Optimisation Uncertainty*

Te Bao[1] Jiaoying Pei[2]

[1]Nanyang Technological University, Singapore [2]University of Cambridge, U.K.

13th November 2026

Latest version here

Abstract

Optimisation uncertainty — the difficulty of mapping a representation of the decision problem into an action — is particularly consequential in strategic games. It is distinct from the encoding uncertainty studied in the literature, which arises from two inputs: noisy beliefs about the world, and noisy access to one's own preference. We elicit confidence by asking participants how certain they are about their contributions in a repeated, randomly rematched public goods game, and remove or control each source of encoding uncertainty. Confidence does not respond to a treatment removing the first, nor does it rise with repetition as the second would predict, leaving optimisation uncertainty as what remains. From choices and confidence reports we then estimate *ideal* (full-confidence) and *default* (zero-confidence) decisions. The ideal, when others do not contribute, is positive for a majority of participants, so contributions are not motivated by confusion alone.

Keywords: Optimisation Uncertainty, Metacognition, Confidence, Efficient Coding, Cognitive Foundations of Decision-Making, Strategic Games, Public Goods Game

*We are especially grateful to Peter Bossaerts for substantive comments. We also thank the members of the LIPNE Lab and participants at various seminars for helpful discussions. This research is supported by the RIE2025 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) (Award I2301E0026), administered by A*STAR, as well as by Alibaba Group and NTU Singapore through the Alibaba-NTU Global e-Sustainability CorpLab (ANGEL), and the Leverhulme International Professorship at the University of Cambridge. This study was approved by the IRB of Nanyang Technological University (IRB-2022-834). An earlier version of this paper was circulated as "Cognitive Uncertainty, GPT, and Contribution in Public Goods Game". The present version has been substantially revised. Te Bao: School of Social Sciences, Nanyang Technological University, Singapore, Email: baote@ntu.edu.sg; Jiaoying Pei: Faculty of Economics, University of Cambridge, CB3 9DD Cambridge, United Kingdom, Email: jp2127@cam.ac.uk.

1 Introduction

A decision can be decomposed into two stages. Figure 1 summarises this decomposition.

The first is *representation*: forming an internal account of the relevant inputs to a choice. These inputs are of two kinds. Some are external, e.g., the state of the world. Some are internal, e.g., one's own preference. A large literature on the cognitive foundations of decision-making argues that both kinds of input are encoded imprecisely, and that many apparent violations of rational-choice axioms reflect this noisy encoding rather than preference anomalies (Glimcher, 2022). Work on rational inattention and efficient coding studies the first kind of noise, showing that the perceived frequency or magnitude of external states, such as outlier events, probabilities, payoffs, is systematically distorted by capacity-limited encoding (Woodford, 2012; Oprea and Vieider, 2024; Frydman and Jin, 2022; Payzan-LeNestour and Woodford, 2022). Work on the imprecise encoding of one's own preferences studies the second, showing that the brain's representation of subjective value is itself noisy (Polanía et al., 2019).

The second stage is *optimisation*: mapping one's representation of the inputs, however precise, into the action that best serves one's preference. Even a decision-maker who encodes both the state of the world and her preference without any noise at all must still solve this mapping problem, and the mapping can be hard on its own terms. We call the difficulty at this second stage *optimisation uncertainty*, to distinguish it from the *encoding uncertainty* that arises at the representation stage.

Confidence is the metacognitive judgment a decision-maker forms about the reliability of her own decision, supported by neural systems distinct from those underlying first-order cognition (Fleming and Daw, 2017). The canonical account in the psychology literature reads confidence as tracking the representation stage alone. Confidence c is defined as the probability that the chosen action a corresponds to the correct state of the world d, given a noisy internal signal about d denoted as $X \sim N(d, \sigma_X^2)$:

$$c = P(d = a \mid X, a, \sigma_X^2), \qquad X \sim N(d, \sigma_X^2). \tag{1}$$

On this account, low confidence can come from one place only: imprecise encoding of the world, reflected in a large σ_X^2 (Fleming and Daw, 2017), and confidence should track the precision of the underlying representation (e.g., Rademaker et al., 2012). Equation (1) is, in the terms just introduced, a definition of confidence for the representation stage only; it has nothing to say about the second stage, because it presumes that once X is known the mapping to an action is immediate.

We argue that this is the wrong account of confidence in strategic settings, and a minimal example makes the distinction concrete. In symmetric matching pennies, two players simultaneously choose a side of a coin. One player wins by matching, and the other wins by mismatching. The encoding view in (1) implies a player can never be fully confident, since without knowing the state

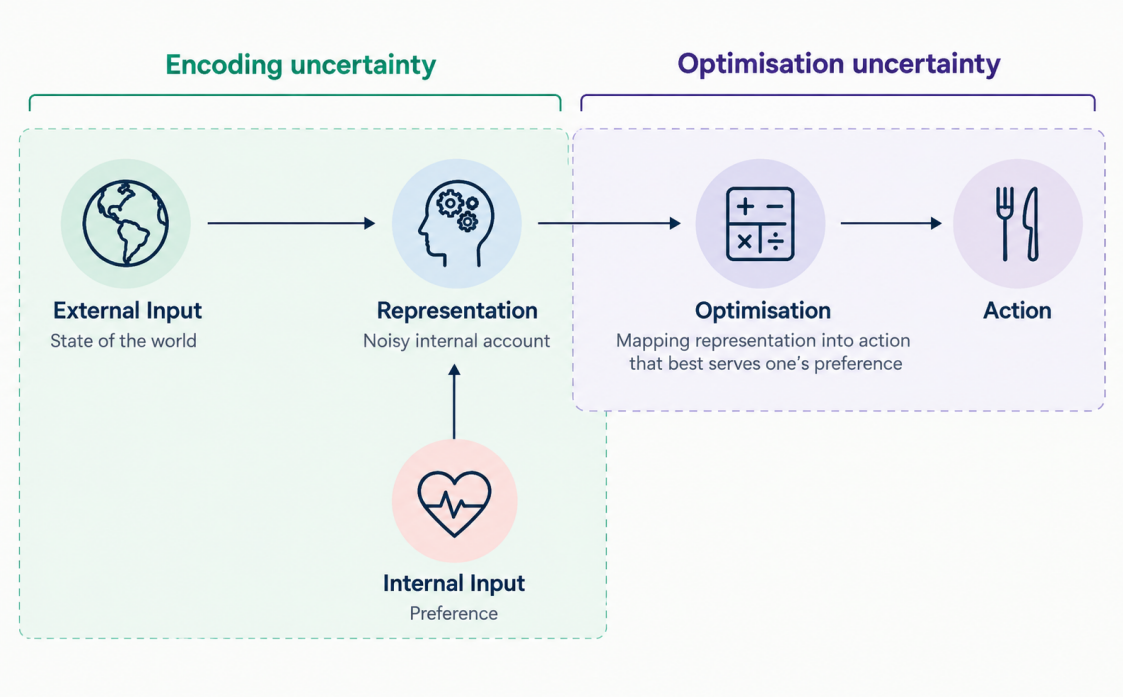


Figure 1. The two-stage decision process. External and internal inputs are encoded respectively into a representation of the decision problem, a process subject to encoding uncertainty. This representation is then mapped into an action through optimisation, a distinct process subject to optimisation uncertainty.

of world, i.e., opponent's choice, no action guarantees a win. Yet a player may reasonably report full confidence — not in winning, but in her strategy — because randomising fifty-fifty is provably optimal here. She expects to win only half the time and is nonetheless certain she is playing as well as she can. Her confidence, in other words, reflects optimisation over what is within her control, not a forecast of the result.

Symmetric matching pennies shows that confidence can be decoupled from knowledge of the world, but it does not show that optimisation itself can be difficult — and hence does not establish the need for a distinct optimisation uncertainty — since the optimal action, once seen, is obvious: mix fifty-fifty. Consider instead an asymmetric version in which the matcher earns more from matching on heads than on tails, while the mismatcher's payoff from mismatching is unaffected by which side the mismatch occurs on. Nash equilibrium still has the matcher randomising fifty-fifty, since her payoff from doing so must leave the mismatcher indifferent between heads and tails; but the mismatcher's own equilibrium strategy is no longer fifty-fifty, and recovering it requires reasoning through the *opponent's* payoffs rather than one's own. The mismatcher can therefore hold a perfectly precise representation of the game (its rules and payoffs, not the matcher's realised action) and of her own preferences, and still be genuinely unsure what to do — not just because the world is uncertain, but because the second stage of the decision, translating that representation into an action, is itself demanding.

Low confidence — equivalently, high metacognitive uncertainty, the term used in the economics literature after Enke and Graeber (2023) — can therefore arise at either stage: from noisy encoding of external or internal inputs, or from optimisation uncertainty at the second stage. Telling these apart requires more than a confidence report on its own; our central methodological contribution is a design that closes each channel in turn and uses confidence to detect what remains. Return to matching pennies for intuition. A player's confidence in her strategy could be low because she is unsure what her opponent will do (encoding uncertainty about external input), because she is unsure of her own objective (encoding uncertainty about internal input), or because, even given both, she cannot see which action is optimal (optimisation uncertainty). The two channels in encoding can in principle be closed one at a time, isolating whatever confidence shortfall optimisation uncertainty alone contributes.

Matching pennies cannot close them in practice, however: the internal input is trivial, since a player's preference is unambiguously to maximise her own monetary payoff, leaving no real scope for uncertainty about what she wants. We therefore turn to the random rematching public goods game, where both inputs are genuinely uncertain and both channels can actually be shut down one at a time. Four players simultaneously choose how much of an endowment to contribute to a public good; contributing benefits the whole group but is individually costly. The internal input, a player's prosocial preference — how much she weighs the group's payoff against her own — is neither observed by the experimenter nor, plausibly, known with precision by the player herself. The external input is equally uncertain. Because groups are randomly rematched each period from a pool of roughly 40 participants, a player has no reliable way to form a precise belief about what her current group will contribute, even if she can form some sense of the population's typical behaviour over time. Optimisation uncertainty is likewise present. Even a participant who knows her own prosocial preference exactly, and sees exactly what her group members contributed, must still settle on one of twenty-one discrete contribution levels, and neither input tells her which one is best. The payoff-maximising contribution is zero regardless of what others do, yet contributions are robustly positive across the literature (Fehr and Gächter, 2000).

Our design has two components, one for each source of encoding uncertainty. A separate, sequential treatment shows each participant her group members' individual contributions before she decides, removing encoding uncertainty about external input entirely. In addition, every treatment repeats the same underlying decision ten times, testing whether confidence sharpens as encoding uncertainty about internal input should predict. In each period, we elicit confidence by asking participants how certain they are that they would actually want to contribute the amount they just chose, rather than how certain they are that this amount maximises their payoff. Confidence reports of this kind are unincentivised, as is standard in the literature, and unincentivised reports have repeatedly been shown to carry information about the underlying decision process (Lu et al., 2025; Enke and Graeber, 2023).

Reported confidence is consistently below its 100% maximum, and we show that this shortfall

reflects optimisation uncertainty rather than encoding uncertainty. Two results carry the argument, and a third is consistent with it. First, confidence is unchanged once encoding uncertainty about external input is removed entirely, in the sequential treatment described above.[1] Encoding uncertainty about external input therefore cannot be the main driver of the shortfall. Second, confidence does not rise with repetition. On the encoding account, it should: confidence tracks the precision of the underlying representation, and pooling ten noisy draws of the same preference should yield a more precise estimate by the end than at the start. Encoding uncertainty about internal input therefore cannot be the main driver of the shortfall either. Third, contributions continue to vary from period to period even holding group behaviour fixed, consistent with participants trying different contribution levels rather than settling on one. Together, the first two results leave optimisation uncertainty as what remains, and the third is consistent with confidence capturing optimisation uncertainty.

Our finding has a direct methodological payoff for experimental economics. Psychology experiments typically elicit preferences directly, so that only encoding uncertainty is in play, and confidence elicitation is already standard practice there for detecting it. Economics experiments are instead built around incentive compatibility, which asks participants to translate a preference into an action calibrated to the experimental environment — precisely the step that generates optimisation uncertainty. If low confidence in our setting comes mainly from optimisation uncertainty, then economists have, if anything, stronger reason than psychologists to elicit confidence. That the reports are informative is not in doubt: although unincentivised, they predict participants' subsequent revisions to their contributions.

The cognitive foundations programme in behavioural economics (Enke and Graeber, 2023; Enke et al., 2025) likewise urges confidence elicitation. It proposes a "dual process" model in the spirit of the elaboration likelihood model in the persuasion literature (Petty and Cacioppo, 1986): a decision converges to a *default*, seemingly safer strategy as confidence falls, and to the *ideal* action the agent believes best serves her preference as confidence rises. Its signature finding is that responses made under low confidence are homogeneously less elastic than those made under high confidence.

Our public goods data tell a different story. We find no evidence of compression in decisions made under low confidence, which implies that defaults respond to others' contributions no less than ideals do in our dataset, and we confirm this directly by estimating the responsiveness of each. This accords with Krosnick (1991), who shows that non-differentiation is *only one* of several strategies respondents use to reach a satisfactory answer without substantial cognitive effort. Because low-confidence responses are not always predictably compressed, and because confidence does not rise with repetition, decisions cannot be corrected in aggregate. Confidence must be read choice by choice to identify which responses to trust.

