially correlated with Hypothetical CC; in that case, shared variance would be absorbed by the strongest predictors, inflating standard errors, and attenuating coefficients on the others. To assess this possibility, Figure 2 in Appendix B reports pairwise Pearson correlations between Hypothetical CC and each aggregated

construct. Hypothetical CC is significantly correlated with all normative judgments (appropriateness of cooperating on cooperation and appropriateness of defecting on cooperation ($p < 0.001$), appropriateness of defecting on defection ($p < 0.01$), appropriateness of cooperating on defection ($p < 0.05$) and Teamwork ($p < 0.05$)).[2]

Overall, we observe that the strongest correlation in terms of magnitude and significance is the correlation between Hypothetical CC and incentivized CC. To contextualize magnitudes, we look at recent papers conducting similar exercises on different preference measures. Falk et al. (2023) report out-of-sample correlations between survey indices and the corresponding incentivized elicitations ranging from approximately $r \approx 0.26$ (trust) to $r \approx 0.59$ (time discounting), with intermediate values for risk (≈0.29) and positive reciprocity (≈0.44). Regarding the preference for competition, Fallucchi et al. (2020) validate a one-item survey measure that correlates $r \approx 0.26$–0.28 with incentivized willingness to compete. More recently, Chapman et al. (2025) reexamine qualitative self-assessments across large and diverse samples and report weaker survey-task links on average (roughly two-thirds the size of the initial (Falk et al., 2023) correlations). Taken together, these references suggest that positive but modest survey-task relationships are common; our estimates (e.g., $r = 0.224$ for Hypothetical CC vs. CC_1) fall within this range.

3.3. Best predictors of conditional cooperation using individual survey items

We now move to the analysis of the individual items from the survey. We have 66 items that we regress each CC classification on. Table VII in Appendix A displays OLS regressions that include all 66 items simultaneously. The raw R^2 values in these full models appear high (0.273–0.284), but the adjusted R^2 is markedly low (0.023–0.034), reflecting the large penalty for including so many additional predictors, which are often collinear. This pronounced difference between R^2 and adjusted R^2 underlines the potential overfitting and instability of the full specification. The F-statistic also suggests that, although some individual variables meet significance criteria, the combined predictive power of all 66 items is diluted by a host of insignificant or correlated predictors.

To address this issue and in line with our pre-registration, we use the linear Least Absolute Shrinkage and Selection Operator (LASSO) approach. This method is well suited for analyzing large sets of potential predictors, as it imposes an L_1 penalty that shrinks certain coefficients to zero, thereby performing both estimation and variable selection. In our study, we initially considered 66 predictors (derived from survey items). Given the high collinearity observed among certain variables, LASSO eliminated multiple correlated predictors. To address this issue, all variables were standardized prior to estimation to ensure comparability of effect sizes and penalty terms.

We determine the optimal level of regularization by performing cross-validation over a range of penalty values. In the literature, three penalty levels (λ) are commonly highlighted: $\lambda_{\min}$, the value of λ that minimizes the cross-validation error; λ_{1se}, the largest λ whose error is within one standard error of the minimum; and $\lambda_{1se}/2$, an intermediate value. Table 5 reports the results of our LASSO analyses under two penalty levels (see Notes to the table for the rationale).

Table 5
Selected items using linear LASSO.

Classification	LASSO $\lambda_{\min}$	LASSO $\lambda_{1se}/2$
CC_1	Hypothetical CC	Hypothetical CC; Appropriateness D on C; GPS—Patience; GPS—Gift choice
CC_2	Hypothetical CC	Hypothetical CC; Appropriateness C on C; Appropriateness D on C; BESSI—item 125, BESSI—item 142; GPS—Patience, GPS—Favor; GPS—Trust, GPS—Gift choice
CC_3	Hypothetical CC	Hypothetical CC; Appropriateness D on C; BESSI—item 142; GPS—Patience; GPS—Favor; GPS—Gift choice

Notes: The values of λ are not fixed but are adaptively selected by the LASSO model through cross-validation. Each dependent variable may yield a different optimal λ due to differences in variance, predictive relationships, and model structure. In our application, λ_{1se} (the largest λ whose cross-validation error lies within one standard error of the minimum) is overly restrictive for all variables except $Hypothetical\ CC$ for CC_2. We focus on $\lambda_{\min}$ (the λ that minimizes the cross-validation error) and the less restrictive compromise $\lambda_{1se}/2$. Slight variations in λ may occur due to the randomness in cross-validation splits and data preprocessing.

