Incentives and Payment Mechanisms in Preference Elicitation

Drichoutis, Andreas C. and Palma, Marco and Feldman, Paul

Agricultural University of Athens, Texas AM University

8 May 2024

Incentives and Payment Mechanisms in Preference Elicitation[*]

Andreas C. Drichoutis[†1], Marco A. Palma[‡2], and Paul Feldman[§2]

[1]Agricultural University of Athens
[2]Texas A&M University

Abstract: Previous literature analyzing the effects of incentive compatibility of experimental payment mechanisms is dominated by theory. With overwhelming evidence of theory violations in a multiplicity of domains, we fill this gap by empirically exploring the effects of different payment mechanisms in induced preference elicitation using a large sample of over 3800 participants across three experiments. In Experiment 1, we collected responses for offer prices to sell a card like in Cason and Plott (2014), systematically varying on a between-subjects basis the way subjects received payments over repeated rounds, by either paying for all decisions (and various modifications) or just one, as well as making the payments certain, probabilistic or purely hypothetical. While we find that the magnitude of the induced value and the range of the prices used to draw a random price significantly affect misbidding behavior, neither the payment mechanism nor the certainty of payment affected misbidding. In Experiment 2, we replaced the BDM mechanism with a second price auction and found similar results, albeit less misbidding rates. In Experiment 3, we examine the effect of payment mechanisms on choice under risk and find portfolio effects (i.e., paying all rounds) when the lottery pairs do not involve options with certainty. Overall, our empirical exercise shows that payment mechanism design considerations should place more weight on the choice architecture rather than on incentive compatibility.

Keywords: Becker-DeGroot-Marschak mechanism, second price auction, risk choices, preference elicitation, choice architecture

[*] Part of this work was completed while Andreas Drichoutis was a Fulbright Visiting Scholar at Texas A&M University. We would like to thank Tim Cason and Uri Gneezy for helpful comments. We received financial support for this project from the Institute for Advancing Health through Agriculture at Texas A&M University.

[†] Associate Professor, Department of Agricultural Economics & Rural Development, School of Applied Economics and Social Sciences, Agricultural University of Athens, Iera Odos 75, 11855, Greece, e-mail: adrihout@aua.gr.

[‡] Professor and Director Human Behavior Laboratory, Department of Agricultural Economics, Texas A&M University, College Station, TX 77843 USA, tel:+1-9798455284 e-mail: mapalma@tamu.edu.

[§] Assistant Professor, Department of Agricultural Economics, Texas A&M University, College Station, TX 77843 USA, tel:+1-9798455284 e-mail: paul.feldman@tamu.edu.

JEL codes: C80, C91, D44; **AEA RCT Registry:** #0009687

1 Introduction

Incentives should matter. This imperative is a core tenet of economics. Insufficient or inadequate incentives can lead to deviations in behavior from what is predicted by economic and behavioral models. Economic experiments often involve multiple decisions of the same task or they incorporate multiple tasks within the same study. Correspondingly, researchers must weigh the potential tradeoffs between different payment mechanisms, the size of incentives, potential payoff externalities, and budget constraints. Paying for every decision increases costs and may induce portfolio and wealth effects; however, these effects can be mitigated by only paying for one random choice. Notwithstanding, paying for one random choice can dilute incentives as the number of choices increases (Beattie and Loomes, 1997; Charness et al., 2016). To complicate matters, random incentive schemes may fail to be incentive-compatible, exhibit menu dependence, and induce risk preferences even in purely deterministic settings. This shortcoming of random incentives has led some researchers to argue for only collecting preferences over a single choice(e.g., Cox et al., 2015; Harrison and Swarthout, 2014) despite how restrictive this approach might be. Overall, the optimal incentive scheme is rarely transparent, making it difficult to develop general guidelines.

Most research on adequate incentives is theoretical and focuses on incentive compatibility (see Charness et al., 2016, for a review). The lack of empirical interest, even from experimental economists (Azrieli et al., 2018), suggests that incentive compatibility concerns may be overstated. More troubling, perhaps, is the clever sleight of hand concerning the unobservability of preferences over binary alternatives under non-expected utility models. Under non-expected utility, the path dependence of choice behavior cannot be assumed away; hence, any binary choice encodes both preference information and the history of all previous choices (Machina, 1989). Further yet, incentive compatibility has little explanatory power for plausible mistakes or heuristics. How, then, can one elicit true preferences and discriminate between different preference incentive mechanisms?

Our study employs the simplest application of *induced values* (Smith, 1976), providing a straightforward empirical testing approach: the true value of money is fixed. Consequently, for all of our extensive experimental treatments, the objective preferences over monetary amounts are clearly defined in terms of their monetary equivalents.[1] Thus, we can focus primarily on the question from the empirical side of the proper incentive scheme. That is, on the effectiveness of different incentive schemes to recover truthful individual preferences from the (induced) preferences. In a nutshell, the true monetary value of $2 is $2, as used by Cason and Plott

[1] This result requires the preposterous assumption that our subjects prefer more money to less.

(2014).

In a sample of these valuation tasks and choices under risk, we consider, broadly, the following three dimensions of incentive mechanisms: 1) the size of the prize, 2) the chance with which payment is determined, and 3) the chance each choice counts towards experimental earnings. In the valuation tasks, we also vary the range of values people can assign to a sure monetary amount and whether the uncertainty determining earnings is strategic. To further understand the mechanisms underlying our results and to validate our online subject pool, we repeat several of our treatments with binary lotteries by replicating a canonical paper in the literature on the effect of incentive schemes over choice under risk (Cox et al., 2015).

In a collective sample of over 3,800 subjects, we find that the magnitude of the incentives and the framing of the problem matter, while chances and reward correlation structure do not. Higher stated monetary rewards result in better-calibrated valuations, while larger value ranges (greater opportunity for deviations) lead to higher deviations from the true value of a dollar. Strikingly, the lack of variation according to the chance of the rewards extends to cases where the rewards are purely hypothetical adding to the literature that finds minimal or no differences between real and hypothetical stakes (e.g., Brañas Garza et al., 2023; Enke et al., 2023; Gneezy et al., 2015; Hackethal et al., 2023; Irwin et al., 1992; Li et al., 2017).[2] One reason why higher incentives may not yield any meaningful effects is that the evaluated tasks (preferences elicitations) are not cognitively demanding or that higher cognitive effort may not lead to a meaningful difference in behavior. Holt and Laury (2002) may be the example that proves the rule. Their risk task exhibits a hypothetical bias and is also cognitively demanding; moreover, cognitively simplifying the task leads to better measures (Charness et al., 2018). We also find strategic uncertainty produces better-calibrated valuations; however, different rewards' chances again play no meaningful role in overall bidding behavior. Although misunderstanding may be a factor explaining misbehavior (Serizawa et al., 2024), we doubt this is driving the results in our case because of all the strict measures to ensure that subjects were attentive, paid attention to the instructions, and understood the procedures.

Our results align with previous findings on risk preferences and incentive schemes. Although they are qualitatively similar to Cox et al. (2015), we do find that subjects are less responsive to different incentive schemes. Specifically, we find no differences when a sure amount of money is available. However, our subjects select safer alternatives if all choices are incentivized and all available choices are uncertain. We conjecture that this certainty effect is driven by choices between non-degenerate lotteries being more complex and thus more likely to be affected by increases in cognitive effort. Consistent with our hypothesis of avoidance of complex lottery

[2] In contrast, Blavatskyy et al. (2022) find that the Allais Paradox is likely to be observed in experiments with, among others, high hypothetical payoffs. Perhaps because most experimentalists could not afford Allais' level of incentives, even if not adjusting for inflation and only recruiting one subject.

choices, a majority of our participants opt for the riskless alternative when one is available.

The paper proceeds as follows. The next section reviews the relevant literature to set the context and motivate our study. We then present the three experiments sequentially by describing the methods and summarizing results from each experiment. We conclude in the final section.

2 Related Literature

This section explores the intricacies of incentives in experimental economics, particularly focusing on value elicitation experiments and choices under risk. It examines the impact of various incentive schemes on participant behavior, exploring how these factors influence bidding in auctions and choices over lotteries. The subsections highlight key studies and findings in the field, revealing the complexities and debates surrounding effective incentive design, the role of cash balances, and the effectiveness of different incentivization strategies. This overview offers insights into how experimental setups and incentive mechanisms can significantly affect economic behavior and decision-making processes.