Our design also speaks to a longstanding debate over what positive contributions actually re-

[1] We confirm that participants used this information, as their contribution is shown to respond to their current group members' contributions rather than to the previous period's.

flect: since any positive contribution is a departure from the payoff-maximising prediction of zero, it calls for an explanation. The public goods literature offers two broad explanations for such departures: that positive contributions reflect genuine prosocial preference, or that they instead reflect confusion, e.g., a failure to understand the incentives or to deduce the dominant, payoff-maximising strategy (Andreoni, 1995). Conditional cooperation, the tendency to contribute more when others are *expected* to contribute more (Fischbacher et al., 2001), sits uneasily between the two: it can be read as a symptom of confusion, e.g., participants mistakenly believing that matching others will lead to higher payoff (Burton-Chellew et al., 2016), or as a factor independent of confusion, e.g., altruism, social norm-following, or self-image concern (Wang et al., 2024).

We estimate each participant's contribution under full confidence, evaluated at a group contributing nothing. Confusion, in the sense used above, is closest to the difficulty of optimisation. Since confidence in our setting tracks optimisation uncertainty, extrapolating to full confidence removes confusion from the estimate, including the confusion-driven reading of conditional cooperation (Burton-Chellew et al., 2016). Evaluating this at a group contributing nothing removes the remaining, confusion-independent reading as well (Wang et al., 2024): an imprecise belief about others, a norm-following disposition, and a reciprocal preference all predict a contribution near zero when there is nothing to match.[2] What remains is the quantity the kindness-versus-confusion debate has sought: contribution attributable to genuine, unconditional preference, net of confusion. It cannot, on its own, separate this from self-image concern, since a participant motivated by either would contribute the same amount regardless of her group.

We find the estimate positive for a majority of participants: direct evidence that genuine prosocial preference — not merely confusion or conditional cooperation — is part of what drives cooperation in this game. A separate result corroborates this: contributions decline but do not converge to zero, and confidence does not rise over the game, contrary to what learning that self-interest is optimal would predict. Behaviour is therefore not driven by self-interest alone.

The paper proceeds as follows. Section 2 reviews the relevant literature. Section 3 describes the experimental design and Section 4 states our hypotheses. Section 5 sets out the empirical strategy, Section 6 reports results, and Section 7 discusses implications and concludes.

[2] For participants without direct information about their current group, the relevant measure is the previous period's average contribution, the best proxy available to them; for those with the information, it is their current group's contribution. We find that both groups condition on their respective measure to a similar degree.

2 Literature Review

2.1 Efficient Coding

Rationality is broadly defined as any behaviour that exhibits logical consistency and utility maximisation (von Neumann and Morgenstern, 1944). Assuming rationality is a standard practice in economics, and any deviation from this is described as a "bias". The recent literature on cognitive foundation of decision-making describes some of the apparent irrationality as the result of efficient use of limited brain resources, a behaviour commonly known as *efficient coding* (Glimcher, 2022)[3]. In particular, efficient coding prescribes that the brain cannot encode value with perfect precision and can represent it only noisily. As a result, a decision-maker acts on this noisy representation rather than the truth, so her choices can systematically depart from the rational-choice prediction.

Encoding Uncertainty about External Input. One example of the consequence of efficient coding is outlier blindness (Payzan-LeNestour and Woodford, 2022), whereby the brain devotes few resources to represent outcomes it expects to encounter only rarely.[4] Other examples include the difficulty of precisely representing numerical information, which gives rise to the description–experience gap (Oprea and Vieider, 2024), as well as features of prospect theory, read as efficient perceptual distortion of objective probabilities and payoffs (Woodford, 2012; Frydman and Jin, 2022). These findings collectively illustrate what we term *encoding uncertainty about external input*: noisy representation of features of the world outside the decision-maker.

Encoding Uncertainty about Internal Input. Efficient coding is not restricted to objective, externally verifiable features of the world. Choice behaviour has also been shown to be well described by an efficient-coding account of subjective, internal value: the subjective value of food, for instance, is represented noisily and adapts efficiently to the range of value a person typically encounters (Polanía et al., 2019). We term this *encoding uncertainty about internal input*: noisy representation of one's own preference, the counterpart to the external-input case just discussed.

2.2 Metacognition

Task performance and the ability to form confidence that tracks task performance are found to be separate abilities. The former is *first-order cognitive ability*, the latter is *metacognitive ability*. Fleming and Daw (2017) formalise the distinction as one between an "actor," who chooses, and

[3] Coding at imperfect precision is regarded as "efficient" instead of "bias" or "mistakes". At current coding precision, the brain—which accounts for roughly 2% of body mass—already consumes about 20% of the body's resting energy. Increasing precision would monotonically raise both metabolic cost and energy intake, yet greater precision is not always what natural selection prioritises (Sacco, 2026).

[4] In particular, after a prolonged period of good returns and low volatility, the brain does not anticipate the occurrence of extreme negative returns and underestimates their magnitude.

a "rater," who evaluates the choice by reporting confidence based on his information. There is empirical evidence on the dissociation, as the two abilities are supported by different brain systems (Fleming et al., 2014; Hoven et al., 2019).

Metacognitive Confidence. Confidence tracks the precision with which task-relevant information is encoded (e.g., Rademaker et al., 2012; Geurts et al., 2022). Confidence is modelled to decrease in (1), when internal signal X the agent forms about the state of the world becomes noisier.[5] While Section 2.1 documents encoding uncertainty about external and internal input through their effects on choice, here we turn to confidence: the metacognitive judgement an agent forms about how precise that same encoding is, still in tasks with a verifiable, experimenter-observable ground truth.

Metacognitive Uncertainty. The recent literature in behavioural economics examines the role of confidence, but largely in the domain of eliciting subjective preference. In particular, the inverse of confidence is termed *cognitive uncertainty* first by Enke and Graeber (2023). When referring to this literature, we use *metacognitive uncertainty* instead, to align with the psychology convention that treats confidence elicitation as a metacognitive rather than a first-order cognition process.

Metacognitive uncertainty in behavioural economics differs from the metacognitive confidence literature discussed before in the nature of the task. There, confidence pertains to decisions with a verifiable ground truth; in subjective tasks, however, accuracy is difficult even to define, since the "true value" is not observable to the experimenter. Instead, participants' choices are regarded as revealing their true preference, when they consistently choose the same options; consistency is found to correlate with high confidence (Polanía et al., 2019). Low confidence, in other words, predicts inconsistency, of which costly information search in perceptual tasks is one manifestation, as an agent unsure of her answer seeks out further evidence before committing to it (Schulz et al., 2023). Since no such external information is available in a subjective-preference task, we test the analogous manifestation: whether low confidence predicts a greater extent of revising one's choice when the same underlying decision recurs.

Although our experiment elicits a subjective preference, we refer simply to *confidence* for most of the paper, reserving *metacognitive uncertainty* for points of direct contact with this literature.

2.3 Optimisation Uncertainty

Both economics and psychology experiment elicit preferences. The prominent difference is that economics would usually not directly ask participants what they want; instead, here we would

[5] Encoding noise may arise either because the agent encodes available information imprecisely or because the information itself is objectively degraded (Bang and Fleming, 2018).

almost always ask participants to translate what they want to what they do. This is known as incentive compatibility, implemented with hope that participants would tell the truth, as their actions are designed to have monetary consequences. We refer to the uncertainty on accurately mapping preference into actions as *optimisation uncertainty*, distinct from the encoding uncertainty discussed in 2.1.[6] As the asymmetric matching pennies example in the 1 illustrates, this mapping can be genuinely uncertain even when both the game and one's own preference are known exactly. We study how much confidence reports capture optimisation uncertainty by removing or controlling encoding uncertainty directly, one channel at a time.

Removing Encoding Uncertainty about External Input. In strategic settings, mapping a preference into an action can depend on what other players do, so encoding uncertainty about external input — here, *strategic uncertainty* (Heinemann et al., 2009) — is a natural concern. The Level-K literature takes this concern seriously, attributing any deviation from Nash entirely to a player's belief about others (Nagel, 1995; Stahl and Wilson, 1995; Camerer et al., 2004; Alaoui and Penta, 2016) — but in doing so, it assumes optimisation uncertainty away entirely.[7] Players in the public goods game face the same strategic uncertainty, since players' contributions are often found to be conditional on their beliefs about others' contributions (Fischbacher et al., 2001), known as conditional cooperative tendency. We take the opposite approach: rather than assuming optimisation uncertainty away, we remove strategic uncertainty directly, and test whether a shortfall in confidence remains. As it will be explained in Section 3.3 , our implementation removes payoff uncertainty at the same time while removing strategic uncertainty. Combined with our test of internal encoding uncertainty below, this lets us attribute what remains, if anything, to optimisation uncertainty.

Controlling for Encoding Uncertainty about Internal Input. Unlike a game such as matching pennies, where a player's preference is unambiguously to maximise her own payoff, the public goods game elicits a genuinely uncertain preference: how much a player weighs the group's payoff against her own is plausibly not known to her with precision, leaving real scope for encoding uncertainty about internal input. Section 4 develops a test for this channel from the encoding account of confidence in psychology literature (Fleming and Daw, 2017), exploiting repeated elicitation of the same underlying preference.

[6] Note that optimisation uncertainty represents the difficulty participants encounter in an optimisation problem that is tractable. The intractable tasks in economics and their implications are discussed otherwise in computational complexity literature (e.g., Bossaerts et al., 2024).

[7] In the p = 1/2 beauty contest game, a participant who guesses 25 is classified as level-1 because 25 is the best response to a belief that others guess 50; the possibility that she holds some other belief and failed to best-respond to it is not entertained. With the belief as the only free parameter, any observed action can be rationalised, and optimisation uncertainty is implicitly assumed to be zero.

2.4 Insensitive Decisions under High Metacognitive Uncertainty

The central claim of the metacogntive uncertainty literature is that choices made under low confidence are *insensitive*, or less elastic with respect to fundamentals (e.g., objective payoff in choice under risk, time delay in intertemporal choice).

Using a Bayesian model, Enke and Graeber (2023) derive the observed decision as a weighted combination of the utility-maximising decision and an insensitive default, with the weight on the former increasing in the precision of the agent's signal. Their measure of low confidence is a composite one, capturing both what we call encoding uncertainty and optimization uncertainty without separating them. Low-confidence decisions are predicted to place more weight on the default, resulting in a compressed response pattern. Enke and Graeber (2023) and Enke et al. (2025) observe these inelastic patterns in choice under risk and intertemporal choice, and hence argue that abnormalities like probability weighting and hyperbolic discounting are not entirely matters of preference, but are at least partly explained by metacognitive uncertainty.

The inelasticity pattern observed there has also been discussed in psychology literature. In survey methodology, satisficing under cognitive load produces non-differentiation on rating scales: when asked to evaluate a series of items, a respondent selects a plausible scale point such as an intermediate option for the first and repeats it thereafter, thereby showing an insensitivity to item content (Krosnick, 1991). Note that Krosnick (1991) highlights anchoring at an intermediate option is only one example of satisficing, known as non-differentiation; participants can select any other answers as their anchoring for each choice. We will turn to this now.

Defaults. The metacognitive uncertainty literature estimates a single population-level default, and find it to be an intermediate option. This however, does not distinguish whether it is a common intermediate default across participants, or there are in fact heterogeneous defaults across both choices and participants that can be extreme but average to an intermediate value.

Evidence from more recent studies that recover the default at the individual level favours the second possibility. Charles et al. (2024) measure it directly at the subject-period level and find that subject fixed effects account for most of its variation, so that defaults differ across participants. Defaults may also differ across decisions for a given participant: Andrikogiannopoulou and Papakonstantinou (2020) find that the reference point determining whether an individual views an outcome as a gain or a loss adapts to past realisations at a rate specific to her.

We estimate the default separately for each participant and allow it to respond to others' contributions. It does, so that a participant's default differs across her decisions as her group's behaviour changes — and we find no evidence that it responds less than her ideal, which is the property on which the compression prediction rests.

Weighting Function. The weighting rule in the metacognitive uncertainty literature is not specific to economics. In the Elaboration Likelihood Model of persuasion (Petty and Cacioppo, 1986; O'Keefe, 2008), judgment mixes scrutiny of issue-relevant arguments with reliance on simpler cues, weighted by the receiver's *elaboration*.

Charles et al. (2024) estimate this specification directly in a laboratory investment task. They find evidence where insensitivity is driven by a noisy cognitive process. They further show that the estimated weights of default (that are directly measured) and subjective expected payoff are not statistically distinguishable from summing to one, hence validating the weighting rule. We therefore adopt the weighting function in our estimation as given.

2.5 The Public Goods Game Literature

The public goods game has been used primarily to test whether human behaviour conforms to the self-interest postulated by standard economic theory, or whether individuals also place weight on the payoffs of others (i.e., kindness). The central empirical finding is that average contributions are robustly positive—commonly 40 to 60 percent of the endowment in early periods (Fehr and Gächter, 2000)—and, under repetition, decline toward but do not reach full free riding.

Kindness vs. Confusion The literature holds competing opinions on what positive contributions implies: whether it is evidence of kindness, or some misunderstandings leading to the appearance of cooperation (e.g., Andreoni, 1995; Houser and Kurzban, 2002; Burton-Chellew et al., 2016). More recently, Wang et al. (2024) argue that, once confusion is minimised through additional participant training, genuine social preferences remain the dominant driver of human cooperation, and use a post-experimental questionnaire to measure the prevalence of distinct candidate motives, including altruism, social norms, and self-image concerns. These literature usually assign observed cooperation to one source or the other, whether by sorting subjects into types or by decomposing aggregate contributions, or removing the recipient of kindness by design so that any remaining contribution must reflect confusion. The weighting function we adopt here, however, does not require that assignment. Instead, confusion and kindness are permitted to operate within the same decision.