Under $\lambda_{\min}$, the strongest predictor of conditional cooperation across all categorizations was the Hypothetical CC classification. However, when the regularization was relaxed to $\lambda_{1se}/2$, additional variables emerged in each categorization.

[2] As a complementary (and non-preregistered) analysis, we also estimate bivariate OLS regressions for each specification, regressing each CC classification on one aggregated construct at a time (see Table VI in Appendix A). Consistent with the multivariate models, Hypothetical CC and Patience are significant predictors of incentivized choices. Beyond these, only the normative rating of the appropriateness of defecting on cooperation is significantly associated with lower conditional cooperation only in the CC_1 specification. This direction is in accordance with our theoretical rationale for including norms: in models with norm-dependent utility, perceived appropriateness enters payoffs and guides behavior; thus, rating defection against a cooperator as more appropriate should predict less conditional cooperation.

Across the three classifications of CC, four items seem to consistently be good predictors: Hypothetical CC, appropriateness of defecting on cooperation, and two GPS survey items, patience (Q17) and gift choice (Q26).

Other variables are uniquely selected in specific comparisons. Specifically, in classification CC_1, the predictors identified at $\lambda_{1se}/2$ were: appropriateness of defecting on cooperation, a measure for patience (GPS—Patience), a measure for generosity (GPS—Gift choice). In the classification CC_2, the variables additionally selected at $\lambda_{1se}/2$ included: appropriateness of defecting on cooperation and of cooperating on cooperation, GPS—Patience, a measure for generosity (GPS—Gift choice), a measure for positive reciprocity (GPS—Favor), a measure of trust (GPS—Trust), two items from BESSI, items 142 and 125. Finally, for the classification CC_3, the relaxed penalty also retained the same set of variables as for CC_2, excluding BESSI, item 125, appropriateness of cooperating on cooperation, and GPS—Trust.

Following these LASSO analyses, we refit unpenalized OLS using only the variables identified as relevant predictors for each classification. To obtain valid post-selection inference for the refit OLS coefficients, we apply the repeated half-sample splitting procedure in the spirit of Meinshausen et al. (2009). In each random half-split, we use one-half of the data to run LASSO (with $\lambda = \lambda_{1se}/2$ as in the main analysis; any tuning occurs within the selection half), and we use the held-out half to refit OLS on the variables selected in that split and compute classical p-values. Repeating this over many random half-splits yields a collection of split-specific p-values for each coefficient, which we then aggregate across splits using Tippett's min-p method (Tippett, 1931). This procedure yields valid p-values for post-selection testing even when λ is tuned adaptively. Intuitively, validity holds because selection and inference are separated in every split, and the aggregation step recombines evidence across splits while preserving error control. Implementation details appear in Table VIII in Appendix A.

Table 6 reports OLS regressions estimated after applying LASSO for variable selection; stars correspond to post-selection p-values, which are reported in Table VIII in Appendix A. The adjusted R^2 values range from 0.076 to 0.106, and the F-tests are highly significant ($p < 0.001$), indicating that a small set of predictors — most notably Hypothetical CC and positive reciprocity items (gift choice and favor) — capture the bulk of the survey–lab relationship while avoiding overfit.