2.1 Incentives in value elicitation experiments

In an early investigation of the winner's curse in common value auctions, Kagel and Levin (1986) generated attention and subsequent discussions regarding payment mechanisms that pay for all decisions in a series of auction periods. Participants in Kagel and Levin (1986) suffered from a winner's curse: bids often exceeded the conditional expected value of the item sold. Hansen and Lott (1991) proposed that deviations from the theoretical bidding equilibrium reported in Kagel and Levin (1986) may have been a perfectly rational response to limited liability and low cash balances (i.e., accumulated earnings as a result of paying multiple rounds). That is, bids that would result in greater losses than available cash balances could lead to higher bidding due to a lack of responsibility. These two studies revealed the importance of cash balances when paying for all decisions in the context of these experiments and their potential to affect bidding behavior in common value auctions and sparked a heated debate around payment mechanisms.

In response, Kagel and Levin (1991) conducted a follow-up experiment that ensured subjects had sufficient cash balances so that deviations from the predicted (risk-neutral) Nash equilibrium could not be explained by the limited liability arguments and still obtained significant overbidding. Ham et al. (2005) argued that cash balances may also affect bidding behavior in private value auctions, and to address this concern, they introduced exogenous variation in cash balances by randomly assigning additional payments while subjects bid in a first price auction.

They found that cash balances also play a statistically significant role in bidding behavior in private value auctions.

While cash balance incentives can be avoided by paying for only one randomly selected trial, Ham et al. (2005) further noted its impact on subjects' incentives, potentially diluting payoffs in two ways. First, expected payoffs can be a function of the compounded probability of a trial being selected multiplied by the payoffs for that trial, which may dilute incentives with an increased number of trials and/or smaller payoffs per trial. Second, since there is only one bidder with earnings in many auction formats, as in a first or a second price auction, effective recruitment of subjects can only be achieved with large fixed show-up fees, which may render the incentives associated with the auctions trivial.

Another strand of the literature has focused on Between-Subject Random Incentive Schemes (BRIS), where only a subset of subjects are randomly selected to realize their decisions and receive a payment (Baltussen, Post, van den Assem, and Wakker, Baltussen et al.). BRIS has been investigated in several domains including fairness (Bolle, Bolle), risk choice (Baltussen, Post, van den Assem, and Wakker, Baltussen et al.) and donations in dictator games (Clot, Grolleau, and Ibanez, Clot et al.). More recently, Ahles et al. (2024) found that a 10% and 1% payment probabilities are effective in eliciting valuations that are statistically indistinguishable from a fully incentivized scheme and that all incentivized conditions can mitigate hypothetical bias, resulting in lower elicited valuations than a purely hypothetical condition.

Harrison (1989) showed that the *opportunity costs* of deviating in experimental first price auctions were minuscule (see also Harrison, 1992, for a specific example for the BDM mechanism). The key insight was that despite bids appearing far from the theoretical predictions, the loss of expected value was actually quite small. The critical issue highlighted by these findings is that the lack of salient or sufficiently meaningful incentives can dilute or eliminate inferences based on induced values. The main criticism this work was addressing was that risk preferences were being challenged on the grounds that they could not explain bidding behavior. The converse was true: bids could not be used to infer risky behavior due to the potential issues with their induced values.

2.2 Incentives in choice under uncertainty

While the handful of value elicitation studies offer valuable insights, they pale compared to the number of studies in the literature surrounding choice under uncertainty. The initial impetus behind examining incentive compatibility more closely was likely driven by the surprising preference reversal phenomenon (Lichtenstein and Slovic, 1971) and its robustness to experimentalists' best efforts to eliminate this discrepancy (Grether and Plott, 1979). The discrepancy between valuation and binary comparisons implying the opposite preferences. That is, a higher valuation is assigned to the least preferred lottery in the binary comparison. This inconsistency challenged the principle of transitivity or weak axiom of revealed preferences underpinning most economic theory (Samuelson, 1938).

Several researchers working on risk preferences found an alternate explanation: either a failure of the reduction of compound lotteries axiom (Segal, 1988) or a failure of the independence axiom (Holt, 1986; Karni and Safra, 1987) could explain these perplexing empirical findings. The intuition for these explanations is straightforward: under non-expected utility preferences, the preferences over a menu of lotteries can be different from the pairwise preferences. To see this intuition, consider the pay-one randomly mechanism. If we elicit binary preference over a single choice, then failures to reduce the menu of choices to their constituent parts—the single choices—can generate this behavior. Alternatively, a failure of independence implies that making the effective choice random can generate different preferences as preferences are no longer *independent* from the probability with which they occur. Hence, non-expected utility preferences are menu-dependent. Perhaps to avoid the complexities that arise from thinking of preferences in this complicating manner, several researchers have argued for single binary choices as the 'gold standard.' Unfortunately, if true preferences are indeed menu-dependent, then binary preferences may be unobservable (see Machina, 1989), and even if they are observable, they may be uninformative about their menu-dependent counterparts. Despite these caveats, we discuss two canonical papers that implicitly assume the researcher is only interested in eliciting binary preferences.

First, Cox et al. (2015) explored how different incentive schemes may affect preference elicitation in choice under risk. They compare eight different incentive schemes: i) pay-all-sequentially as subjects make choices (PAS), ii) pay-all at the end with independent draws for each decision (PAI), iii) pay-one randomly with prior information; that is, see all options in advance before choosing one (PORpi), iv) pay-one randomly with no prior information (PORnp), v) combining POR with PAS; options are played out sequentially as in PAS before the option that is relevant for payoff is randomly selected (PORpas), vi) pay-all correlated; pay all with one realization of the world at the end (PAC), vii) pay-all correlated but divide the payment by the number of choices in order to scale down payoffs similar to POR (PACn), viii) only one task is performed and paid (OT).[3] Their findings indicate that individual behavior is significantly affected by the

[3] POR mechanisms are strongly incentive (consistent for binary preferences) compatible for theories that assume the reduction of compound lotteries and independence axioms. The reduction and dual independence axioms imply that PAC and PACn are weakly incentive (consistent for a specific menu) compatible for comonotonic lotteries. In theory, PAS should not be incentive-compatible under the expected utility over terminal wealth (EUTW) model. PAI should only be incentive-compatible under risk neutrality. Therefore, some payment mechanisms for binary choice are theoretically incentive-compatible under more restrictive assumptions. Because with the OT mechanism each subject has to respond to only one choice task which is played out for real, it is considered the gold standard by which to compare all other mechanisms albeit there is no consensus over it (see Johnson et al., 2021, and references therein). Moreover, Brown and Healy (2018) argue that an OT treatment confounds incentive compatibility failures with framing effects.

payoff mechanism, and this phenomenon is not unique to the pay-one-randomly mechanism (POR).

Second, Azrieli et al. (2018) study general conditions for assumptions on preferences to result in an incentive-compatible mechanism for binary preferences. For example, they show that if there are no negative complementarities at the top–making a bundle with every preferred element from a set of choices suboptimal–then pay-all mechanisms are the only incentive-compatible schemes. They also show that if preferences are event/state-wise monotonic and consistent with the reduction of compound lotteries, then the only incentive-compatible mechanism is POR.[4] However, they note that if preferences satisfy both 'reduction' and 'monotonicity', then preferences are consistent with the independence axiom and, therefore, rule out non-expected utility preferences (Azrieli et al., 2020). As an alternative, Starmer and Sugden (1991) and Cubitt et al. (1998) show theoretically and empirically that if each decision is treated in *isolation*, i.e., not reduced, then POR is incentive compatible for non-expected utility preferences.