3 Experimental Design

This paper's primary goal is to identify whether optimisation uncertainty is present in strategic games, distinct from the encoding uncertainty documented in the literature. We use the public goods game because, unlike matching pennies, both a player's internal preference and her belief about others are genuinely uncertain, giving encoding uncertainty real scope to operate and so a

real confound to rule out. The mapping from preference to action is also more demanding: a player must settle on one of twenty-one discrete contribution levels, rather than choosing between two options as in matching pennies, leaving more room for optimisation uncertainty as well. We identify optimisation uncertainty by removing or controlling for each source of encoding uncertainty in turn: external input, via a treatment that shows a participant what her group members contributed before she decides, and internal input, via repeated elicitation of the same underlying preference. This section describes the design that lets us do both; Section 4 states the hypotheses each design feature is meant to test.

3.1 The Game

We implement the without-punishment public goods game of Fehr and Gächter (2000), a social dilemma in which participants can benefit the group at their own cost. The interface is shown in Figure A.1, top left panel.

In each period t ($t = 1, \ldots, 10$), each of the $N = 4$ participants in a group receives an endowment of 20 points and decides how to allocate it between a private and a public account.[8] The marginal cost of contributing to either is 1. The private account yields a marginal return of 1, while each contribution to the public account is multiplied by $M = 1.6$ and shared evenly among the N members, so that the marginal return from the public account is $m = M/N = 0.4$. The payoff of participant i in period t is

$$\pi_t^i = 1 \cdot (20 - p_t^i) + 0.4 \cdot p_t^i + 0.4 \cdot \sum_{j=1}^{3} p_t^j.$$

A participant who cares only about her own payoff contributes nothing. Any positive contribution is read as evidence of prosocial preferences.

Group composition changes each period and participants are told so, which rules out direct reciprocity, though it cannot rule out a generalized reciprocity norm applied to strangers. And only a participant's own payoff is displayed, for ten seconds, which limits inequality aversion.[9] The payoff interface is shown in Figure A.1, bottom right panel.

[8] Participants were told that the game would consist of ten rounds. The session in fact continued with a surprise restart of another ten rounds, designed to address a different set of research questions and announced only after round ten, as in Andreoni (1988). Since it was unannounced at the beginning of the experiment, it cannot have affected the decisions we study, and data from the restart are reported separately.

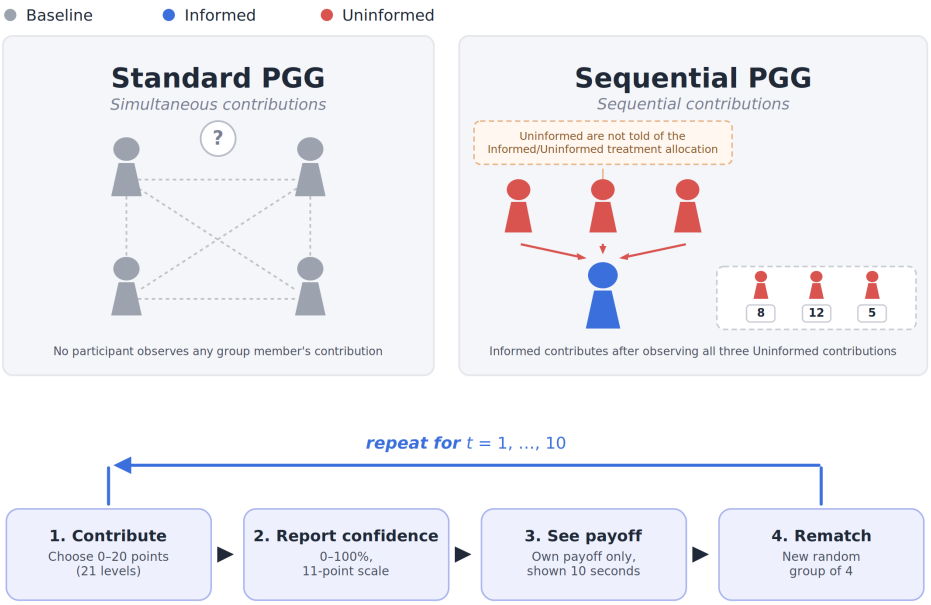
[9] The group's average payoff could still be inferred from one's own. The ten-second display limits the opportunity for such calculations.

3.2 Confidence Elicitation

After each contribution decision, and before seeing the payoff, participants report how certain they are that they would want to contribute the amount they chose. A participant who entered 5 is asked how certain she is that she would actually want to contribute between 4 and 6 points, and answers on an eleven-point scale running from 0% ("very uncertain") to 100% ("completely certain"). The interface is shown in Figure A.1, bottom left panel. Confidence is not incentivised, so as not to add further complexity to the task. Unincentivised elicitation is standard in both the psychology and the economics literature.

A given confidence report, on its own, does not reveal which of the three sources discussed in Section 2.3 it reflects. The next two subsections describe how our design separates them.

3.3 Removing Encoding Uncertainty about External Input



Treatment	# Participants	Experiment	Date	# Sessions	Color	Avg Payoff (Points)
Baseline	140	Standard PGG	2023-06-22	3	Gray	22.78
Informed	39	Sequential PGG	2023-10-18, 2025-04-18	4	Blue	24.97
Uninformed	117	Sequential PGG	2023-10-18, 2025-04-18	4	Red	22.99

Figure 2. Experimental treatment assignment and group composition.

Under random rematching, a participant has no way of forming a precise belief about what her current group will contribute, so the strategic uncertainty discussed in 2.3 remains. We remove it in the *Informed* treatment of the *Sequential PGG* experiment. Each Informed participant is placed in a group with three *Uninformed* participants and observes all three of their individual contributions before submitting her own. This also renders the payoff consequence of every possible

contribution level deterministic, thereby removing their *payoff uncertainty*. Treatment assignment is fixed throughout. Uninformed participants are not told that their contribution is observed by another group member before they decide, and receive instructions identical to those given in the *Baseline* treatment of the *Standard PGG* experiment, in which all participants play the game described in Section 3.1.[10]

There are therefore three treatments in total: Baseline, Informed, and Uninformed (see Figure 2).[11]

3.4 Controlling for Encoding Uncertainty about Internal Input

Every treatment repeats the same underlying decision ten times ($t = 1, \ldots, 10$), with groups rematched each period. This repetition is the design's control for encoding uncertainty about internal input: as discussed in 2.3, if a participant's preference is fixed but noisily accessed, confidence should rise across periods as her estimate sharpens. Section 4 states this as a hypothesis.

4 Testable Hypotheses

4.1 What Does Confidence Reflect?

Reported confidence in our data is consistently below its maximum, as it will be shown later. Section 3 described how each design component closes one source of encoding uncertainty; here we state the resulting predictions, treating whatever confidence shortfall survives both as evidence of optimisation uncertainty.

4.1.1 Encoding Uncertainty about External Input

Informed treatment removes encoding uncertainty about external input of strategic and payoff uncertainties for Informed participants, described in Section 3.3. If the confidence shortfall is attributable to it, they should report higher confidence than participants who face the same decision without that information; if it is not, confidence should not differ.

[10] The design involves no deception; it relies on the selective withholding of information. Had any participant asked whether their contribution would be observed by others before they decided, the experimenter was instructed to reply: "This is not part of the instructions, and I am unable to tell you." No such question was raised. This may be attributable to the online implementation: in a conventional laboratory it would be easier for Uninformed participants to infer that other group members were operating under different instructions, or were waiting on their decisions. The experimenter did receive queries from Uninformed participants asking why their screen continued to display "Waiting for the other participants" after a prolonged wait. The scripted reply was: "The experiment is configured so that it cannot proceed until all participants in the session (around 40) have completed their decisions. Please be patient and try refreshing the page." Requiring all participants to finalise their decisions before any may proceed is standard practice in settings where group composition is randomised each period.

[11] The full instructions can be found in Appendix C.

Hypothesis 1. *If the confidence shortfall reflects encoding uncertainty about external input, then reported confidence is higher in Informed than in Baseline and Uninformed.*

Granting Informed participants their group members' exact contributions does not guarantee that they use them: they may not attend to the numbers, or may not trust them. Rejecting Hypothesis 1 is therefore ambiguous along a second dimension as well: between removing strategic and payoff uncertainty leaving confidence unchanged, and the information never being used, so that these uncertainties was never actually removed. We check this on contributions:

Manipulation Check 1. *If Informed participants use the information available to them, then their contributions condition on the contemporaneous contributions only they can observe, rather than on the previous period's average, which is the best proxy available to participants in the other treatments.*

4.1.2 Encoding Uncertainty about Internal Input

The game is repeated ten times, and each period asks the participant to consult the same underlying prosocial preference. According to (1), if that preference is fixed and each encounter yields an independent noisy signal of it, and those signals are pooled across periods, then precision rises with the number of draws, and confidence, which tracks precision, should rise with it.

The test assumes that the preference a participant consults is stable across periods, and that she carries forward what earlier draws have told her about it rather than consulting it afresh each time. This is a natural assumption for a preference that does not change from period to period, and nothing in the design gives her reason to discard what earlier periods have told her about it.

Hypothesis 2. *If the confidence shortfall reflects encoding uncertainty about internal input — a preference that is stable across periods — then reported confidence rises with repetition.*

The same prediction follows from a second account, which the repetition test therefore also speaks to. Suppose participants care only about their own payoff but do not know which contribution maximises it, so that they face a search problem and arrive at the answer by trial and error, as in a bandit problem. Learning of this kind is exhausted with experience, so repetition should again resolve their uncertainty: contributions should converge towards zero and confidence should rise as they do.[12]

[12] A further possibility is that participants do not initially know how their contribution maps into their payoff, even when others' contributions are known, and learn that mapping through repetition. This would be an additional source of encoding uncertainty, which we call *mapping uncertainty*. We regard it as unlikely, since the control questions participants answered before the tasks covered eight cases of this mapping. It is in any case inconsistent with what we find later: confidence does not rise with repetition even among Informed participants, for whom others' contributions and hence each level of their contribution is deterministic.

4.1.3 Optimisation Uncertainty

The external channel is closed simply by handing the participant the input she was missing: the Informed treatment gives her the information directly, with nothing for her to do beyond reading it. The internal channel is closed passively as well: each period gives her another noisy read of the same fixed preference, and pooling repeated reads sharpens the estimate without any need to act and check whether the action taken was right. Optimisation uncertainty is different in kind. What is uncertain there is not an input but the mapping from a known input and a known preference to an action, and there is in general no way to read that mapping off directly. Discovering it requires deliberately trying candidate actions and inferring, from how the outcome feels relative to one's preference, whether it was the right one.

Recall the asymmetric matching pennies example from 1: a mismatcher who knows the matcher's payoffs exactly, and who takes it as given that the matcher plays her Nash strategy, cannot read off her own equilibrium mixing probability by inspection. This is not to say the problem is unsolvable: the correct mixing probability is fully determined by the two payoff matrices, in the same tractable sense as the rest of the game. The question is whether a participant, without training in solving for it, reliably arrives at that answer rather than merely being capable of it in principle. Discovering that she should play, say, tails two-thirds of the time plausibly requires experimenting with different frequencies and tracking whether her payoff improves — a process with no analogue in either encoding channel, where the right answer is either given outright or converged upon by simply averaging.

Note that the matching penny example is also stripped of any encoding uncertainty on internal input: the mismatcher's own preference is simply to maximise her own payoff, leaving nothing to encode noisily. If trial and error is needed even there, it should be needed at least as much in a setting like the public goods game, where the preference being served is itself richer and only imprecisely known to the participant herself. In other words, the mapping problem does not become easier once a genuinely uncertain preference is added to it, and plausibly becomes harder.

Optimisation uncertainty should therefore leave a behavioural trace: a participant who does not know which contribution best serves her preference should keep trying different ones, so her contributions should not settle. Variability alone cannot separate this from noisy encoding, which would also generate it, so we treat it as a consistency check and not as an identifying test.

Hypothesis 3. *If the confidence shortfall reflects optimisation uncertainty, then contributions should continue to vary from period to period, controlling for group members' contributions, even once Hypotheses 1 and 2 are both rejected.*

4.1.4 Choice Variability under Confidence

A further check asks whether confidence in our setting behaves as it does in the literature. There, low confidence predicts greater variability in choices across repetitions of the *same* problem (Enke and Graeber, 2023), read as a form of information seeking: an agent unsure of her answer tries another and observes the outcome (Schulz et al., 2023). This is distinct from Hypothesis 3, which concerns whether contributions continue to vary at all, as a consistency check on optimisation uncertainty; here we ask instead whether that variability is concentrated specifically in low-confidence periods, as the literature's information-search account implies.

Whether the same relationship holds in our setting is worth establishing, but its interpretation will not carry over: in the literature, the same choice is repeated unchanged, so a different answer means revising a response to a fixed problem. Here, the group changes every period, so a different contribution may just track a different group rather than reflect any revision at all — hence we control for it.

Hypothesis 4. *If low confidence in our setting operates as it does in the literature, then it predicts greater variability in contributions across periods than high confidence does, controlling for group members' contributions.*

4.2 How Does Low Confidence Distort Contributions?

If decisions made under high confidence differ systematically from those made under low confidence, then a contribution figure alone is an incomplete description of what a participant wanted, and eliciting confidence carries information that the choice does not.

The metacognitive uncertainty literature's signature finding is that low-confidence choices are less responsive to the fundamentals of the decision than high-confidence choices are. We ask whether this replicates in our setting, where the relevant fundamental is others' behaviour, which conditional cooperation makes a leading determinant of contributions.

Hypothesis 5. *If the literature's compression pattern replicates in our setting, then contributions are less responsive to group members' contributions under low confidence than under high confidence.*

The literature explains this pattern with a specific mechanism: as confidence falls, a decision is increasingly blended towards a common *default* that does not itself track the fundamentals of the problem. That explanation rests on an assumption about the default itself — that it is an intermediate option, common across participants, and by construction unresponsive to the specifics of the decision. Because we elicit confidence at the level of the individual decision, we can estimate the default and the *ideal* — the choice a participant makes when she reports full confidence — separately for each participant and test this assumption directly.