Overall, the post-LASSO refits in Table 6 isolate a small set of predictors that remain statistically robust under post-selection inference. Hypothetical CC is the strongest predictor, with a positive coefficient that is statistically significant in all specifications, indicating that being a conditional cooperator in the survey predicts conditional cooperation in the lab. GPS—Gift choice is positive and significant across all three models, and GPS—Favor is significant in CC_2 and CC_3 specifications. By contrast, Patience and the norm of defecting on cooperation move in the theoretically expected directions but are not consistently significant once we account for selection. Taken together, these patterns suggest that the hypothetical PD carries most of the survey-based signal for incentivized CC, with a small number of complementary items adding incremental predictive content.

Table 6
OLS regressions post LASSO for three CC classifications.

	CC_1	CC_2	CC_3
Constant	0.771***	1.281***	1.498***
	(0.148)	(0.298)	(0.241)
Hypothetical CC	0.211***	0.127*	0.167**
	(0.067)	(0.069)	(0.065)
Appropriateness D on C (Q8)	−0.050	−0.042	−0.052
	(0.032)	(0.032)	(0.031)
GPS – Patience (Q17)	−0.032	−0.025	−0.025
	(0.014)	(0.013)	(0.013)
GPS – Gift choice (Q26)	0.033**	0.039**	0.038**
	(0.016)	(0.016)	(0.016)
GPS – Favor (Q21)		−0.069**	−0.058**
		(0.023)	(0.022)
BESSI – item 142 (Q53)		−0.061	−0.056
		(0.029)	(0.028)
GPS – Trust (Q23)		0.017	
		(0.010)	
BESSI – item 125 (Q50)		0.035	
		(0.026)	
Appropriateness C on C (Q7)		0.039	
		(0.047)	
Observations	259	250	259
R^2	0.090	0.138	0.115
Adjusted R^2	0.076	0.106	0.093
Residual Std. Error	0.399 (df = 254)	0.373 (df = 240)	0.379 (df = 252)
F Statistic	6.316*** (df = 4; 254)	4.274*** (df = 9; 240)	5.435*** (df = 6; 252)

Notes: Coefficients and SEs are from unpenalized OLS refit on the subset of variables selected by LASSO. Stars correspond to post- selection p-values from repeated half-sample splitting reported in Table VIII, Appendix A. $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

4. Summary and conclusions

Our study investigates the ability of survey-based measures to predict conditional cooperation in incentivized settings. We find that, among the survey measures, the hypothetical PD consistently emerges as the strongest predictor, reaching a significance of at least the 5% level in all our specification models. The predictive power of other measures, such as patience or normative evaluations, lacks robustness across different model specifications, which suggests that their interpretation as potential predictors should be cautious.

We also find some, though not pervasive, evidence of hypothetical bias. In particular, more than two-thirds of the participants remain in the same behavioral type across hypothetical and real-stakes contexts, suggesting substantial consistency. However, we also see that a small fraction switches from conditional cooperation to free-riding when payoffs become real—a pattern consistent with stated-preference evidence of a positive but imperfect link between hypothetical and actual behavior (for recent reviews, see Haghani et al., 2021a, 2021b).

This duality explains why the raw distributions of types in the survey and the lab differ, yet regression analyses still reveal a positive, statistically significant association ($\alpha = 0.05$) at the individual level between hypothetical and incentivized cooperation. Most participants behave similarly across both settings, driving the robust association in our models, while fewer participants who switch strategies generate the shift in the overall distribution of types.

Interestingly, patience is negatively associated with conditional cooperation. One possible explanation is linked to the concept of intuitive cooperation (Rand et al., 2014), which suggests that cooperation often emerges as a fast, automatic response, while deliberation may favor more self-interested choices. Since impatience may be associated with a greater reliance on intuitive decision-making, impatient individuals may be more likely to cooperate instinctively, without engaging in strategic deliberation. In contrast, more patient individuals, who tend to think more deliberatively, might override intuitive cooperative tendencies in favor of self-interested behavior. However, this explanation should be taken cautiously, as some of the intuitive cooperation results are shown not to be robust (see Kvarven et al., 2020). In a closely related setting, Espín et al. (2012) investigate whether punishment in one-shot cooperation games is implemented out of a deliberate action or out of visceral impulses and show that cooperators are more likely to punish if they are patient (indicating deliberation) and free-riders are more likely to punish if they are impatient (indicating intuition).