However, it is not always clear at which level isolation holds. For instance, Brown and Healy (2018) find that showing all decisions together in a list may compromise incentive compatibility (or produce framing effects), but randomizing rows and presenting them on separate screens may restore it. Freeman and Mayraz (2019) consistently find more risk-taking when a choice is embedded in a choice list than when it is presented on its own, and this difference persists when they inform subjects of the paid choice in advance. This implies that isolation can fail not because of random incentives but simply because the choice appears in a list together with others. Following Azrieli et al. (2018), this may or may not be a challenge for incentive compatibility, which is entirely dependent on the assumptions placed on preferences. For example, Feldman and Ferraro (2023) model preferences as being reference-dependent expected utility within every choice list, while the reference point changes across every choice list. Hence, POR is incentive compatible for these preferences as long as each choice list is treated in isolation. Treating each choice in the choice list in isolation would be inconsequential from an incentive compatibility perspective, while integrating multiple choice lists would be incentive incompatible.

There are other papers that examine incentive compatibility.[5] For example, Li (2021) find that the Accumulative Best Choice ('ABC') mechanism is incentive compatible for all rational (complete and transitive) risk preferences. Subjects face N choices; the selected best alternative is carried over to the next menu; and subjects only get paid for the last round. It is important however, to emphasize Azrieli et al. (2020) cautionary words (and footnote):

[4] Formally, event-wise independence implies that if we replace an outcome in a lottery with a preferred outcome, then this new lottery is preferred. That is, contrasting with the independence axiom, this monotonicity is the same as independence over outcomes (degenerate lotteries).

[5] Charness et al. (2016) reviews several studies that randomize who is paid and how many decisions are made.

> Behaviorally, we speculate that mechanisms that pay based on surely-identified sets [like in Li (2021)] or that use negative weights are excessively complicated and may lead to more confusion and mistakes by subjects.[10]
>
> [10] Our theory assumes a deterministic preference relation and does not allow for mistakes or stochastic choice. Though this would be an important direction to study, even the definition of incentive compatibility becomes unclear when random behavior is permitted.

Increasingly, for the reasons stated above, researchers have focused on behavioral incentive compatibility. Cason and Plott (2014) showed how misunderstanding the incentive scheme can lead to overvaluations. Further, Danz et al. (2022) argues that failures to improve the elicitation when more information about the mechanism is given to participants or when the incentives of the mechanism are represented as incentives only (replacing choices by their (lottery) payments), can show that the mechanism is not incentive compatible. Hence, behavioral incentive compatibility is primarily an empirical question, albeit one informed by theory.

We now focus on the footnote; currently, there is ample evidence that individuals can exhibit (deliberate) stochastic behavior (Agranov and Ortoleva, 2017; Dwenger et al., 2018; Feldman and Rehbeck, 2022) and that they also make mistakes (Benjamin et al., 2020; Breig and Feldman, 2023; Martínez-Marquina et al., 2019). As an interesting example, consider McGranaghan et al. (2024), where valuation tasks (which are not incentive compatible under non-expected utility preferences) are utilized to estimate choice errors in paired binary tasks. They argue that even with a biased preference elicitation, valuation tasks can be used to control for randomness in binary choice behavior. Relatedly, Buschena and Zilberman (2000) showed that if preferences exhibit stochasticity, then assumptions on preferences and assumptions on their stochastic component are not independent and cannot be separately identified using binary choices. Again, these imprecisions in (single) binary choices challenge their status as the 'gold standard.'

To recapitulate, our survey of the incentivization literature reveals that incentives are an empirical and theoretical question. Failure to incorporate insights from either camp is unlikely to be fruitful in determining optimal incentive schemes.

3 Experiment 1: Preference elicitation with the BDM mechanism

3.1 Methods and Experimental Design

We designed and executed the experiment online via Qualtrics. Subjects were panelists from Forthright Access, an online research company that handles their own recruitment through a variety of direct advertising channels. All potential panelists are processed through a multi-step, double opt-in procedure to both ensure informed consent, and collect basic profile information. Once participants are in the panel, the company continues to capture new profiling metrics and monitor their data for quality control. All Forthright panelists participate in surveys where they are shown the rewards in dollar amounts, and over half have validated their personal phone numbers with the company that allows them to receive instant rewards for their participation.[6] This study and the studies described in the next sections were preregistered with the AEA's RCT registry (AEARCTR-0009687).

Participants were offered a $2.5 reward for a 20 min study. Subjects were informed they could also gain additional rewards after entering the study. We employed several quality controls to ensure subjects' attention and comprehension based on a pilot study with 78 subjects (Haaland et al., 2023). First, all the instruction screens included minimum timers and subjects were excluded if at any point they rushed out to the next screen (less than 7 secs).[7] Second, we included two attention check questions. In the first question, subjects were explicitly asked to skip the question without providing an answer. At a later point in the study, a second question asked subjects to indicate if they agreed with a nonsensical proposition. Failure to disagree indicates low attention to the study. Subjects who failed both attention checks were excluded from the study and received no payment.

The instructions included several examples explaining how the BDM mechanism works, followed by a series of true/false and open-ended comprehension questions. If a subject indicated or typed the wrong answer, they received an explanation about the correct answer and were asked to explicitly click or state the correct answer to proceed. All experimental instructions, test questions, and attention checks have been deposited with the Open Science Framework: `https://osf.io/2qpnw/?view_only=8152fec1eb48401283995375e6e5840d`.

After screening out inattentive subjects for one of the reasons explained above, the sample included 2,575 subjects.[8] In addition to the participation fee, subjects earned an average of $2.67 (min=$0, max=$29.4) from this part of the study.[9]

Experiment 1 involved preference elicitation over an induced value (IV) via the BDM mechanism. As in Cason and Plott (2014), subjects were endowed with a card worth a known IV and were asked to state their offer price to sell the card back to the experimenter.[10] Subjects were

[6] A stream of studies has opted to examine the performance of panels from various survey providers (Chandler et al., 2019; Litman and Robinson, 2021; Peer et al., 2017, 2022) particularly after concerns about diminishing data quality on MTurk (Ahler et al., 2021; Chmielewski and Kucker, 2020). The Forthright panel has been found to have good representativeness and quality of data when screening for attentiveness (Stagnaro et al., 2024).

[7] At the beginning of the study, subjects received a warning that they would be excluded if they failed quality control checks or if they did not pay attention to the instructions.

[8] The second part of the study (not analyzed in this paper) explored preferences for sustainable meat consumption, so we filtered out about 3.9% of the sample (223 subjects) that were vegan or vegetarian.

[9] This part of the study was combined with a second part on incentivized preference elicitation for steaks so that in total, subjects received on average $6.08 from both parts of the experiment (min=$2.5, max=$33.9).

[10] We interchangeably refer to offer prices as bids and vice versa.

informed that their offer price would be compared to a fixed offer that would be randomly drawn from the interval of $[0, X]$ where X was varied from task to task depending on the payment scheme.[11] We varied the IV at a low and a high level of \$1 and \$3 and varied the maximum bid range, X, at \$4, \$5, and \$6. Consequently, each subject participated in six tasks; all possible combinations of the IV and the upper level X of the support of the distribution. The order of the six preference elicitation tasks was randomized across subjects.

Our experimental design also varied on two between-subjects dimensions. To test for potential diluted incentives, we had three probabilities of decisions being paid: subjects either had a 100% chance of getting monetary rewards associated with their decisions, a 50% chance, or a 1% chance. After collecting data for these treatments, we found a null effect of differences between the treatments, so we decided to run two additional boundary conditions: a 0.2% chance treatment of getting monetary rewards and a purely hypothetical treatment. Thus, the first dimension of payments had five distinct probabilities of decisions being paid. Every subject was given information about the probability of their decisions being paid in two different screens; one at the beginning of the study and one right before eliciting their preferences with the BDM mechanism. For the hypothetical treatment subjects were informed multiple times at different points of the study that although monetary rewards would be shown in various screens, they would only receive fixed compensation and that none of the stated monetary amounts would count toward their earnings. Text was modified appropriately for the rest of the treatments varying the probability of payments. Complete scripts are shown in the Online Appendix.