Hypothesis 6. *If the default is the intermediate, unresponsive anchor the literature describes, then its responsiveness to group members' contributions must be smaller than that of the ideal.*

4.3 The True Prosocial Preference

Finally, because confidence is reported at the level of the individual decision, we are able to recover the *net ideal*: the contribution a participant would make at full confidence were her group members to contribute nothing. The net ideal is net of both metacognitive uncertainty and conditional cooperation, so it isolates genuine, unconditional preference from the confusion and belief-driven conditioning that have competed to explain positive contributions in the literature. This lets us state directly the question the kindness-versus-confusion debate has posed: whether cooperation in this game is prosocially motivated, or is instead an artifact of confusion and conditional cooperation once those are stripped away.

Hypothesis 7. *If prosocial preference, rather than confusion or conditional cooperation, is part of what drives cooperation in this game, then the net ideal is positive.*

5 Empirical Strategy

The main specifications are estimated by linear panel regression. This section sets out the four principal methodological considerations, tailored to the experimental design and hypotheses of this paper.

First, standard errors are clustered at the session level. Participants are randomly rematched into new groups each period, so a shock experienced by one participant in a given period — being paired with unusually low or high contributors, say — may influence her subsequent behaviour and propagate to future group members through the rematching process. This generates cumulative interdependence among all participants within a session.[13]

Second, our design yields only seven sessions, and cluster-robust standard errors are biased downward when the number of clusters is small, so that conventional critical values over-reject. We therefore report, alongside each specification, standard errors clustered at the participant level in B.1 in the appendix. ($N = 296$), where the small-cluster problem does not arise. This is a weaker assumption on the dependence structure than the session-level clustering we prefer, and it is not a correction for the downward bias; it establishes only that our conclusions do not depend on which level of clustering is adopted. Our results are qualitatively unchanged across the two.

Third, each participant makes ten decisions, and these are strongly correlated within participants: an intercept-only random effects specification gives $\rho = 0.54$ for reported confidence, and

[13] Figure A.2 plots mean contributions over time for two representative sessions from the Sequential PGG experiment. Session 7 exhibits a steeper decline in contributions than Session 4, indicating pronounced within-session correlation.

$\rho = 0.48$ for contribution, so that over half its variance lies between rather than within participants. This has two implications for our analysis. In the regressions, we employ a random effects estimator to capture unobserved heterogeneity across participants.[14] In the non-parametric tests, we compare participant-level means across treatments rather than pooling individual decisions, since treating a participant's ten reports as independent observations would inflate the effective sample size and understate standard errors accordingly.

Fourth, an important component of the analysis concerns whether participants in different treatments behave differently. We therefore introduce interaction terms between treatment indicators and covariates. Because these additional parameters increase model complexity and reduce statistical power, we use the Bayesian Information Criterion to compare specifications with and without them. All specifications incorporating interaction terms yield higher BIC values, indicating that the additional parameters are not justified by the improvement in fit. We therefore report the more parsimonious specifications in the main text, with the interacted specifications reported in full in Table B.1 in the appendix.

6 Results

We begin by asking which of the three sources set out in Section 4.1 — encoding uncertainty about external input, encoding uncertainty about internal input, or optimisation uncertainty — the confidence shortfall reflects. We then examine how decisions made under low confidence depart from those that would express preference alone.

6.1 What does Confidence Reflect?

6.1.1 Confidence falls short of its maximum

The top left panel of Figure 3 plots mean reported confidence in the two experiments, Standard PGG and Sequential PGG. Despite the difference in design, mean confidence is close to 78% in both experiments and statistically indistinguishable across them (rank-sum $p = 0.91$)[15].

Result 1. *Mean reported confidence is approximately 78% in both experiments, roughly 22 percentage points below its maximum, and does not differ across experiments.*

[14] Random effects requires the participant effect to be uncorrelated with the regressors; this is plausible here by design, since group composition is randomised each period and participants are therefore not systematically matched with group members who share their own contribution tendencies. We additionally run all of our analyses with a fixed effects specification and find similar results.

[15] Figure A.3 reports the full distribution of reported confidence. Confidence of 100 is the modal response, chosen far more often than a uniform distribution over the scale's eleven points would predict (roughly 37% of reports, 27% of remove participants who always report full confidence against 9% under uniformity), in every treatment. The shortfall from the maximum therefore does not reflect endpoint avoidance; it reflects a right-skewed distribution with a meaningful tail of lower reports below an otherwise dominant ceiling response.

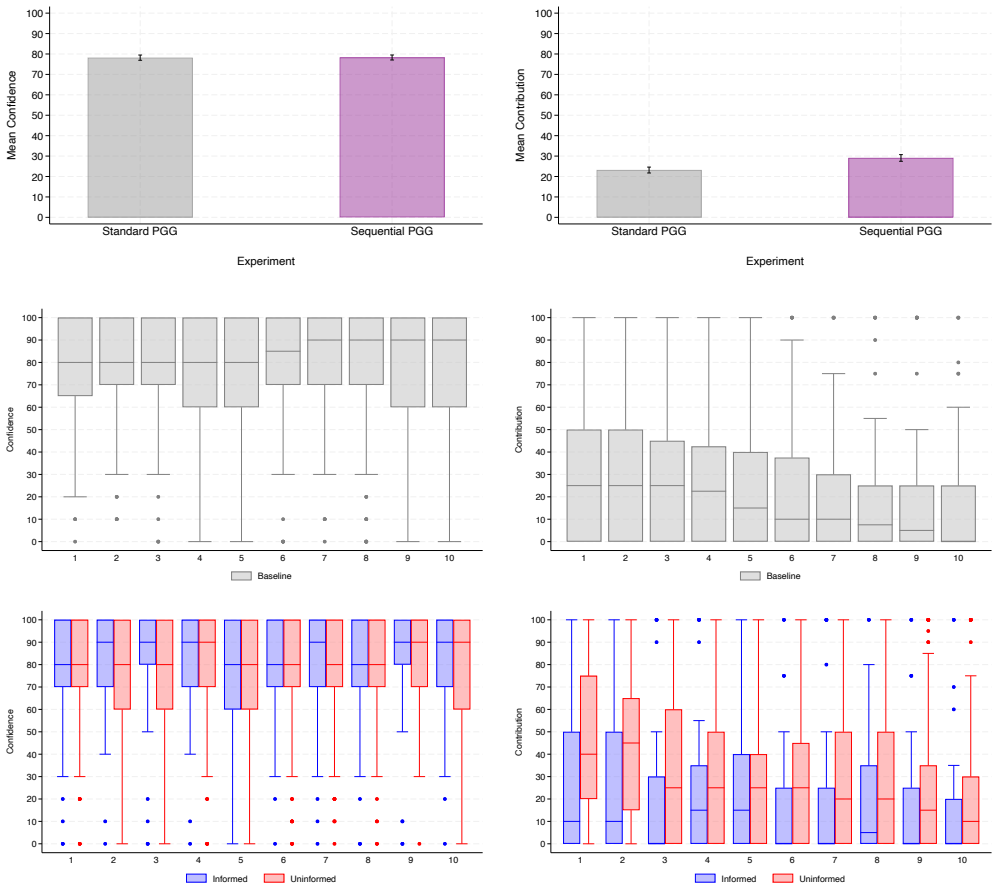


Figure 3. Confidence and contributions by experiment, treatment, and period. The left column reports confidence and the right column contributions. The top panel gives pooled means by experiment, with error bars spanning ± 1.96 standard errors of the mean. The middle panel shows the Baseline treatment of the Standard PGG experiment; the bottom panel shows the Informed and Uninformed treatments of the Sequential PGG experiment. The middle and bottom panels report the distribution across the ten periods as box plots, displaying the interquartile range (IQR) bounded by the first and third quartiles, the median (horizontal line inside the box), adjacent values defined as the most extreme observations within $Q1 - 1.5 \times IQR$ and $Q3 + 1.5 \times IQR$ (whiskers), and outliers beyond these bounds (dots).

Table 1. Linear modelling of dependent variable on variables of interest and controls, allowing for random effects (intercept) at participant level in Models (1)–(5) and fixed effects (intercept) at participant level in Model (6). Robust standard errors clustered at session level ($N = 7$) in parentheses. p-values in brackets.

Model	(1)	(2)	(3)	(4)	(5)	(6)
Dependent Variable	p_t	p_t	$c_{t+1} - c_t^i$	$\lvert p_{t+1} - p_t \rvert$	p_t	p_t
	Informed	Baseline, Uninformed				
Variable of Interest						
Constant, $\bar{\mu}$	21.21	36.25	-0.02	21.18	35.29	33.56
	(2.61)	(3.08)	(0.16)	(1.88)	(2.49)	(24.18)
	[0.000]	[0.000]	[0.907]	[0.000]	[0.000]	[0.165]
c_t^i, β_1	-0.16	-0.24		-0.12	-0.23	
	(0.03)	(0.05)		(0.02)	(0.04)	
	[0.000]	[0.000]		[0.000]	[0.000]	
c_t^i, β_1^i						-0.19
						(0.28)
						[0.492]
$\bar{p}_t^j$, β_2	0.32					
	(0.09)					
	[0.000]					
$\bar{p}_{t-1}^j$, β_3	0.09	0.30				
	(0.09)	(0.02)				
	[0.315]	[0.000]				
$\bar{p}^j$, β_4					0.25	0.26
					(0.09)	(0.03)
					[0.008]	[0.000]
$(c_t^i \times \bar{p}^j)$, β_5					0.00	
					(0.00)	
					[0.520]	
Control						
$\Delta\bar{p}^j$, γ_1			0.05			
			(0.02)			
			[0.002]			
$\lvert\Delta\bar{p}^j\rvert$, γ_2				0.09		
				(0.03)		
				[0.007]		
$p_{t+1} - p_t$, γ_3			-0.20			
			(0.03)			
			[0.000]			
Interaction with Treatment	No	No	No	No	No	No
BIC	3263.42	20985.42	21462.13	21300.47	24578.92	NA
Wald p-value	0.000	0.000	0.000	0.000	0.000	NA
Observations	351	2,312	2,406	2,406	2,702	2,702
Number of participants	39	257	296	296	296	296

In neither experiment does confidence approach its maximum. The remainder of 6.1 asks what the shortfall reflects.

6.1.2 Manipulation Check

Since our test of Hypothesis 1 turns on a comparison of reported confidence, we first establish on contributions alone that Informed participants used the information they were additionally given.

The bottom right panel of Figure 3 provides a first indication that the information was used. Informed participants contribute less than Uninformed participants (rank sum $p < 0.001$ at both decision and participant level)[16].

A difference in levels shows that behaviour changed, but not that it changed because participants conditioned on what they saw. We therefore test the channel directly, exploiting an informational asymmetry in the design. Informed participants observe their current group members' contributions before deciding, and so can condition on them if they believe the information; otherwise they must fall back on the preceding period's average, which is all that participants in the other treatments have available. We therefore estimate, for Informed participants,

$$p_t^i = \mu^i + \beta_1 c_t^i + \beta_2 \bar{p}_t^j + \beta_3 \bar{p}_{t-1}^j + \varepsilon_t^i, \qquad i \in \text{Informed}, \tag{2}$$

and, for participants in the remaining treatments,

$$p_t^i = \mu^i + \beta_1 c_t^i + \beta_3 \bar{p}_{t-1}^j + \varepsilon_t^i, \qquad i \in \{\text{Baseline}, \text{Uninformed}\}. \tag{3}$$

In both, μ^i is a participant-level random effect, ε_t^i an idiosyncratic error, and standard errors are clustered at the session level. We control for reported confidence c_t^i, since whether contributions and confidence are related is the question this paper sets out to answer, and we therefore hold it fixed when estimating how contributions respond to the information available; we return to this relationship in 6.2.2.

Table 1 Columns (1) and (2) report the results. Informed participants condition on the contemporaneous contributions that only they can observe ($\hat{\beta}_2 = 0.32$, $p < 0.001$), and not on the preceding period's average ($\hat{\beta}_3 = 0.09$, $p = 0.32$). Participants in Baseline and Uninformed, for whom the preceding period is the only information available, condition on it to a similar degree ($\hat{\beta}_3 = 0.30$, $p < 0.001$). Informed participants therefore substituted better information for worse, rather than becoming more responsive to others overall[17].

[16] The permutation test reported in Figure A.5 confirms that it is not an artefact of unbalanced sample sizes.

[17] We confirm this in a pooled specification reported Table B.1 Column (2), in which contributions are regressed on the information available to each participant — the contemporaneous group average for Informed participants and the preceding period's average for the others — interacted with treatment indicators. Neither interaction is statistically

Manipulation Check Result 1: Informed participants condition their contributions on their current group members' contributions rather than on the previous period's average. The additional information was used, so strategic and payoff uncertainties were removed as intended.

6.1.3 Not Encoding Uncertainty about External Input

We now test Hypothesis 1: whether the confidence shortfall reflects encoding uncertainty arising from external input of strategic and payoff uncertainties. We established that Informed participants used the contributions they were shown, so for them this uncertainty was removed. If this uncertainty were the source of the shortfall, their confidence should exceed that of participants for whom it remains.

The bottom left panel of Figure 3 plots the distribution of reported confidence in the two treatments of the Sequential PGG experiment. Mean confidence is 78.82 in Informed and 78.10 in Uninformed, a difference of 0.72 percentage points that is not statistically distinguishable (rank-sum $p = 0.50$). Baseline lies between the two, plotted in the middle left panel of Figure 3, at 78.21, and differs from neither (Informed: $p = 0.70$; Uninformed: $p = 0.74$). Mean confidence across all three treatments therefore spans less than one percentage point. The permutation tests reported in Figure A.6 confirm that these results are not an artefact of unbalanced sample sizes.