Beyond hypothetical gameplay, we find limited predictive power for other factors—such as personality traits, social preferences, and normative evaluations of appropriateness within the same PD game. These results are broadly consistent with recent studies (e.g., Chapman et al., 2025; Kosfeld et al., 2025), which indicate that the predictive relationships of established survey constructs do not always generalize robustly across different contexts.

Future work should examine the robustness of these patterns across different social dilemmas, population samples, and cultural contexts. Given the ongoing shift toward large-scale online data collection, high-powered studies using diverse participant pools may further elucidate how hypothetical tasks and alternative survey measures perform under varying conditions. Such efforts can help refine the design and implementation of survey-based proxies, ultimately enhancing their reliability as tools for understanding cooperative behavior.

Finally, if the research objective shifts from prediction to mitigation of hypothetical bias, the literature on stated preference offers simple, low-cost approaches that can be adapted to our design. In our experimental setting, options include: (i) cheap talk—a brief preamble that warns about hypothetical overstatement and asks participants to answer as if choices had real consequences (Cummings & Taylor, 1999); (ii) a consequentiality statement—one or two lines indicating that responses inform subsequent tasks or decisions, increasing perceived payoff relevance and effort (Carson & Groves, 2007; Vossler et al., 2012); (iii) an ex-ante honesty oath—a short pledge to answer truthfully and carefully (Jacquemet et al., 2013); and (iv) an ex-post certainty scale—a 0–10 confidence rating used to calibrate or down-weight low-certainty answers (Champ et al., 1997). Future implementations can assess whether these approaches not only change mean responses, but also improve out-of-sample predictive accuracy and calibration of hypothetical PD relative to incentivized PD.

Acknowledgments

This work was supported by the Italian Ministry of Universities and Research [PRIN PNRR 2022 grant “Conditional cooperation, social norms, and sustainable behavior”, project code: P2022JASLC]. The replication materials for the study are available at https://osf.io/ah5e8/. We thank the Editors and two anonymous referees for helpful comments.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.joep.2026.102901.

References

Allcott, H. (2011). Social norms and energy conservation. *Journal of Public Economics*, *95*(9), 1082–1095, Special Issue: The Role of Firms in Tax Systems.

Andre, P., Boneva, T., Chopra, F., & Falk, A. (2024). Misperceived social norms and willingness to act against climate change. *The Review of Economics and Statistics*, 1–46.

Baillon, A., Bleichrodt, H., & Granic, G. D. (2022). Incentives in surveys. *Journal of Economic Psychology*, *93*, Article 102552.

Bowker, A. H. (1948). A test of symmetry in contingency tables. *Journal of the American Statistical Association*, *43*(244), 572–574.

Brañas-Garza, P., Jorrat, D., Espín, A. M., et al. (2023). Paid and hypothetical time preferences are the same: lab, field and online evidence. *Experimental Economics*, *26*(2), 412–434.

Buser, T., & Dreber, A. (2016). The flipside of comparative payment schemes. *Management Science*, *62*(9), 2626–2638.

Buser, T., Niederle, M., & Oosterbeek, H. (2024). Can competitiveness predict education and labor market outcomes? Evidence from incentivized choice and survey measures. *The Review of Economics and Statistics*, 1–45.

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*(1–3), 7–42.

Carson, R. T., & Groves, T. (2007). Incentive and informational properties of preference questions. *Environmental and Resource Economics*, *37*(1), 181–210.

Carter, J. R., & Irons, M. D. (1991). Are economists different, and if so, why? *Journal of Economic Perspectives*, *5*(2), 171–177.

Champ, P. A., Bishop, R. C., Brown, T. C., & McCollum, D. W. (1997). Using donation mechanisms to value nonuse benefits from public goods. *Journal of Environmental Economics and Management*, *33*(2), 151–162.