The second dimension of payments we varied, was the number of decisions paid, the correlation between those payments, and whether we adjusted for the magnitude of the number of decisions paid. Our base payment is the Pay-One-Randomly (POR) mechanism, where only one of the six tasks is randomly selected for payment. We compare the POR mechanism with four additional payment mechanisms previously used by Cox et al. (2015): (a) the Pay-All-Correlated (PAC) mechanism, (b) the PAC mechanism adjusted for the number of tasks (PACn), (c) the Pay-All-Independently (PAI) mechanism and (d) the PAI mechanism adjusted for the number of tasks (PAIn). For the PAC mechanism, subjects were paid for all six preference elicitation tasks, but the fixed offer was determined with a single draw for all tasks as follows: a random number would be drawn between 0% and 100%, and the randomly drawn percentage would be multiplied by the upper support of the distribution of allowed offers which would determine the fixed offer per task, albeit with just one draw. An arithmetic example illustrated this mechanism to subjects. The PACn mechanism was explained in a similar fashion, albeit subjects were

[11] Because of several behavioral biases associated with the BDM mechanism regarding the minimum and maximum values of the support distribution (they determine expectations, i.e., the probability of getting the induced value conditional on one's offer price, as well as they serve as price and loss anchors), we varied the support distribution between subjects instead of keeping it constant to a predetermined level (see Vassilopoulos et al. (2018) and references therein for a discussion as well as Mamadehussene and Sguera (2022)).

aware they would receive one-sixth of the total payoffs (therefore, payoffs were divided by the number of tasks).

In the PAI mechanism, subjects received an independent draw per task as follows: subjects were informed that the computer would choose a percentage number for each task that would be multiplied by the upper support of the distribution of allowed offers, which would determine a different fixed offer per task. An arithmetic example illustrated this mechanism to subjects. The PAIn mechanism was explained in a similar way, albeit subjects were aware they would receive one-sixth of the total payoffs.

Table 1 summarizes the experimental design and number of subjects assigned to each treatment arm. Our target of 100 subjects/treatment is large enough to detect minimum differences in absolute bid deviations ($(|bid - IV|/IV)$) of 0.05 or larger with 80% power. Sample size calculations, instructions, examples, and final payoff screens have been deposited with the Open Science Framework: https://osf.io/2qpnw/?view_only=8152fec1eb48401283995375e6e5840d.

Table 1: Experimental design and number of subjects per treatment

		Payment mechanism					
		PAC	PACn	PAI	PAIn	POR	**Total**
Incentives	Hypothetical	100	101	101	100	102	*504*
	0.20%	101	100	100	101	101	*503*
	1%	113	100	105	100	101	*519*
	50%	101	115	100	99	103	*518*
	100%	100	99	104	120	108	*531*
	Total	*515*	*515*	*510*	*520*	*515*	**2,575**

Notes: PAC, PAI and POR stand for pay-all-correlated, pay-all-independently and pay-one-randomly, respectively. n indicates sum of payoffs are divided by the number of tasks.

3.2 Experiment 1 Results

Before discussing the results of Experiment 1, it is customary to check the balance of subjects' observable characteristics across treatments. While many researchers use statistical tests to check for balance of observable characteristics between treatments, the literature points to some pitfalls of this practice (e.g., Briz et al., 2017; Deaton and Cartwright, 2018; Ho et al., 2007; Moher et al., 2010; Mutz and Pemantle, 2015). Following this literature, we report in Table A2 standardized differences across treatments (Imbens and Rubin, 2016; Imbens and Wooldridge, 2009). Since the differences are pairwise comparisons of all treatment cells, for brevity we only present comparisons between the payment-probabilities treatments. Cochran and Rubin's (1973) rule of thumb is that the standardized difference should be less than 0.25. The last column in Table A2 also compares the demographics between the pooled sample with the sample of inattentive subjects filtered out of the study. None of the variables show an imbalance.

Figure 1 shows bid deviations from the IV (Panel a) and relative absolute deviations from IV (Panel b). It is clear that more misbidding occurs for the lower IV.[12] The upper panel also shows that overbidding is more common than underbidding. Only 15.50% of all bids are exactly equal to the IV and 24.71% (30.89%) of all bids are within 5% (10%) of the IV. Cason and Plott (2014) report that without training 16.7% of subjects have bids within 5 cents (2.5%) of their induced value of $2, which is similar to our findings. Moreover, Brown et al. (2023) find similar patterns of misbidding that are fairly constant across various elicitation formats that are strategically-equivalent but cognitively simpler than the BDM mechanism.

Table 2 shows descriptive statistics (mean, standard deviation, median) for the relative absolute deviations by incentives and payment mechanism. The values of the deviations in this table are remarkably stable across treatments at around 0.5, supporting a null effect of both incentives and payment mechanism.

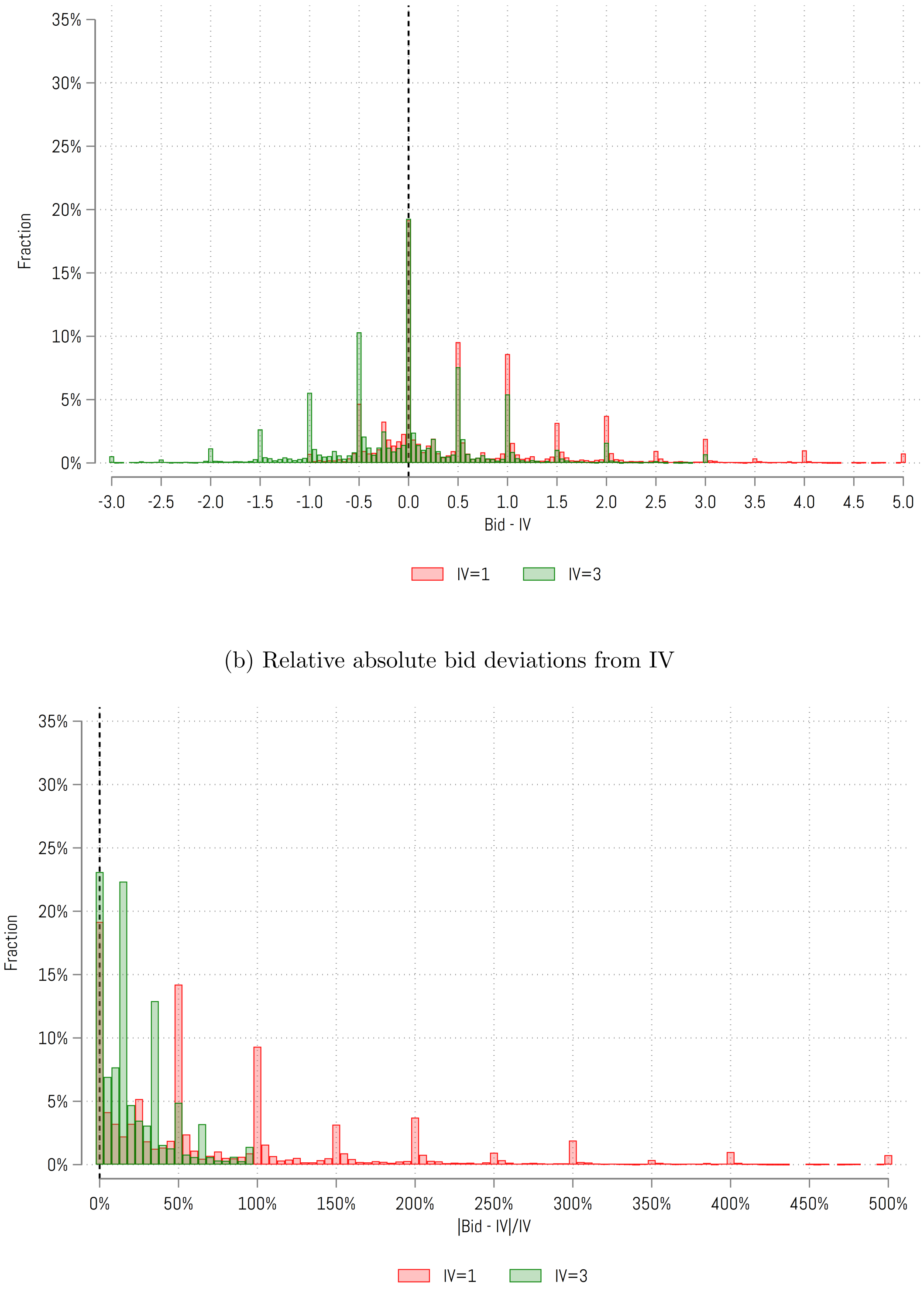
Table 2: Descriptive statistics of $|Bid - IV|/IV$ by payment mechanism and incentive scheme