The magnitudes are what carry the argument. Removing strategic and payoff uncertainty entirely moves mean confidence by less than one percentage point, against a shortfall of 22 points. Whatever residual difference the data permit is far too small to account for the gap.

Result 2. *Removing strategic and payoff uncertainty leaves reported confidence essentially unchanged. Confidence therefore does not reflect encoding uncertainty about external input.*

6.1.4 Not Encoding Uncertainty about Internal Input

We now test Hypothesis 2: if the shortfall reflects encoding uncertainty about internal input, confidence should rise with repetition as participants' estimates of their own preference sharpen.

The middle and bottom left panels of Figure 3 plot the distribution of reported confidence by period. The median is stable across the ten periods in all three treatments, and the interquartile range does not narrow. We test this more formally by estimating

$$c^i_{t+1} - c^i_t = \mu^i + \gamma_1 \Delta \bar{p}^j + \gamma_3 (p_{t+1} - p_t) + \varepsilon^i_t \tag{4}$$

where

distinguishable from zero ($p = 0.66$ for Informed, $p = 0.73$ for Uninformed), so the rate at which participants condition on the information they have does not differ across treatments.

$$\Delta\bar{p}^j = \begin{cases} \Delta\bar{p}^j_t = \bar{p}^j_t - \bar{p}^j_{t-1} & \text{if } i \in \text{Informed} \\ \Delta\bar{p}^j_{t-1} = \bar{p}^j_{t-1} - \bar{p}^j_{t-2} & \text{if } i \in \{\text{Baseline, Uninformed}\}. \end{cases}$$

The dependent variable is the change in reported confidence from period t to $t+1$, and μ^i is a participant-level random effect. The constant $\bar{\mu}$ is the object of interest: it is the average per-period change in confidence, so that a positive value indicates confidence rising over the course of the game.[18] ε^i_t is an idiosyncratic error, and standard errors are clustered at the session level.

The specification controls for two quantities that co-move with confidence. The first is the change in the participant's own contribution, $p_{t+1} - p_t$. Whether contributions and confidence are related is the question this paper sets out to answer, so we hold the contribution change fixed when estimating the trend in confidence. The second is the change in group members' contributions, $\Delta\bar{p}^j$, since conditional cooperation makes others' behaviour a determinant of a participant's own contribution and hence, potentially, of her confidence in it. Here $\bar{p}^j$ denotes $\bar{p}^j_t$ for Informed participants and $\bar{p}^j_{t-1}$ otherwise, reflecting the informational asymmetry confirmed in Manipulation Check.

Table 1 Column (1) reports the results. The estimated per-period change in confidence is $\hat{\bar{\mu}} = -0.02$, so that over the ten periods of the game confidence declines by 0.2 percentage points in total. The point estimate is not merely insignificant but negative, and its magnitude is negligible against the 22-point shortfall we are trying to explain. Nor is this a matter of imprecision: the 95% confidence interval, $[-0.32, 0.29]$, is consistent with a cumulative increase of at most 2.6 points across the nine transitions, around one-eighth of the shortfall we observe.

In Table B.1 Column (1), we augment the specification with interactions between the regressors and treatment indicators, yields a small negative trend rather than a zero one. The two are economically equivalent: the point estimate implies a cumulative decline of 3 percentage points across all ten periods. No statistically significant heterogeneity is detected across treatments.

In conclusion, neither specification shows the increase the encoding account predicts.

Result 3. *Reported confidence does not rise with repetition, so that confidence does not reflect encoding uncertainty about internal input.*

[18] We also estimate the specification with period dummies in place of the constant,

$$c^i_{t+1} - c^i_t = \mu^i + \sum_{s=2}^{9} \delta_s \, \mathbf{1}\{t = s\} + \gamma_1 \Delta\bar{p}^j + \gamma_3 (p_{t+1} - p_t) + \varepsilon^i_t,$$

allowing the per-period change to differ across periods rather than imposing a constant drift. Figure A.4 plots the predicted change in confidence for each period: eight of the nine transitions are indistinguishable from zero, and only one (period 8) is estimated above it, with no cumulative pattern across the game. Estimated separately by treatment, the period dummies are jointly indistinguishable from zero in all three (Informed: $\chi^2(3) = 0.63$, $p = 0.89$; Uninformed: $\chi^2(3) = 1.84$, $p = 0.61$; Baseline: $\chi^2(2) = 0.23$, $p = 0.89$).

This also rejects the trial-and-error account set out in Section 4.1.2, which predicts the same pattern. Neither part of it holds: contributions decline over the game but remain positive in the final period (t-test $p < 0.01$ in all treatments), and confidence is flat throughout.

6.1.5 Optimisation Uncertainty

Results 2 and 3 reject Hypotheses 1 and 2, while reported confidence remains 22 percentage points below its maximum in every treatment. By the argument given in Section 4.1.3, what remains is optimisation uncertainty.

Hypothesis 3 makes this more than an inference by elimination: if the shortfall reflects optimisation uncertainty, contributions should exhibit meaningful period-to-period variability even after controlling for group members' contributions, the signature of experimentation rather than passive noise around a fixed target. We therefore estimate

$$|p^i_{t+1} - p^i_t| = \mu^i + \beta_1 c^i_t + \gamma_2 |\Delta \bar{p}^j| + \varepsilon^i_t, \tag{5}$$

with $\Delta \bar{p}^j$ defined as in (4). The constant μ^i is the revision at zero confidence when others' contributions do not change, and β_1 is the change in that revision per point of confidence. Since we have argued that confidence tracks optimisation uncertainty, μ^i is the search under full optimisation uncertainty, and $100\beta_1$ is the part of the revision that depends on confidence. We control for $|\Delta \bar{p}^j|$, since conditional cooperation predicts that a participant adjusts her contribution when her group's contributions change, and this adjustment is not itself evidence of experimentation.

Table 1 Column (4) reports the results. At zero confidence, which we have argued reflects full optimisation uncertainty, participants revise their contributions by 21.2 percentage points on average ($p < 0.001$) when others' contributions do not change. The revision falls with confidence ($\hat{\beta}_1 = -0.12$) but remains about 9 percentage points at full confidence. Given Results 2 and 3, we read the roughly 12 points that move with confidence as variability associated with optimisation uncertainty, and the remainder as variability that does not depend on confidence, such as execution noise. A revision of this size at zero confidence, controlling for group behaviour, is consistent with participants trying different contribution levels rather than reporting a noisy but fixed action, although, as we discuss below, we cannot establish that it reflects search.[19]

Result 4. *We reject both encoding-based hypotheses. Contributions continue to vary substantially from period to period, controlling for group behaviour: by about 21 percentage points at zero confidence and about 9 at full confidence. Reported confidence in this strategic setting therefore points*

[19] Table B.1 Column (3) augments the specification with interactions between the regressors and treatment indicators. The estimated revision at zero confidence, when others' contributions do not change, is 17.67 percentage points in Baseline, 20.25 in Informed ($p = 0.599$ for the difference from Baseline), and 25.51 in Uninformed ($p = 0.010$). All three are large and of a similar order of magnitude, so the finding holds within each treatment.

to optimisation uncertainty.

6.1.6 Choice Variability under Confidence

Hypothesis 4 asks whether confidence in our setting behaves as it does in the literature. The coefficient of interest in (5) is β_1: a negative value indicates that lower confidence predicts larger revisions, as the literature reports.

Table 1 Column (4) reports the results. The estimated $\hat{\beta}_1 = -0.12$ ($p < 0.001$) implies that a fall in confidence from its maximum to zero is associated with a revision 12 percentage points larger, an increase of well over half the baseline established in Result 4. Confidence and subsequent revision are therefore related in our setting in the direction the literature reports.[20] What the association establishes is that the unincentivised confidence reports are informative and predict subsequent behaviour, rather than being noise.

Result 5. *Lower confidence predicts larger subsequent revisions to contributions.*

6.2 Is Low Confidence Associated with Distorted Contributions?

6.2.1 Does Low Confidence Compress Contributions?

We now test Hypothesis 5: the signature finding of the metacognitive uncertainty literature, that choices made under low confidence are less elastic with respect to fundamentals. In our setting the relevant fundamental is others' behaviour, since conditional cooperation makes group members' contributions a central determinant of a participant's own. We estimate

$$p_t^i = \mu^i + \beta_1 c_t^i + \beta_4 \bar{p}^j + \beta_5 \left(c_t^i \times \bar{p}^j\right) + \varepsilon_t^i, \tag{6}$$

where contributions and reported confidence are both measured on a scale from 0 to 100, and

$$\bar{p}^j = \begin{cases} \bar{p}_t^j & \text{if } i \in \text{Informed} \\ \bar{p}_{t-1}^j & \text{if } i \in \{\text{Baseline, Uninformed}\}. \end{cases}$$

The literature's finding implies $\beta_5 > 0$, so that

[20] Again, Table B.1 Column (3) augments the specification with interactions between the regressors and treatment indicators. Neither interaction with confidence is statistically distinguishable from zero ($p = 0.58$ and $p = 0.30$), so the relationship between confidence and subsequent revision does not differ across treatments. Uninformed participants revise their contributions more on average and respond somewhat less strongly to changes in their group's contributions. Nevertheless, the coefficient on confidence is stable across specifications (-0.10 against -0.12 in the parsimonious specification here).

$$\left.\frac{\partial p_t^i}{\partial \bar{p}^j}\right|_{c_t^i=0} < \left.\frac{\partial p_t^i}{\partial \bar{p}^j}\right|_{c_t^i=100}.$$

Column (5) of Table 1 reports the estimates. Contributions respond strongly to others' contributions ($\hat{\beta}_4 = 0.25$, $p = 0.008$), confirming conditional cooperation. The interaction, however, is small and imprecisely estimated: $\hat{\beta}_5 = 0.0007$ ($p = 0.52$), so that moving from zero to full confidence changes the slope by $100 \times \hat{\beta}_5 = 0.07$, against a base slope of 0.25.[21] We therefore find no evidence of the compression the literature documents.

What compression requires. The literature models an observed decision as a weighted combination of two quantities, with the weight given by reported confidence (Enke and Graeber, 2023; Charles et al., 2024): the *ideal*, $p^*_{i,t}$, the choice a participant makes under full confidence, and the *default*, $p^o_{i,t}$, the choice she makes under zero confidence. Writing confidence as a proportion of its maximum,

$$p_{i,t} = \frac{c_{i,t}}{100}\, p^*_{i,t} + \left(1 - \frac{c_{i,t}}{100}\right) p^o_{i,t}. \tag{7}$$

Our manipulation check shows that observed contributions respond to $\bar{p}^j$. Since the observed contribution is a weighted average of the ideal and the default, that response could originate in either component, or in both. We therefore write both as functions of $\bar{p}^j$, with participant-specific coefficients, and let the estimation determine where the responsiveness lies:

$$\begin{cases} p^*_{i,t} = \theta^i_1 \bar{p}^j + \theta^i_0 \\ p^o_{i,t} = \phi^i_1 \bar{p}^j + \phi^i_0, \end{cases} \tag{8}$$

where θ^i_1 and ϕ^i_1 govern how the ideal and the default respond to group members' contributions, and θ^i_0 and ϕ^i_0 are their levels when those contributions are zero. Allowing these to vary across participants is consistent with Charles et al. (2024), who elicit the default directly and find that subject fixed effects account for the bulk of its variation. It departs from Enke and Graeber (2023), who estimate a single population-level default and find it to be intermediate ($\hat{d} = 0.43$ in choice under risk and 0.52 in belief updating), leaving no scope for the default to vary across participants or to respond to the fundamentals of the problem.

Substituting (8) into (7) yields an expression linear in $c_{i,t}$, $\bar{p}^j$ and their interaction, with the

[21] The 95% confidence interval on $\hat{\beta}_5$ implies a change in slope of between -0.15 and 0.30, which is wide relative to the base slope of 0.25; the estimate is uninformative about effects of that size. Table B.1 Column (4) augments the specification with interactions between all regressors and treatment indicators, and neither three-way interaction is statistically distinguishable from zero ($p = 0.78$ and $p = 0.16$).

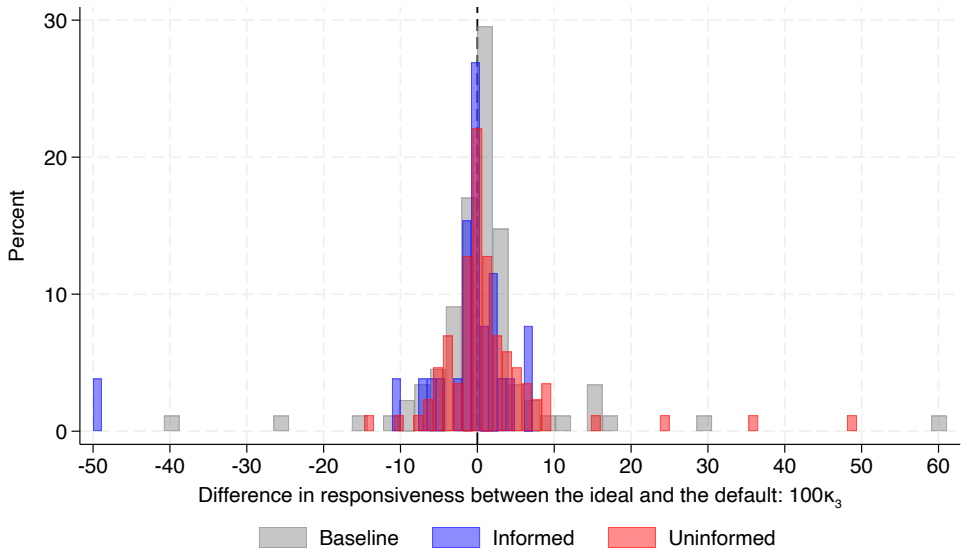


Figure 4. Distribution of $100 \times \hat{\kappa}_3^i$, the difference in responsiveness to group members' contributions between the ideal and the default by treatment. Each observation is one participant's coefficient on the interaction between reported confidence and group members' contributions in (10), estimated separately on that participant's own decisions. The dashed vertical line marks zero, at which the ideal and the default respond equally; positive values are those consistent with the assumption underlying the compression prediction. Estimates are shown for the 200 participants for whom all four coefficients are identified.

interaction term carrying the coefficient $\left(\theta_1^i - \phi_1^i\right)/100$.[22] Compression — a response to fundamentals that is weaker under low confidence than under high, so that $\partial^2 p_{i,t}/\partial \bar{p}^j \partial c_{i,t} > 0$ — therefore requires

$$\theta_1^i > \phi_1^i : \tag{9}$$

the ideal must respond more strongly to others' contributions than the default does. This is what β_5 in (6) estimates, pooled across participants. We now test Hypothesis 6 by estimating it for each participant separately.