Chapman, J., Ortoleva, P., Snowberg, E., Yariv, L., & Camerer, C. F. (2025). *Reassessing qualitative self-assessments and experimental validation*: *Working paper w33520*, National Bureau of Economic Research.

Chen, D. L., Schonger, M., & Wickens, C. (2016). OTree—An open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, *9*, 88–97.

Cummings, R. G., & Taylor, L. O. (1999). Unbiased value estimates for environmental goods: A cheap talk design for the contingent valuation method. *American Economic Review*, *89*(3), 649–665.

d'Adda, G., Drouvelis, M., & Nosenzo, D. (2016). Norm elicitation in within-subject designs: Testing for order effects. *Journal of Behavioral and Experimental Economics*, *62*, 1–7.

Dasgupta, U., Gangadharan, L., Maitra, P., & Mani, S. (2017). Searching for preference stability in a state dependent world. *Journal of Economic Psychology*, *62*, 17–32.

Dechezleprêtre, A., Fabre, A., Kruse, T., Planterose, B., Sanchez Chico, A., & Stantcheva, S. (2025). Fighting climate change: International attitudes toward climate policies. *American Economic Review*, *115*(4), 1258–1300.

Dohmen, T., Falk, A., Huffman, D., Sunde, U., Schupp, J., & Wagner, G. G. (2011). Individual risk attitudes: Measurement, determinants, and behavioral consequences. *Journal of the European Economic Association*, *9*(3), 522–550.

Ehmke, M. D., Lusk, J. L., & List, J. A. (2008). Is hypothetical bias a universal phenomenon? A multinational investigation. *Land Economics*, *84*(3), 489–500.

Eichenseer, M., & Moser, J. (2020). Conditional cooperation: Type stability across games. *Economics Letters*, *188*, Article 108941.

Espín, A. M., Brañas-Garza, P., Herrmann, B., & Gamella, J. F. (2012). Patient and impatient punishers of free-riders. *Proceedings of the Royal Society B: Biological Sciences*, *279*(1749), 4923–4928.

Falk, A., Becker, A., Dohmen, T., Enke, B., Huffman, D., & Sunde, U. (2018). Global evidence on economic preferences*. *The Quarterly Journal of Economics*, *133*(4), 1645–1692.

Falk, A., Becker, A., Dohmen, T., Huffman, D., & Sunde, U. (2023). The preference survey module: A validated instrument for measuring risk, time, and social preferences. *Management Science*, *69*(4), 1935–1950.

Falk, A., & Zimmermann, F. (2017). Consistency as a signal of skills. *Management Science*, *63*(7), 2197–2210.

Fallucchi, F., Luccasen III, R. A., & Turocy, T. L. (2019). Identifying discrete behavioural types: a re-analysis of public goods game contributions by hierarchical clustering. *Journal of the Economic Science Association*, *5*(2), 238–254.

Fallucchi, F., Nosenzo, D., & Reuben, E. (2020). Measuring preferences for competition with experimentally-validated survey questions. *Journal of Economic Behavior and Organization*, *178*, 402–423.

Fehr-Duda, H., & Fehr, E. (2016). Sustainability: Game human nature. *Nature*, *530*(7591), 413–415.

Feraco, T., Casali, N., Pellegrino, G., et al. (2024). The Italian behavioral, emotional, and social skills inventory (BESSI-I). *Journal of Personality Assessment*, *106*(6), 750–764.

Fischbacher, U., & Gächter, S. (2010). Social preferences, beliefs, and the dynamics of free riding in public goods experiments. *American Economic Review*, *100*(1), 541–556.

Fischbacher, U., Gächter, S., & Fehr, E. (2001). Are people conditionally cooperative? Evidence from a public goods experiment. *Economics Letters*, *71*(3), 397–404.