	PAC	PACn	PAI	PAIn	POR	Total
Hypothetical	0.603 (0.912) [0.315]	0.512 (0.741) [0.250]	0.580 (0.853) [0.250]	0.548 (0.855) [0.250]	0.502 (0.790) [0.177]	0.549 (0.832) [0.250]
0.2%	0.443 (0.687) [0.200]	0.477 (0.731) [0.250]	0.477 (0.720) [0.238]	0.530 (0.741) [0.260]	0.569 (0.833) [0.305]	0.499 (0.745) [0.250]
1%	0.582 (0.881) [0.247]	0.439 (0.637) [0.183]	0.470 (0.656) [0.240]	0.554 (0.833) [0.260]	0.496 (0.693) [0.218]	0.510 (0.751) [0.240]
50%	0.500 (0.726) [0.247]	0.549 (0.822) [0.245]	0.483 (0.695) [0.250]	0.553 (0.785) [0.275]	0.539 (0.811) [0.257]	0.525 (0.771) [0.250]
100%	0.516 (0.799) [0.200]	0.537 (0.763) [0.283]	0.439 (0.636) [0.200]	0.555 (0.832) [0.250]	0.508 (0.740) [0.252]	0.512 (0.759) [0.250]
Total	0.530 (0.809) [0.240]	0.504 (0.744) [0.250]	0.489 (0.716) [0.250]	0.548 (0.810) [0.250]	0.522 (0.775) [0.250]	**0.519** **(0.772)** **[0.250]**

Notes: Table shows means, standard deviations in parenthesis and medians in brackets. PAC, PAI and POR stand for pay-all-correlated, pay-all-independently and pay-one-randomly, respectively; n indicates sum of payoffs are divided by the number of tasks.

[12] Table A1 shows descriptive statistics (mean, standard deviation, median) for the relative absolute deviations by IV and upper limit of the support distribution. A larger support limit and a smaller IV, increase relative absolute bid deviations.

Figure 1: Histograms of bid deviations from IV (BDM)



(a) Bid deviations from IV

(b) Relative absolute bid deviations from IV

Table 3 shows estimates from regression models with clustered standard errors at the individual level, using either bid deviations ($Bid-IV$) or relative absolute deviations ($|Bid-IV|/IV$) as the dependent variable and the treatments dummies as independent variables.[13]

As shown in Table 3, compared to the baseline of the 100% incentives with a POR payment mechanism, none of the incentives schemes nor the payment mechanism significantly affect misbidding behavior. None of the coefficients are statistically different to the baseline. On the other hand, both the IV and the support level of the distribution affect deviations from the induced value. More specifically, the upper panel of Table 3 shows that a larger induced value reduces deviations from the IV and that this reduction is moderated by the level of the support of the distribution. For the lower IV of $1, a larger support increases relative absolute misbidding by 0.17 to 0.39. The larger IV of $3 reduces misbidding behavior but with a larger support this reduction shrinks. For example, model (1) shows that misbidding declines by 0.76 for a $4 support, but it is only reduced by 0.31 for the larger level of support of $6.[14]

We can also classify subjects depending on whether they submitted a bid lower, equal or larger than their induced value and estimate ordered logit regressions on the treatment variables. Table A6 in the Online Appendix shows raw coefficient estimates for four different models that increase the tolerance of grouping subjects as bidding under/over or around their assigned IV. The heading of Table A6 indicates the corresponding definition for the three ordinal categories. For example, model (4) in Table A6 classifies a subject as bidding close to the IV if the respondent submitted a bid within ±10% of their IV and as an underbidder (overbidder) if they submitted a bid smaller (larger) than 90% (110%) of their IV. Table A7 in the Online Appendix shows that estimates are roughly similar when adding demographic control variables in the specifications. In general, relaxing the thresholds of classifying subjects as under- or over-bidders does not change any of the conclusions.

The pattern of results in Tables A6 and A7 is similar to the regressions reported in Table 3. None of the incentive schemes or payment mechanism treatments have an effect on bidding behavior in the BDM mechanism. On the other hand, both the magnitude of the IV and the upper level of the support have a statistically significant effect on bidding. A larger IV reduces

[13] Table A5 in the Online Appendix Supplementary Material show estimates where we add demographic controls to the specifications albeit we lose some observations due to missing values.

[14] To explore whether sliders induced a different bidding behavior than letting subjects freely submit a bid into a box, we included an additional test at the end of the study for a subset of 503 subjects randomly assigned on a between-subjects basis to a treatment where they had to submit a bid using a slider or to another treatment where they would submit a bid using an input box. In this part of the study, subjects faced an induced value of $2 and the support was varied within-subjects at two levels: $3 or $4, so that subjects participated in two bidding rounds. Subjects were paid for one randomly selected round on top to any other earnings. We regressed bid deviations from IV or absolute bid deviations from IV on the treatment variables and a set of demographic controls. We find that the slider does not have a statistically significant effect on bid deviations as compared to the box ($\hat{b} = -0.049$, $se = 0.057$ for the $2 support; $\hat{b} = -0.116$, $se = 0.074$ for the $3 support) and that a higher support level induces higher misbidding ($\hat{b} = 0.328$, $se = 0.043$ for the box; $\hat{b} = 0.261$, $se = 0.041$ for the slider). Similar results are in place if one uses absolute bid deviations as the dependent variable.

Table 3: Regressions of bid deviations on treatment variables

	$Bid - IV$		$\lvert Bid - IV \rvert / IV$	
	(1)		(2)	
Constant	0.479^{***}	(0.070)	0.624^{***}	(0.043)
IV = 1 & Support = 5	0.190^{***}	(0.017)	0.174^{***}	(0.015)
IV = 1 & Support = 6	0.416^{***}	(0.022)	0.392^{***}	(0.020)
IV = 3 & Support = 4	-0.765^{***}	(0.019)	-0.438^{***}	(0.014)
IV = 3 & Support = 5	-0.562^{***}	(0.018)	-0.428^{***}	(0.013)
IV = 3 & Support = 6	-0.309^{***}	(0.018)	-0.398^{***}	(0.012)
Hypothetical & PAC	0.018	(0.116)	0.095	(0.069)
Hypothetical & PACn	-0.103	(0.104)	0.004	(0.062)
Hypothetical & PAI	0.083	(0.107)	0.072	(0.065)
Hypothetical & PAIn	0.003	(0.104)	0.040	(0.063)
Hypothetical & POR	-0.024	(0.099)	-0.007	(0.063)
0.2% & PAC	-0.089	(0.094)	-0.065	(0.058)
0.2% & PACn	-0.172^{*}	(0.103)	-0.031	(0.059)
0.2% & PAI	-0.048	(0.097)	-0.031	(0.060)
0.2% & PAIn	-0.001	(0.099)	0.022	(0.059)
0.2% & POR	0.022	(0.104)	0.061	(0.063)
1% & PAC	0.073	(0.101)	0.074	(0.064)
1% & PACn	-0.022	(0.093)	-0.069	(0.058)
1% & PAI	-0.069	(0.091)	-0.038	(0.055)
1% & PAIn	0.044	(0.109)	0.046	(0.065)
1% & POR	-0.036	(0.094)	-0.013	(0.058)
50% & PAC	-0.113	(0.101)	-0.008	(0.060)
50% & PACn	0.073	(0.097)	0.040	(0.059)
50% & PAI	-0.052	(0.100)	-0.025	(0.060)
50% & PAIn	0.066	(0.101)	0.044	(0.060)
50% & POR	-0.031	(0.099)	0.030	(0.061)
100% & PAC	0.034	(0.096)	0.008	(0.060)
100% & PACn	0.075	(0.099)	0.029	(0.059)
100% & PAI	-0.134	(0.087)	-0.069	(0.054)
100% & PAIn	0.076	(0.097)	0.047	(0.061)
Observations	15450		15450	
R^2	0.154		0.181	
Adj. R^2	0.153		0.180	
F-stat. (p-value)	86.223	(< 0.001)	59.250	(< 0.001)

Notes: Clustered standard errors in parentheses. * p<0.1, ** p<0.05 *** p<0.01. Base categories for the treatment variables are: IV = 1 & Support = 4, 100% & POR. PAC, PAI and POR stand for pay-all-correlated, pay-all-independently and pay-one-randomly, respectively. n indicates sum of payoffs are divided by the number of tasks.

the likelihood of overbidding by as much as 27.3% for model (1) of Table A6 for the lowest support level of \$4 and the effect is still negative but slightly smaller for a larger support level of \$6 (13.4%). In general, a larger IV and lower support level elicit bids that are closer to the IV i.e., less misbidding behavior.