Estimation. We estimate the specification in (6) on each participant's own decisions,

$$p_t^i = \kappa_0^i + \kappa_1^i c_t^i + \kappa_2^i \bar{p}^j + \kappa_3^i \left(c_t^i \times \bar{p}^j\right) + \varepsilon_t^i, \tag{10}$$

which recovers, for participant i, the *net ideal* — the ideal evaluated at a group contributing nothing, $\theta_0^i = \hat{\kappa}_0^i + 100\hat{\kappa}_1^i$; and the *net default*— the default evaluated at a group contributing nothing, $\phi_0^i = \hat{\kappa}_0^i$; along with their responsiveness to group behaviour, $\theta_1^i = \hat{\kappa}_2^i + 100\hat{\kappa}_3^i$ and

[22] In full, $p_{i,t} = \phi_0^i + \frac{c_{i,t}}{100}\left(\theta_0^i - \phi_0^i\right) + \phi_1^i \bar{p}^j + \frac{c_{i,t}\bar{p}^j}{100}\left(\theta_1^i - \phi_1^i\right)$.

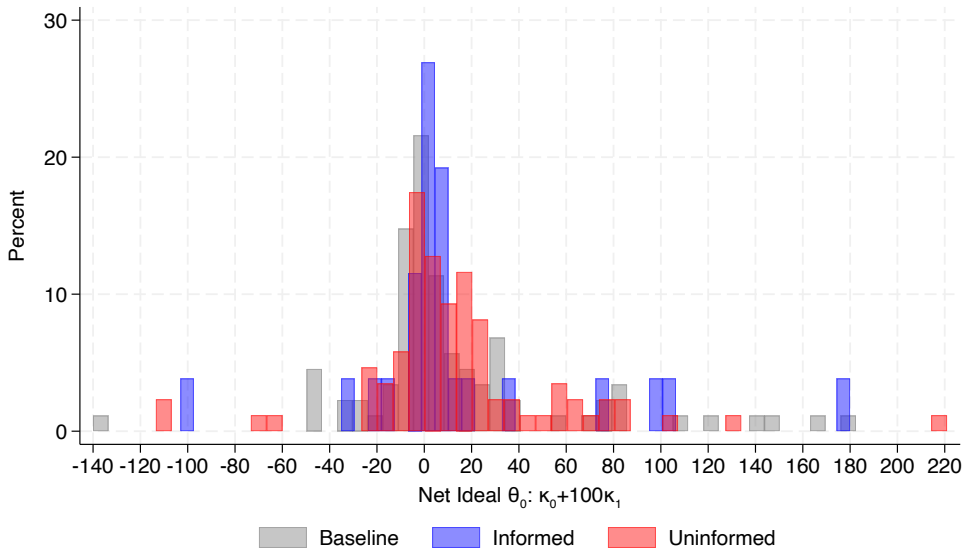


Figure 5. Distribution of the net ideal $\hat{\theta}_0^i$, by treatment. Each observation is one participant's contribution at full confidence evaluated at $\bar{p}^j = 0$, recovered from (10) estimated separately on that participant's own decisions. Estimates are shown for the 200 participants for whom all four coefficients are identified.

$\phi_1^i = \hat{\kappa}_2^i$[23]. Since $\theta_1^i - \phi_1^i = 100\hat{\kappa}_3^i$, condition (9) holds for participant i if and only if $\hat{\kappa}_3^i > 0$. We therefore examine the distribution of $\hat{\kappa}_3^i$ across participants.

Each participant input ten decisions, so the individual estimates are imprecise and we do not impose conventional significance thresholds on them. Doing so would set most estimates to zero and build the conclusion into the construction. The four coefficients in (10) are identified jointly, and require variation in both a participant's contributions and her reported confidence. They are not identified for 96 of 296 participants, whose contributions or reported confidence are constant or nearly so across periods. Our question is how confidence affects a participant's contribution, which cannot be answered for a participant whose contribution or whose confidence does not move. All individual-level analysis is therefore restricted to the 200 participants for whom all four coefficients are identified.

Figure 4 plots the distribution separated by treatment. The mean of $\hat{\kappa}_3^i$ is 0.008 and the median 0.002, with 53% of participants having a positive estimate. The difference is not statistically distinguishable from zero, by either a signed-rank test ($p = 0.22$) or a t-test ($p = 0.26$).[24] Nor is the magnitude informative: the 95% confidence interval admits a difference in responsiveness of anywhere between -0.6 and $+2.1$. Our test does not rule out that the default is substantially less responsive than the ideal; it only fails to find that it is.

Result 6. *We find no evidence that contributions made under low confidence are less responsive*

[23] This is clear from footnote 22 and (10).

[24] The same holds within each treatment: the signed-rank test does not reject in Baseline ($p = 0.34$, $N = 88$), Informed ($p = 0.42$, $N = 26$) or Uninformed ($p = 0.16$, $N = 86$).

to others' contributions than those made under high confidence. The assumption that generates compression is not supported in our data, as we find no evidence that the default is less responsive than the ideal.

6.2.2 Low-Confidence Contributions Are Not Uniformly Distorted

Compression is not the only way low confidence could distort observed choices. It might instead shift contributions in a common direction. Columns (1) and (2) of Table 1, estimated separately on Informed and on the remaining treatments, both show higher confidence associated with lower contributions: moving from zero to full confidence lowers contributions by 16 percentage points of the endowment among Informed participants and by 24 among the others. If such a shift were uniform across participants, choice-level confidence elicitation would be unnecessary. A researcher could recover preferences from observed choices by applying a single correction, without knowing how confident any individual was in any particular decision.

Whether the shift is uniform is an assumption of the specification rather than something we have established: (2), (3) and (6) all constrain β_1 to be identical across all 296 participants. We relax this by allowing participant-specific slopes:

$$p_t^i = \mu^i + \beta_1^i c_t^i + \beta_4 \bar{p}^j + \varepsilon_t^i, \tag{11}$$

with $\bar{p}^j$ defined as in (6), estimated by fixed effects with individual-specific slopes.

Column (6) of Table 1 reports the results. The average partial effect of confidence is -0.19, close to the estimate obtained under the homogeneity restriction. Its standard error, however, rises from 0.04 to 0.28, and the effect is no longer statistically distinguishable from zero ($p = 0.49$).[25] The same holds for the intercept: the estimated contribution at zero confidence and zero group contribution is 33.56, comparable to the estimates obtained under homogeneity, but likewise imprecise ($p = 0.17$). Together these suggest that confidence does not shift contributions in a common direction across participants.

Result 7. *The effect of confidence on contributions differs in sign across participants, and the average effect is not statistically distinguishable from zero once this heterogeneity is allowed for. Because confidence's effect is not uniform across participants, observed choices cannot be corrected*

[25] The standard error necessarily rises when a slope is estimated for each participant, so the loss of significance need not by itself indicate heterogeneity: it may reflect only the imprecision of estimating 296 slopes from at most ten observations each. Figure A.7 plots the distribution of the participant-specific slopes and shows that this is not the explanation. The estimates differ in sign as well as in magnitude: 21% of participants contribute more when they report higher confidence rather than less, and individual slopes range from -4.85 to 3.90, an order of magnitude wider in either direction than the average partial effect of around 0.2, and the mass is concentrated near zero. Nor is the dispersion an artefact of estimating a slope from few observations per participant: allowing the effect of confidence to vary randomly across participants improves the fit of the model decisively (likelihood-ratio $\chi^2(2) = 123.29$, $p < 0.001$), so the variance of the confidence slope is not zero.

ex post for low confidence without choice-level confidence reports. Nor can a participant's final decision serve as a proxy for her most confident one: confidence does not rise over the course of the game (Result 3), so later choices are not systematically better calibrated than earlier ones.

6.3 The True Prosocial Preference

We now test Hypothesis 7 using the net ideal, $\hat{\theta}_0^i$, recovered in Section 6.2.1: the contribution a participant would make at full confidence were her group members to contribute nothing.

The net ideal is positive. Figure 5 plots the distribution of net ideal. The net ideal is positive for 61% of participants, with a median of 3.8 and a mean of 14.4, and is significantly greater than zero (signed-rank $p < 0.001$; t-test $p < 0.001$).[26]

A participant with a positive net ideal would contribute something in a group contributing nothing, at full confidence and with no scope for reciprocation. Positive contributions in our data are therefore not made out of confusion alone.

The net ideal does not vary across treatments. Median net ideals are 1.7 in Baseline, 2.4 in Informed and 8.0 in Uninformed, but these differences are not significant (Kruskal–Wallis $\chi^2(2) = 0.84$, $p = 0.66$; all pairwise rank-sum $p > 0.38$). Nor do the distributions differ in shape rather than in location (all Kolmogorov–Smirnov test $p > 0.3$). The same holds on the censored measure.

This is what should be observed if $\hat{\theta}_0^i$ measures preference. Our treatments vary what a participant knows about her group members' contributions, not what she wants to contribute, and random assignment should leave the three groups with the same preferences to begin with. That the recovered ideals are indistinguishable is therefore a check on the estimation: had they differed, $\hat{\theta}_0^i$ would be capturing something about the information environment rather than about the participant.

Result 8. *The net ideal does not differ across treatments, as should be the case if it measures preference. It is positive for* 61% *of participants and remains positive on average, so positive contributions are not motivated by confusion or conditional cooperation alone.*

[26] The distribution has long tails, since each estimate extrapolates from ten observations to a counterfactual in which group members contribute nothing — a situation participants encountered in 6.9% of periods. Some estimates consequently fall outside the $[0, 100]$ interval within which a contribution must lie. We therefore also repeat analyses on estimates censored to that interval, treating negative net ideals as zero and those above 100 as 100; the conclusion is unchanged.

7 Conclusion

We argue that the definition of confidence in psychology does not consider optimisation uncertainty in strategic games, i.e., difficulty acting on what one wants, as distinct from encoding uncertainty about the representations themselves — what one wants or what one faces. We show that the 22-point shortfall from maximal confidence is not explained by encoding uncertainty, since it does not fall once strategic and payoff uncertainty is removed or once repetition gives preferences time to sharpen, which leaves optimisation uncertainty; consistent with this, contributions continue to vary from period to period even holding group behaviour fixed, and more so at lower confidence.

This finding makes confidence elicitation particularly relevant for economics experiments, as optimisation uncertainty is likely to arise wherever incentive compatibility requires a participant to convert a preference into a specific, possibly unfamiliar, action. Our results also have a direct implication for how confidence should be used once elicited. We find no evidence that low confidence compresses contributions toward a common default, and its effect on contribution levels varies in sign across participants. A single population-level correction, as in the existing literature, would therefore misclassify many participants. Moreover, since confidence does not rise over the game, no single elicitation — early or late — can substitute for asking at the moment of decision.

We also show how choice-by-choice confidence can be put to use in a public goods game. We recover each participant's net ideal — her contribution at full confidence and absent any reciprocation. We find it positive for 61% of participants and stable across treatments that vary information. With choice-level confidence reports, we can conclude that positive contributions in public goods games are not made out of confusion alone.

References

Alaoui, L. and Penta, A. (2016). Endogenous depth of reasoning. *The Review of Economic Studies*, 83(4):1297–1333.

Andreoni, J. (1988). Why free ride? strategies and learning in public goods experiments. *Journal of Public Economics*, 37:291–304.

Andreoni, J. (1995). Cooperation in public-goods experiments: kindness or confusion? *The American Economic Review*, pages 891–904.

Andrikogiannopoulou, A. and Papakonstantinou, F. (2020). History-dependent risk preferences: Evidence from individual choices and implications for the disposition effect. *The Review of Financial Studies*, 33(8):3674–3718.

Bang, D. and Fleming, S. M. (2018). Distinct encoding of decision confidence in human medial prefrontal cortex. *Proceedings of the National Academy of Sciences*, 115(23):6082–6087.

Bossaerts, P., Ioannidis, K., Woods, R., and Yadav, N. (2024). How general equilibrium in markets with indivisible goods obtains at the cost of computational complexity. Technical report, Technical report. Cambridge Faculty of Economics Technical Report.

Burton-Chellew, M. N., El Mouden, C., and West, S. A. (2016). Conditional cooperation and confusion in public-goods experiments. *Proceedings of the National Academy of Sciences*, 113(5):1291–1296.

Camerer, C. F., Ho, T.-H., and Chong, J.-K. (2004). A cognitive hierarchy model of games. *The quarterly journal of economics*, 119(3):861–898.