Fischbacher, U., Gächter, S., & Quercia, S. (2012). The behavioral validity of the strategy method in public good experiments. *Journal of Economic Psychology*, *33*(4), 897–913.

Frank, R. H., Gilovich, T., & Regan, D. T. (1993). Does studying economics inhibit cooperation? *Journal of Economic Perspectives*, *7*(2), 159–171.

Frey, B. S., & Meier, S. (2004). Social comparisons and pro-social behavior: Testing "conditional cooperation" in a field experiment. *American Economic Review*, *94*(5), 1717–1722.

Gächter, S., Lee, K., Sefton, M., & Weber, T. O. (2024). The role of payoff parameters for cooperation in the one-shot Prisoner's Dilemma. *European Economic Review*, *166*, Article 104753.

Gerlach, P. (2017). The games economists play: Why economics students behave more selfishly than other students. *PLoS One*, *12*(8), Article e0183814.

Golsteyn, B., & Schildberg-Hörisch, H. (2017). Challenges in research on preferences and personality traits: Measurement, stability, and inference. *Journal of Economic Psychology*, *60*, 1–6.

Gosling, S. D., Rentfrow, P. J., & Swann, J. (2003). A very brief measure of the big five personality domains. *Journal of Research in Personality*, *37*, 504–528.

Greiner, B. (2015). Subject pool recruitment procedures: organizing experiments with ORSEE. *Journal of the Economic Science Association*, *1*, 114–125.

Haghani, M., Bliemer, M. C., Rose, J. M., Oppewal, H., & Lancsar, E. (2021a). Hypothetical bias in stated choice experiments: Part I. Macro-scale analysis of literature and integrative synthesis of empirical evidence from applied economics, experimental psychology and neuroimaging. *Journal of Choice Modelling*, *41*, Article 100309.

Haghani, M., Bliemer, M. C., Rose, J. M., Oppewal, H., & Lancsar, E. (2021b). Hypothetical bias in stated choice experiments: Part II. conceptualisation of external validity, sources and explanations of bias and effectiveness of mitigation methods. *Journal of Choice Modelling*, *41*, Article 100322.

Harrison, G. W., & Rutström, E. E. (2008). Experimental evidence on the existence of hypothetical bias in value elicitation methods. *Handbook of Experimental Economics Results*, *1*, 752–767.

Hertwig, R., & Ortmann, A. (2001). Experimental practices in economics: A methodological challenge for psychologists? *Behavioral and Brain Sciences*, *24*(3), 383–403.

Holm, H., & Nystedt, P. (2008). Trust in surveys and games – a methodological contribution on the influence of money and location. *Journal of Economic Psychology*, *29*(4), 522–542.