4 Experiment 2: Preference elicitation with the Second Price Auction

To test whether the preference elicitation mechanism has an effect on elicited preferences, in Experiment 2 we replaced the BDM mechanism with the Second Price Auction (SPA). Both BDM and SPA are theoretically incentive compatible, but the SPA has a different strategic uncertainty about the bidding strategy coming from other bidders rather than a randomly drawn price. Although this should not influence behavior, by replacing the BDM with and an SPA allows us to empirically test the source of strategic uncertainty as a potential explanation of the insensitivity we observe in Experiment 1 regarding payment mechanisms and incentive schemes. We reduced the treatment arms of the experimental design of Experiment 1 to fit budget constraints and selected to test a subset of treatments that are most widely used and provide boundary conditions since they may be more likely to affect bidding behavior. With respect to payment incentives we administered a purely hypothetical treatment and a treatment that pays with 100% certainty. With respect to the payment mechanism, we selected the POR and the Pay-All divided by the number of rounds (PAn) in order to keep incentives comparable.[15] In summary, we implement a 2×2 between-subjects design in Experiment 2.

4.1 Methods and Experimental Design

We designed and executed our experiment online via Qualtrics using SMARTRIQS (Molnar, 2019) which allows interactive online experiments. Subjects were panelists from Forthright Access, none of which had participated in Experiment 1. We offered a \$2 reward for a 15 min study. Subjects that were not assigned to a hypothetical treatment, were informed they could also gain additional rewards after entering the study.

We implemented the same quality controls as in Experiment 1, and the Experimental Instructions have been deposited with the Open Science Framework: https://osf.io/2qpnw/?view_only=8152fec1eb48401283995375e6e5840d. One particular feature to this experiment is that recruitment was done on a limited time window within a day so that we have a large

[15] Note that a pay-all correlated or pay-all independent mechanism do not make sense in the context of the SPA since the second price is determined endogenously in the auction group, unlike the exogenous random draw of the BDM mechanism.

number of participants entering simultaneously and achieve good matching of people to the auction groups. Four subjects would form an auction group but if more than 3.5 minutes had elapsed without fulfilling a group, then we used bots to complement a group. If bots were used in a particular group, the participants were informed about it. The main results present responses from subjects who were matched in groups of humans only. The results including the subjects matched with bots are presented in the Online Appendix and are similar to what we present here. Subjects were informed about the number of bots they were matched with, if any, and while we control for the number of bots in the regressions shown in the Online Appendix, we find that this information does not significantly affect our results. All subjects in a group were assigned to the same treatment for the entire experiment.[16]

The final sample with complete responses includes 637 subjects, albeit 209 of them were matched with one or more bots. On top to their participation fee, subjects received an average of \$1.06 (min=\$0, max=\$3). Table 4 shows the number of subjects per treatment. In the main regressions we only use observations from subjects that were not matched to bots but we controlled for the number of bots in additional specifications shown in the Online Appendix and all of our results hold.

Table 4: Experimental design, number of subjects and number of bots per treatment

		N of bots		
Incentives	Payment mechanism	0	$\geq$1	**Total**
Hypothetical	PAn	96	39	*135*
Hypothetical	POR	116	75	*191*
100%	PAn	96	43	*139*
100%	POR	120	52	*172*
Total		*428*	*209*	**637**

Notes: PA and POR stand for pay-all and pay-one-randomly, respectively; n indicates sum of payoffs are divided by the number of tasks.

Similar to Experiment 1, subjects were endowed with a card worth a known IV and were asked to state their offer price to sell the card back to the experimenter with the understanding that they were assigned to a group of four subjects and their offer is compared to all other offers and the lowest offer is accepted, but the second lowest offer is the binding price. Subjects experienced four different IVs that were selected to be in the same range as in Experiment 1: \$1, \$1.7, \$2.4 and \$3. Subjects experienced all the IVs in a random order and at any given round only one subject was assigned to each IV so that all four IVs were assigned at any round.

Before participating in the SPA, all subjects went through similar instructions, compre-

[16] We first launched only one of the treatments to assess the sign-up rates and timing and realized that we had to massively sent out invitations during a short time window in order to avoid subjects always being matched with bots. As a result more observations were collected in one of the treatments.

hension questions and quality checks as in Experiment 1. All experimental instructions, test question and attention check questions have been deposited with the Open Science Framework: https://osf.io/2qpnw/?view_only=8152fec1eb48401283995375e6e5840d.

4.2 Experiment 2 Results

Table A3 in the Online Appendix shows standardized differences of observable characteristics between the treatments and comparisons with subjects that did not finish the study. Because we observe a few small differences in some of the treatments, we control for these characteristics in subsequent analysis.

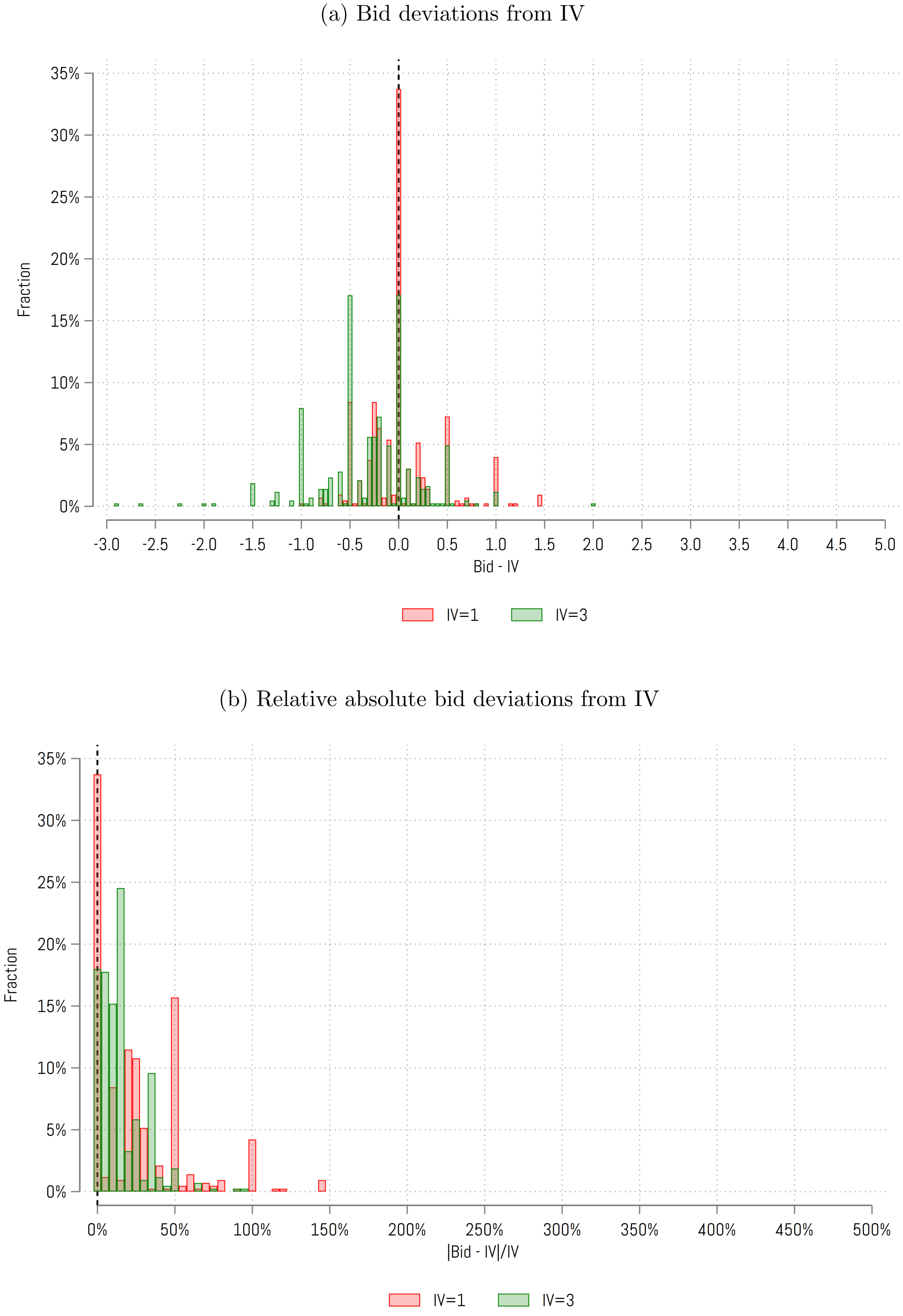
Figure 2 shows bid deviations from IV (Panel a) and relative absolute deviations from IV (Panel b) for two of the IVs. We purposefully keep the scale of the x-axis similar to Figure 1 so that the reader can easily visualize the differences to the BDM mechanism in Experiment 1. The results show evidence that the SPA leads to less misbidding than the BDM and that a larger IV reduces misbidding. In the SPA, 19.98% of all bids are exactly equal to the IV and 27.45% (42.93%) of all bids are within 5% (10%) of the IV. This is a substantial improvement compared to the BDM mechanism in Experiment 1.