Charles, C., Frydman, C., and Kilic, M. (2024). Insensitive investors. *The Journal of Finance*, 79(4):2473–2503.

Enke, B. and Graeber, T. (2023). Cognitive uncertainty. *The Quarterly Journal of Economics*, 138(4):2021–2067.

Enke, B., Graeber, T., and Oprea, R. (2025). Complexity and time. *Journal of the European Economic Association*, 23(5):1838–1867.

Fehr, E. and Gächter, S. (2000). Cooperation and punishment in public goods experiments. *American Economic Review*, 90(4):980–994.

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

Fleming, S. M. and Daw, N. D. (2017). Self-evaluation of decision-making: A general bayesian framework for metacognitive computation. *Psychological review*, 124(1):91.

Fleming, S. M., Ryu, J., Golfinos, J. G., and Blackmon, K. E. (2014). Domain-specific impairment in metacognitive accuracy following anterior prefrontal lesions. *Brain*, 137(10):2811–2822.

Frydman, C. and Jin, L. J. (2022). Efficient coding and risky choice. *The Quarterly Journal of Economics*, 137(1):161–213.

Geurts, L. S., Cooke, J. R., van Bergen, R. S., and Jehee, J. F. (2022). Subjective confidence reflects representation of bayesian probability in cortex. *Nature human behaviour*, 6(2):294–305.

Glimcher, P. W. (2022). Efficiently irrational: deciphering the riddle of human choice. *Trends in cognitive sciences*, 26(8):669–687.

Heinemann, F., Nagel, R., and Ockenfels, P. (2009). Measuring strategic uncertainty in coordination games. *The review of economic studies*, 76(1):181–221.

Houser, D. and Kurzban, R. (2002). Revisiting kindness and confusion in public goods experiments. *American Economic Review*, 92(4):1062–1069.

Hoven, M., Lebreton, M., Engelmann, J. B., Denys, D., Luigjes, J., and van Holst, R. J. (2019). Abnormalities of confidence in psychiatry: an overview and future perspectives. *Translational psychiatry*, 9(1):268.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied cognitive psychology*, 5(3):213–236.

Lu, X., Murawski, C., Bossaerts, P., and Suzuki, S. (2025). Estimating self-performance when making complex decisions. *Scientific Reports*, 15(1):3203.

Nagel, R. (1995). Unraveling in guessing games: An experimental study. *The American economic review*, 85(5):1313–1326.

O'Keefe, D. J. (2008). Elaboration likelihood model. *The international encyclopedia of communication*.

Oprea, R. and Vieider, F. M. (2024). Minding the gap: On the origins of probability weighting and the description-experience gap. Technical report, working paper, UC Santa Barbara.

Payzan-LeNestour, E. and Woodford, M. (2022). Outlier blindness: A neurobiological foundation for neglect of financial risk. *Journal of Financial Economics*, 143(3):1316–1343.

Petty, R. E. and Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. In *Advances in experimental social psychology*, volume 19, pages 123–205. Elsevier.

Polanía, R., Woodford, M., and Ruff, C. C. (2019). Efficient coding of subjective value. *Nature neuroscience*, 22(1):134–142.

Rademaker, R. L., Tredway, C. H., and Tong, F. (2012). Introspective judgments predict the precision and likelihood of successful maintenance of visual working memory. *Journal of vision*, 12(13):21–21.

Sacco, P. L. (2026). The metabolic logic of cognitive biases. *Trends Open*.

Schulz, L., Fleming, S. M., and Dayan, P. (2023). Metacognitive computations for information search: Confidence in control. *Psychological Review*, 130(3):604.

Stahl, D. O. and Wilson, P. W. (1995). On players models of other players: Theory and experimental evidence. *Games and Economic Behavior*, 10(1):218–254.

von Neumann, J. and Morgenstern, O. (1944). *Theory of Games and Economic Behavior*. Princeton University Press, Princeton, NJ.

Wang, G., Li, J., Wang, W., Niu, X., and Wang, Y. (2024). Confusion cannot explain cooperative behavior in public goods games. *Proceedings of the National Academy of Sciences*, 121(10):e2310109121.

Woodford, M. (2012). Prospect theory as efficient perceptual distortion. *American Economic Review*, 102(3):41–46.

Appendices

A Additional Figures

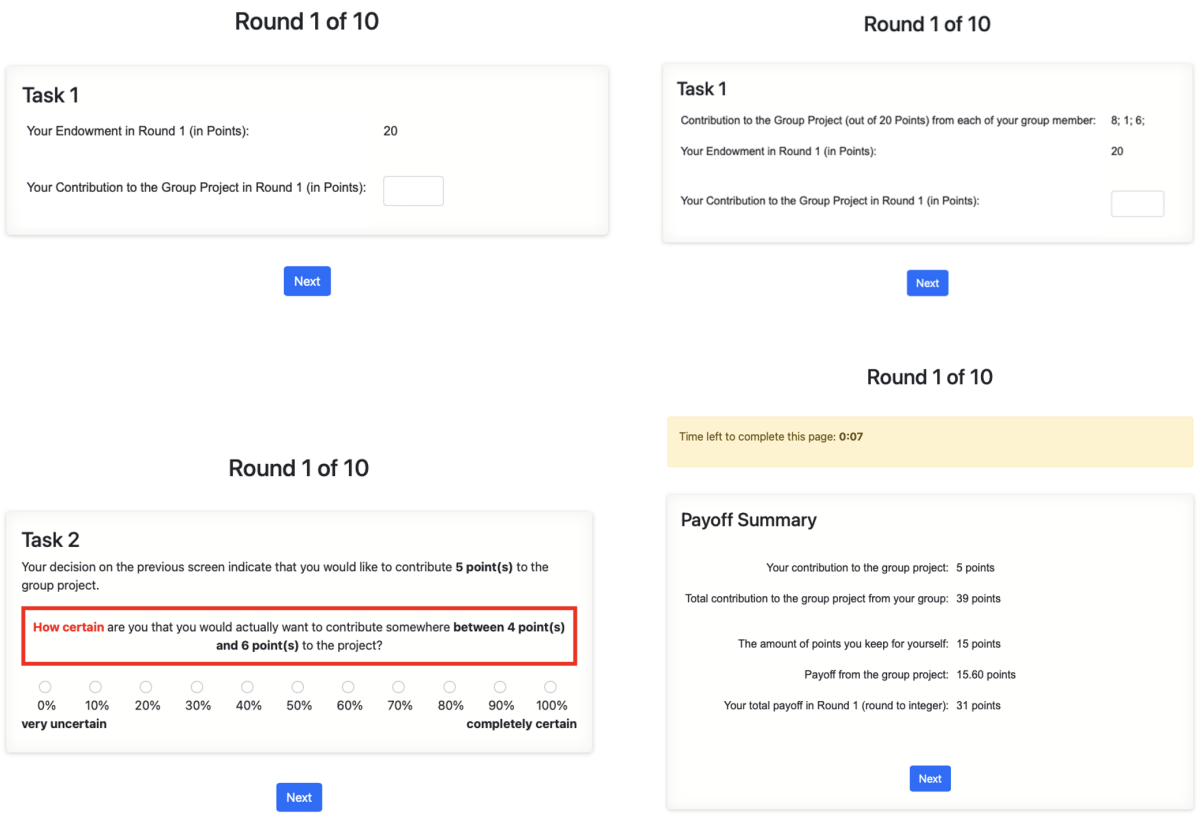


Figure A.1. Sample interface in contribution stage in *Baseline* and *Uninformed* (top left) and in *Informed* (top right). Confidence elicitation (bottom left) and the subsequent payoff interface (bottom right) are identical to all participants.

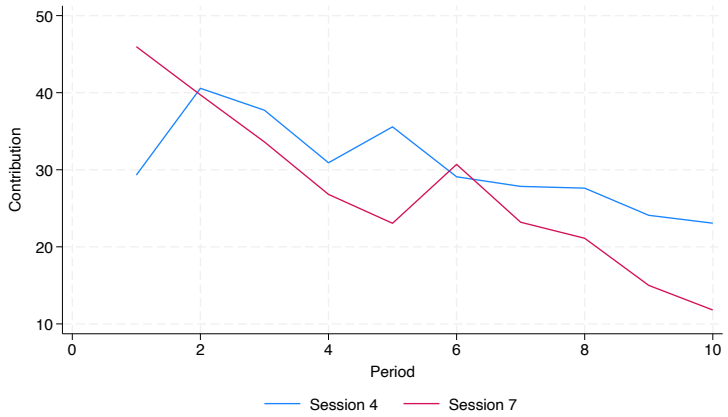


Figure A.2. Trajectory of average contributions from two representative sessions in Sequential PGG.

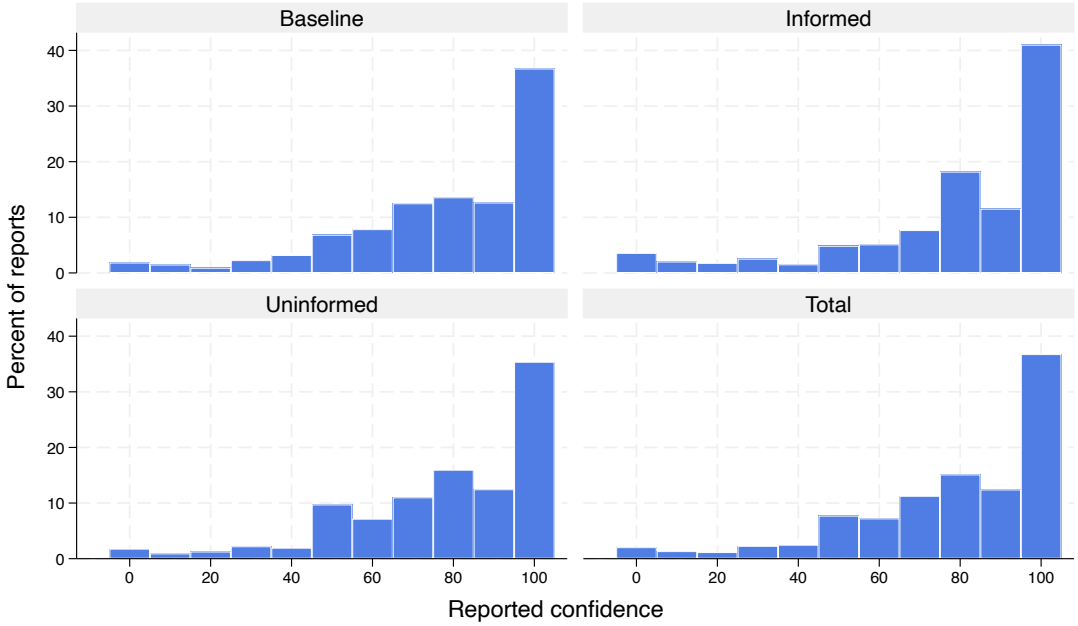


Figure A.3. Distribution of reported confidence, pooled and by treatment. Each panel shows the percentage of confidence reports at each of the eleven values of the 0–100 scale (in increments of ten), for Baseline, Informed and Uninformed participants and for the pooled sample (Total).

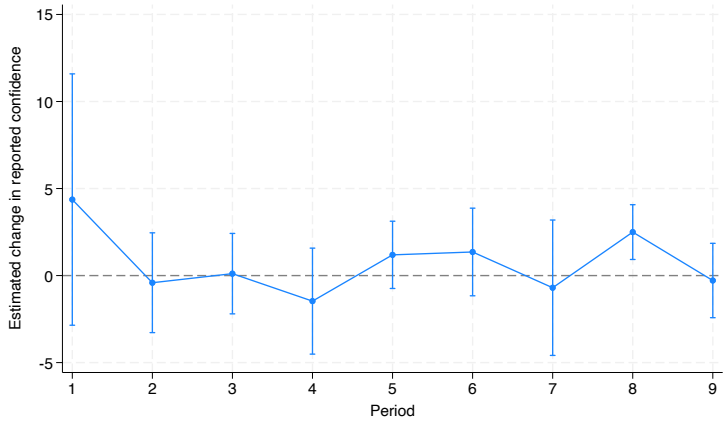


Figure A.4. Estimated change in reported confidence by period. Predictive margins from (4) estimated with additional period dummies, shown with 95% confidence intervals. Period t denotes the change in confidence from period t to period $t+1$, so that the nine points correspond to the nine transitions of the ten-period game. Standard errors are clustered at the session level.

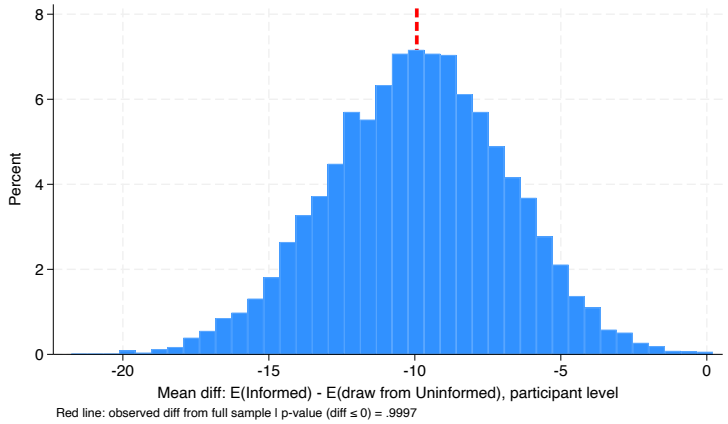


Figure A.5. Empirical distribution of the permutation test for mean contribution differences across the two groups in Sequential PGG. Specifically, the group with more participants (Uninformed) is randomly subsampled so that its sample size matches that of the smaller group (Informed). The mean contribution levels are then compared and recorded. This procedure is repeated 10,000 times. The p-value is computed as the proportion of repetitions in which the mean difference is less than or equal to zero. Dotted red line represents the mean difference using full sample from the two groups.