Jacquemet, N., Joule, R.-V., Luchini, S., & Shogren, J. F. (2013). Preference elicitation under oath. *Journal of Environmental Economics and Management*, *65*(1), 110–132.
Kimbrough, E. O., & Vostroknutov, A. (2018). A portable method of eliciting respect for social norms. *Economics Letters*, *168*, 147–150.
Kölle, F., & Quercia, S. (2021). The influence of empirical and normative expectations on cooperation. *Journal of Economic Behavior and Organization*, *190*, 691–703.
Kölle, F., Quercia, S., & Tripodi, E. (2025). Social preferences under the shadow of the future. *Experimental Economics*, 1–21.
Kosfeld, M., Sharafi, Z., Sontag González, M., & Zou, N. (2025). *Measuring economic preferences with surveys and behavioral experiments*: *Technical report 11631*, CESifo Working Paper Series.
Krupka, E. L., & Weber, R. A. (2013). Identifying social norms using coordination games: Why does dictator game sharing vary? *Journal of the European Economic Association*, *11*(3), 495–524.
Kvarven, A., Strømland, E., Wollbrant, C., Andersson, D., Johannesson, M., Tinghög, G., Västfjäll, D., & Myrseth, K. O. R. (2020). The intuitive cooperation hypothesis revisited: a meta-analytic examination of effect size and between-study heterogeneity. *Journal of the Economic Science Association*, *6*, 26–42.
Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174.
List, J. A., & Gallet, C. A. (2001). What experimental protocol influence disparities between actual and hypothetical stated values? *Environmental and Resource Economics*, *20*(3), 241–254.
List, J. A., & Shogren, J. F. (1998). Calibration of the difference between actual and hypothetical valuations in a field experiment. *Journal of Economic Behavior and Organization*, *37*, 193–205.
Matousek, J., Havranek, T., & Irsova, Z. (2022). Individual discount rates: a meta-analysis of experimental evidence. *Experimental Economics*, *25*(1), 318–358.
Meinshausen, N., Meier, L., & Bühlmann, P. (2009). P-Values for High-Dimensional regression. *Journal of the American Statistical Association*, *104*(488), 1671–1681.
Pace, D. D., Imai, T., Schwardmann, P., & van der Weele, J. J. (2025). Uncertainty about carbon impact and the willingness to avoid CO2 emissions. *Ecological Economics*, *227*, Article 108401.
Rand, D. G., Peysakhovich, A., Kraft-Todd, G. T., Newman, G. E., Wurzbacher, O., Nowak, M. A., & Greene, J. D. (2014). Social heuristics shape intuitive cooperation. *Nature Communications*, *5*(1), 3677.
Rustagi, D., Engel, S., & Kosfeld, M. (2010). Conditional cooperation and costly monitoring explain success in forest commons management. *Science*, *330*(6006), 961–965.
Selten, R. (1967). Die strategiemethode zur erforschung des eingeschränkt rationalen verhaltens im rahmen eines oligopolexperimentes. In *Beiträge zur experimentellen wirtschaftsforschung* (pp. 136–168). Sauermann, H..
Smith, V. L. (1982). Microeconomic systems as an experimental science. *The American Economic Review*, *72*(5), 923–955.
Soto, C. J., Napolitano, C. M., & Roberts, B. W. (2021). Taking skills seriously: Toward an integrative model and agenda for social, emotional, and behavioral skills. *Current Directions in Psychological Science*, *30*, 26–33.
Stantcheva, S. (2023). How to run surveys: A guide to creating your own identifying variation and revealing the invisible. *Annual Review of Economics*, *15*(1), 205–234.
Thielmann, I., Spadaro, G., & Balliet, D. (2020). Personality and prosocial behavior: A theoretical framework and Meta-Analysis. *Psychological Bulletin*, *146*(1), 30–90.
Thöni, C., Tyran, J.-R., & Wengström, E. (2012). Microfoundations of social capital. *Journal of Public Economics*, *96*(7), 635–643.
Thöni, C., & Volk, S. (2018). Conditional cooperation: Review and refinement. *Economics Letters*, *171*, 37–40.
Tibshirani, R. (2018). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *58*(1), 267–288.
Tippett, L. H. C. (1931). *Methods of Statistics*. London: Williams and Norgate.
Van Landeghem, B. (2019). Stable traits but unstable measures? Identifying panel effects in self-reflective survey questions. *Journal of Economic Psychology*, *72*, 83–95.
Volk, S., Thöni, C., & Ruigrok, W. (2011). Personality, personal values and cooperation preferences in public goods games: A longitudinal study. *Personality and Individual Differences*, *50*(6), 810–815.
Vossler, C. A., Doyon, M., & Rondeau, D. (2012). Truth in consequentiality: Theory and field evidence on discrete choice experiments. *American Economic Journal: Microeconomics*, *4*(4), 145–171.
Vranka, M., & Houdek, P. (2024). Moral hypocrisy and the dichotomy of hypothetical versus real choices in prosocial behavior. *Journal of Economic Psychology*, *105*, Article 102772.
Yezer, A. M., Goldfarb, R. S., & Poppen, P. J. (1996). Does studying economics discourage cooperation? Watch what we do, not what we say or how we play. *Journal of Economic Perspectives*, *10*(1), 177–186.