Table 5 shows estimates from regression models with clustered standard errors at the individual level where we regressed either bid deviations ($Bid - IV$) or relative absolute deviations ($|Bid - IV|/IV$) on the treatments dummies. The sample is restricted to subjects that were not matched with a bot for the SPA.[17]

With respect to the payment mechanism and incentive schemes, the results are similar to the general pattern we observe with the BDM mechanism. Higher IVs reduce the level of misbidding while misbidding is unresponsive to the payment mechanism and whether the treatment is hypothetical or real.

[17] Table A8 in the Online Appendix extends estimations by including demographic controls while Table A9 includes in the estimations subjects that were matched with bots and adds the number of bots as an additional control. Results are robust to these additional specifications. Serizawa et al. (2024) find minimal effects in bidding behavior when subjects play against bots versus humans.

Figure 2: Histograms of bid deviations from IV (BDM)


(a) Bid deviations from IV

(b) Relative absolute bid deviations from IV

Table 5: Regressions of bid deviations on treatment variables for the SPA

	$Bid - IV$		$\lvert Bid - IV\rvert/IV$	
	(1)		(2)	
Constant	0.053	(0.033)	0.258***	(0.018)
IV = 1.7	-0.149***	(0.021)	-0.123***	(0.017)
IV = 2.4	-0.265***	(0.022)	-0.137***	(0.018)
IV = 3	-0.331***	(0.030)	-0.124***	(0.017)
Hypothetical & PAn	-0.034	(0.047)	0.016	(0.018)
100% & PAn	-0.034	(0.046)	0.027	(0.017)
100% & POR	-0.040	(0.044)	0.009	(0.017)
Observations	1712		1712	
R^2	0.081		0.056	
Adj. R^2	0.077		0.052	
F-stat. (p-value)	32.691	(< 0.001)	12.341	(< 0.001)

Notes: Clustered standard errors in parentheses. * p<0.1, ** p<0.05 *** p<0.01. Base categories for the treatment variables are: IV = 1, Hypothetical and POR. PA and POR stand for pay-all and pay-one-randomly, respectively; n indicates sum of payoffs are divided by the number of tasks.

4.3 The BDM mechanism vs. the SPA

In this section we further explore differences in bidding behavior between the two elicitation mechanisms explored in Experiment 1 and Experiment 2. Average difference of $Bid - IV$ is 0.29 in the BDM and -0.16 in the SPA indicating that subjects on average overbid in the BDM mechanism and underbid in the SPA. The difference is statistically significant using a t-test or a Wilcoxon-Mann-Whitney test (both p-values < 0.001). In terms of absolute relative deviations ($|Bid - IV|/IV$), subjects deviate on average 51.9% in the BDM mechanism and around 17.4% in the SPA indicating a substantially lower level of misbidding in the SPA (p-value < 0.0001 according to a t-test or a Wilcoxon-Mann-Whitney test). The magnitude of the improvement with the SPA is large.

We also regress bid deviations on the SPA dummy and demographic controls, and confirm that the SPA elicits smaller deviations from IVs ($\hat{b} = -0.446, se = 0.023$). Similarly for absolute bid deviations, the SPA elicits 33.9% smaller bids than the BDM ($se = 0.012$).

4.4 Decision-making noise and subjects' misconceptions

Misbidding in the BDM could arise if agents experience decision-making noise. We use maximum likelihood methods to fit our data in a model where subjects choose to maximize their expected payoff but make logit errors (as in Cason and Plott, 2014). Subjects' probability of submitting an offer b_j can be defined as:

$$\text{Prob}(\text{offer} = b_j) = \frac{e^{\lambda E[\pi|b_j]}}{\sum_{k\epsilon K} e^{\lambda E[\pi|b_k]}} \tag{1}$$

where K is the set of possible offers and λ bounds the cases where subjects are insensitive to differences in expected payoffs ($\lambda = 0$) or where subjects choose the offer that maximizes their expected payoff with the highest probability ($\lambda \to \infty$). A higher level of λ indicates a better fit, requiring less noise to characterize subject's choices according to that particular model. The λ parameter comes from the Quantal Response Equilbrium approach where the probability of taking an action is modeled as a multinomial logit.

In Equation 1, using $E^{opt}[\pi|b_j] = IV \times \text{Prob}(b_j > p) + E(p|b_j < p) \times \text{Prob}(b_j < p)$, then λ characterizes the optimal offers model without any misconceptions about the payoff function of the BDM. We then define the log-likelihood function as:

$$\ln L(\lambda; y_i) = \sum_i \ln \frac{y_i e^{\lambda E[\pi|b_j]}}{\sum_{k\epsilon K} e^{\lambda E[\pi|b_k]}} \tag{2}$$

where y_i is an indicator that the offer is b_j.[18]

We also estimate a mixture specification that allows some choices to be consistent with a First Price Auction-Game Form Misconception (FPA-GFM) model with probability M (that is, we allow the expected payoff to be $E^{gfm}[\pi|b_j] = IV \times \text{Prob}(b_j > p) + b_j \times \text{Prob}(b_j < p)$ instead of $E^{opt}[\pi|b_j]$, i.e., subjects believe they are getting paid their bid amount, and not the posted price) and consistent with the optimal offer model of BDM with probability $1 - M$.

Table 6 shows estimates from the optimal BDM model and the mixture model (Models 1 and 2). As a general remark, none of the estimated parameters is affected by the probability of realization of the payment or the payment mechanism treatments. Estimates from model (1) and model (2) are very similar, which is reasonable given that the estimated probability M is just 3.8%, not giving much support for the FPA-BDM model. The estimated parameters for λ show that with a larger IV subjects become more sensitive to differences in expected payoffs (the effect on λ is positive). A larger support has a negative effect on λ for the low IV, which can be interpreted as subjects becoming less sensitive to differences in expected payoffs. For the large IV, λ increases for the \$5 support, and it remains relatively stable for a further increase of the support to \$6 based on model (2).

Model (3) estimates an optimal model with the SPA data assuming subjects think that win-

[18] Because the expression in the denominator of Equation 1 becomes extremely large when one uses the lowest possible division of 1 cent (i.e., the expression involves the summation of $(\bar{p} \times 100 + 1)$ summands, where $\bar{p}$ is the upper limit of the support distribution), the maximum likelihood estimations are performed with y_i indicating offers are in a bin within X cents of b_j, where X is the lowest division that our estimation software would accommodate given the length of the expression involved. Cason and Plott (2014); Drichoutis and Nayga Jr (2022) use a similar strategy.