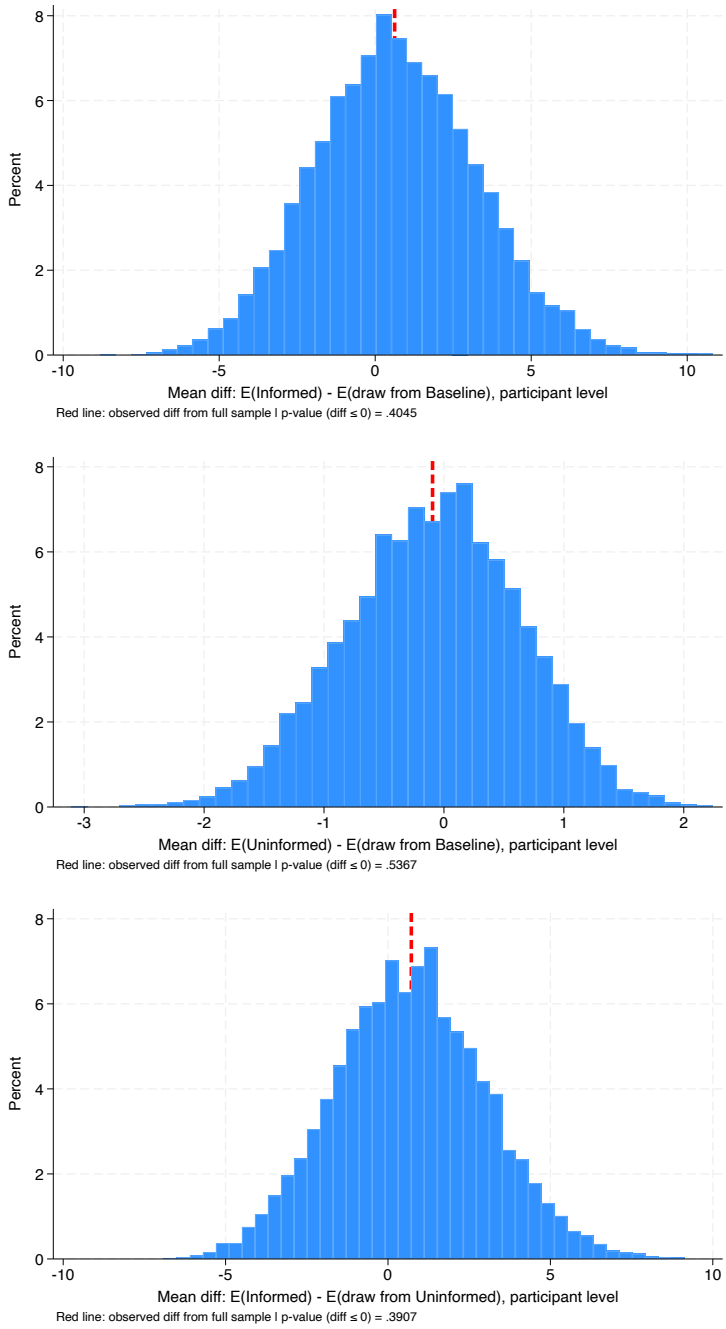


Figure A.6. Empirical distribution of the permutation test for mean confidence differences across the three treatment groups. In each test, the treatment group with more participants is randomly subsampled so that its sample size matches that of the smaller group. The mean confidence levels are then compared and recorded. This procedure is repeated 10,000 times. The p-value is computed as the proportion of repetitions in which the mean difference is less than or equal to zero. Dotted red line represents the mean difference using full sample from the two treatment groups.

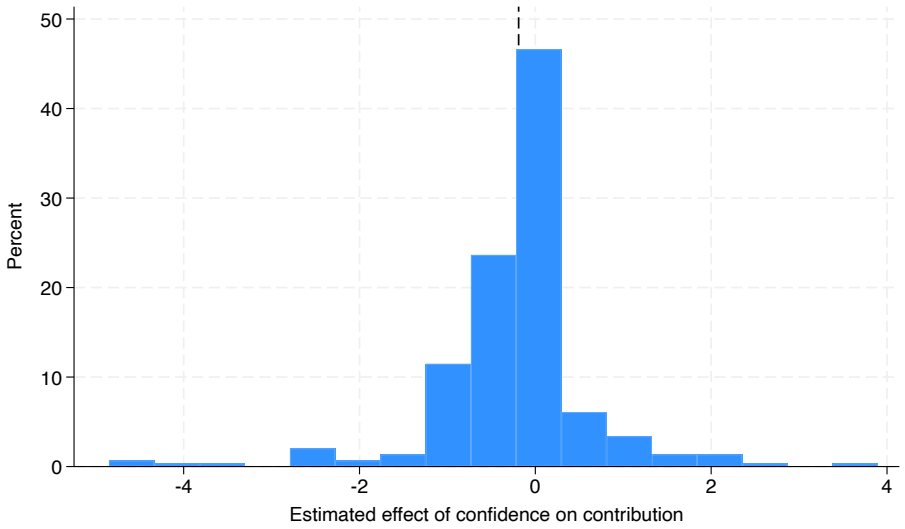


Figure A.7. Distribution of participant-specific effects of confidence on contributions. Each observation is the coefficient on reported confidence from a separate regression of contributions on confidence and group members' contributions, estimated on that participant's own decisions ($N = 296$ participants, ten decisions each). The dashed vertical line marks the average partial effect of -0.19 reported in Column (6) of Table 1. The median is -0.03 and the interquartile range $[-0.50, 0.00]$.

B Additional Tables

Table B.1. Linear modelling of dependent variable on variables of interest and controls, additionally allowing them to interact with treatment indicators, with random effects (intercept) at participant level in Models (1)–(4). Each model extends its corresponding specification in Table 1 by additionally allowing all variables to interact with treatment indicators. Estimates of control variables and their interactions with treatment are omitted for brevity. Robust standard errors clustered at session level ($N = 7$) in parentheses. p-values in brackets.

	(1) p_t	(2) $c_{t+1} - c_t$	(3) $\lvert p_{t+1} - p_t \rvert$	(4) p_t
Variable of Interest				
Constant, $\hat{\mu}$	31.53	-0.34	17.67	31.40
	(3.07)	(0.13)	(2.85)	(3.48)
	[0.000]	[0.008]	[0.000]	[0.000]
c_t, β_1	-0.22		-0.10	-0.21
	(0.04)		(0.03)	(0.06)
	[0.000]		[0.004]	[0.000]
$\bar{p}^j$, β_2	0.30			0.36
	(0.03)			(0.14)
	[0.000]			[0.010]
$c_t \times \bar{p}^j$, β_3				-0.00
				(0.00)
				[0.689]
Treatment Interaction with Variable of Interest				
Informed, δ_1	-3.60	0.51	2.58	-15.05
	(3.03)	(0.54)	(4.89)	(5.89)
	[0.235]	[0.350]	[0.599]	[0.011]
Uninformed, δ_2	7.10	0.56	7.84	15.71
	(2.45)	(0.37)	(3.03)	(4.46)
	[0.004]	[0.126]	[0.010]	[0.000]
Informed $\times c_t^i$, δ_3			-0.02	0.14
			(0.04)	(0.10)
			[0.585]	[0.135]
Uninformed $\times c_t^i$, δ_4			-0.04	-0.11
			(0.04)	(0.07)
			[0.303]	[0.139]
Informed $\times \bar{p}^j$, δ_5	0.04			0.12
	(0.08)			(0.24)
	[0.657]			[0.610]
Uninformed $\times \bar{p}^j$, δ_6	-0.02			-0.32
	(0.04)			(0.23)
	[0.727]			[0.168]
Informed $\times \bar{p}^j \times c_t^i$, δ_7				-0.00
				(0.00)
				[0.777]
Uninformed $\times \bar{p}^j \times c_t^i$, δ_7				0.00
				(0.00)
				[0.160]
Control and Treatment Interaction with Controls	Yes	Yes	Yes	Yes
BIC	24594.67	21498.13	21331.7	24624.3
Observations	2,702	2,406	2,406	2,702
Number of participant	296	296	296	296

Table B.2. Linear modelling of dependent variable on variables of interest and controls, allowing for random effects (intercept) at participant level in Models (1)–(5). Robust standard errors clustered at participant level ($N = 296$) in parentheses. p-values in brackets.

Model	(1)	(2)	(3)	(4)	(5)
Dependent Variable	p_t	p_t	$c_{t+1} - c_t^i$	$\lvert p_{t+1} - p_t \rvert$	p_t
	Informed	Baseline, Uninformed			
Variable of Interest					
$\bar{\mu}$	21.21	36.25	-0.02	21.18	35.29
	(8.45)	(3.08)	(0.19)	(2.43)	(4.20)
	[0.012]	[0.000]	[0.923]	[0.000]	[0.000]
c_t^i, β_1	-0.16	-0.24		-0.12	-0.23
	(0.08)	(0.05)		(0.03)	(0.05)
	[0.033]	[0.000]		[0.000]	[0.000]
$\bar{p}_t^j$, β_2	0.32				
	(0.09)				
	[0.000]				
$\bar{p}_{t-1}^j$, β_3	0.09	0.30			
	(0.07)	(0.02)			
	[0.176]	[0.000]			
$\bar{p}^j$, β_4					0.25
					(0.11)
					[0.029]
$(c_t^i \times \bar{p}^j)$, β_5					0.00
					(0.00)
					[0.592]
Control					
$\Delta\bar{p}^j$, γ_1			0.05		
			(0.02)		
			[0.015]		
$\lvert\Delta\bar{p}^j\rvert$, γ_2				0.09	
				(0.03)	
				[0.004]	
$p_{t+1} - p_t$, γ_3			-0.20		
			(0.04)		
			[0.000]		
Interaction with Treatment	No	No	No	No	No
Wald p-value	0.000	0.000	0.000	0.000	0.000
Observations	351	2,312	2,406	2,406	2,702
Number of participants	39	257	296	296	296

C Instruction

Below are the instructions. The section titled "Additional Information" is shown only to participants in the Full Information. Participants in the Baseline will not see this section. All other parts of the instructions are identical across the two treatment groups.

There are 10 rounds of games.

In each round, you will be randomly assigned to interact in a group with 3 of the other participants.

There are two tasks in each round.

Task 1

Each of the 4 participants in your group will receive an endowment of 20 points.

Every one of you must decide how much of your endowment of 20 points to contribute to the "group project". The contribution must be an integer between 0 to 20 (including 0 and 20).

All points contributed to the "group project" are multiplied by 1.6 and then split evenly among the 4 group members.

Therefore, in each round, your payoff will be the sum of the part of your 20 points endowment you keep for yourself, plus your share of the group project, i.e.,

Your Payoff in Points
$= (20 - \text{your contribution to group project}) + 0.4 \times (\text{total contributions to the group project})$

Additional Information

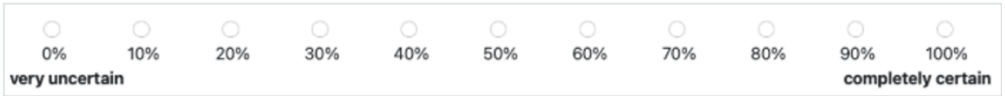
You have been randomly allocated to the "Full Information" treatment in this experiment.

This means that you are able to see all your group members' contributions to the group project in each round before you make your own contribution decision.

Please note that all other members of your group cannot see others' contributions like you can in each round. Furthermore, they are unaware that you are able to view their contributions.

Task 2

Next, you will indicate how certain you are that you would actually want to contribute to the group project at the amount you indicate.



Result

After all members have made their decisions, you will be able to view the information about your payoff from the current round.

Random Grouping in Each Round

Then, the next round will begin, with the composition of the groups changing for each round, which means that you will be playing with different participants in every round.

Payoff

At the end of the experiment, the sum of your 10 payoffs will be converted to Singapore dollars (SGD) at a rate of 1 Point = 0.025 SGD, which is your final earning (that excludes show-up fee) for the experiment.

You will be asked some demographic questions at the end of the experiment. Then, you will be paid in a private manner.

Control question 1/5

1. Each of the 4 group members has an endowment of 20 points. Suppose nobody (including you) contributes any points to the group project.

What is your payoff (in points)? _____

What is the payoff of each of the other group members (in points)? _____

Hints:

Your Payoff
= (20 − your contribution to group project) + 0.4 × (total contributions to the group project)

Control question 2/5

2. Each of the 4 group members has an endowment of 20 points. Suppose you contribute 20 points to the group project. All other 3 group members each contribute 20 points to the group project.

What is your payoff (in points)? _____

What is the payoff of each of the other group members (in points)? _____

Hints:

Your Payoff
= (20 − your contribution to group project) + 0.4 × (total contributions to the group project)

Control question 3/5

3. Each of the 4 group members has an endowment of 20 points. Suppose the other 3 group members contribute a total of 30 points to the group project.

What is your payoff (in points) if you do not contribute any points to the group project? _____

What is your payoff (in points) if you contribute 15 points to the group project? _____

Hints:

Your Payoff
= (20 − your contribution to group project) + 0.4 × (total contributions to the group project)

Control question 4/5

4. Each of the 4 group members has an endowment of 20 points. Suppose you contribute 8 points to the group project.

What is your payoff (in points) if all other group members together contribute a total of 7 points to the group project? _____

What is your payoff (in points) if all other group members together contribute a total of 22 points to the group project? _____

Hints:

Your Payoff
$= (20 - \text{your contribution to group project}) + 0.4 \times (\text{total contributions to the group project})$

Control question 5/5

5. Suppose you are 80% certain that you would actually want to contribute to the group project the amount you indicate. Which button should you click in this case?