Table 6: Maximum likelihood estimates of Decision-making noise and misperception models

	(1)		(2)		(3)		(4)	
	Noise-only model		CP mixture model		SPA		Pooled model	
Constant	3.218***	(0.418)	3.634***	(0.578)	24.728	(16.650)	2.098***	(0.087)
IV = 1, Support = 5	-0.808***	(0.125)	-0.901***	(0.176)				
IV = 1, Support = 6	-1.368***	(0.130)	-1.529***	(0.196)				
IV = 3, Support = 4	1.014***	(0.229)	0.496	(0.304)				
IV = 3, Support = 5	3.161***	(0.264)	2.820***	(0.320)				
IV = 3, Support = 6	2.653***	(0.209)	2.961***	(0.282)				
IV = 1.7					28.954**	(13.372)		
IV = 2.4					13.486	(12.271)		
IV = 3					-7.498	(11.920)		
Hypothetical, PAC	-1.069**	(0.510)	-1.291**	(0.628)				
Hypothetical, PACn	-0.102	(0.607)	-0.184	(0.757)				
Hypothetical, PAI	-0.619	(0.537)	-0.768	(0.661)				
Hypothetical, PAIn	-0.526	(0.548)	-0.685	(0.676)				
Hypothetical, POR	-0.199	(0.603)	-0.280	(0.758)	6.893	(9.366)		
Hypothetical, PAn					-0.816	(8.536)		
0.2%, PAC	0.774	(0.647)	1.200	(1.124)				
0.2%, PACn	0.025	(0.601)	-0.034	(0.766)				
0.2%, PAI	0.250	(0.672)	0.287	(0.884)				
0.2%, PAIn	-0.171	(0.522)	-0.241	(0.658)				
0.2%, POR	-0.598	(0.528)	-0.777	(0.648)				
1%, PAC	-0.664	(0.506)	-0.800	(0.632)				
1%, PACn	1.182	(0.867)	1.835	(1.482)				
1%, PAI	0.452	(0.577)	0.521	(0.762)				
1%, PAIn	-0.649	(0.550)	-0.814	(0.674)				
1%, POR	0.029	(0.573)	-0.018	(0.729)				
50%, PAC	-0.104	(0.581)	-0.220	(0.721)				
50%, PACn	-0.437	(0.523)	-0.553	(0.653)				
50%, PAI	0.298	(0.607)	0.449	(0.840)				
50%, PAIn	-0.345	(0.528)	-0.430	(0.663)				
50%, POR	-0.348	(0.552)	-0.479	(0.688)				
100%, PAC	-0.267	(0.534)	-0.345	(0.677)				
100%, PACn	-0.173	(0.550)	-0.214	(0.702)				
100%, PAI	0.940	(0.646)	1.258	(0.984)				
100%, PAIn	-0.497	(0.519)	-0.605	(0.653)				
100%, PAn					2.616	(8.507)		
SPA							26.823***	(3.519)
IV high							3.192***	(0.167)
	M			0.038***	(0.012)			
Observations	15450		15450		1712		17162	
Log-Likelihood	-55904.837		-49625.709		-5017.004		-49443.067	
AIC	111869.675		99313.419		10048.008		98892.135	
BIC	112099.036		99550.425		10086.125		98915.386	

Notes: Clustered standard errors in parentheses. * p<0.1, ** p<0.05 *** p<0.01. Base categories for the treatment variables are: IV = 1 & Support = 4, 100% & POR. PA, PAC, PAI and POR stand for pay-all, pay-all-correlated, pay-all-independently and pay-one-randomly, respectively; n indicates sum of payoffs are divided by the number of tasks.

ning prices are coming from a uniform distribution. The intercept is not statistically significant at the conventional significance levels indicating that at the base levels of the dummy variables, we fail to reject the null hypothesis that subjects are insensitive to differences in payoffs. Note that the magnitude of the estimated coefficients for the SPA are much larger than those estimated with the BDM data, indicating a better fit.

Model (4) pools observations from models (1) and (3). The effect of the SPA dummy is positive and statistically significant indicating that subjects become more sensitive to differences in expected payoffs in the SPA compared to the BDM.

While we find no differences in the payment mechanisms and incentives schemes (i.e., probability of realization), design features such as the IV and the support of the distribution have an impact on the bidding behavior. Furthermore, the results clearly indicate that the SPA induces behavior that is much closer to the IV and reduces the likelihood of misbidding compared to the BDM mechanism.

5 Experiment 3: Preference elicitation in choice under risk

Since Experiment 1 and Experiment 2 indicate the absence of any effect of payment mechanisms on bidding behavior, in Experiment 3 we revisit the effect of payment mechanisms on choice under risk following Cox et al. (2015). With respect to payment incentives we administered a purely hypothetical treatment and a treatment that always pays with 100% certainty. With respect to the payment mechanism, we selected the pay one randomly, the pay-all correlated and the pay-all independently treatments divided by the number of choice tasks, in order to keep incentives comparable across treatments. Experiment 3 consists of a 3×2 between-subjects design.

5.1 Methods and Experimental Design

We designed and executed our experiment online via Qualtrics. Subjects were panelists from Forthright Access, none of which had participated in Experiment 1 or Experiment 2. We offered a $1 reward for a 5 min study. Subjects that were not assigned to a hypothetical treatment were informed they could also gain additional rewards after entering the study.

Similar quality controls to Experiment 1 and 2 were enforced and all Experimental Instructions have been deposited with the Open Science Framework: `https://osf.io/2qpnw/?view_only=8152fec1eb48401283995375e6e5840d`. The final sample with complete responses includes 610 subjects. On top of their participation fee, subjects in the fully incentivized treatment received an average of $1.88 (min=$0, max=$5.5). Table 7 shows the number of subjects

Table 7: Experimental design and number of subjects

	Hypothetical	Real	Total
POR	99	100	*199*
PACn	109	101	*210*
PAIn	101	100	*201*
Total	*309*	*301*	**610**

Notes: POR, PACn, PAIn stand for pay-one-randomly, pay-all-correlated and pay-all-independently, respectively; n indicates sum of payoffs are divided by the number of tasks.

per treatment.

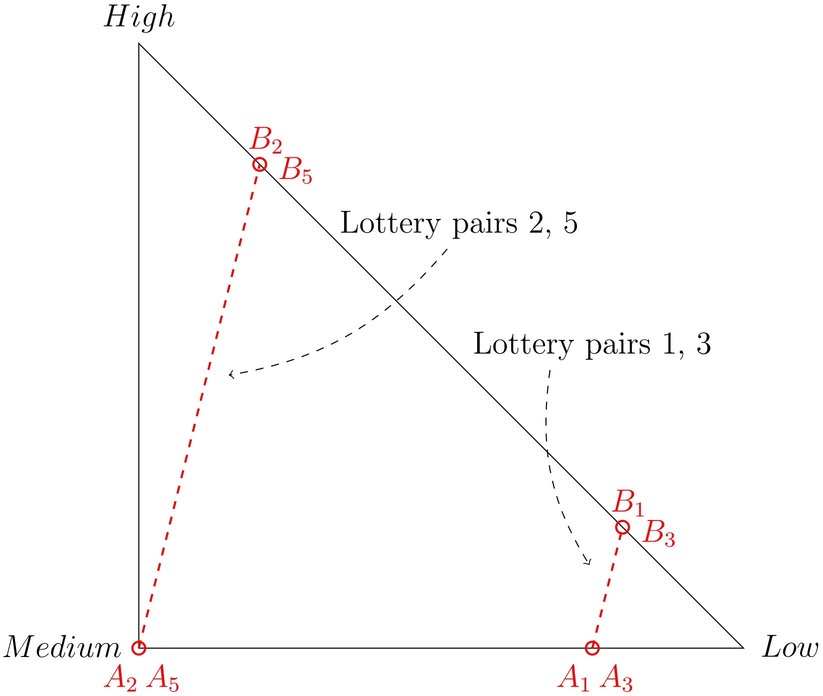
Subjects were asked to choose between the lottery pairs shown in Table 8 in the form of $(p, x_1; x_2)$ i.e., probability p of receiving x_1 and $1 - p$ of receiving x_2. Lotteries are scaled down versions (by a factor of four) in terms of monetary amounts of the lotteries in Cox et al. (2015) to manage budgetary constraints. In Table 8 we keep the same numbering of pairs as in Cox et al. (2015) to facilitate comparisons but group them together not in sequence to highlight their relationship to Allais' Paradox (Allais, 1953). Figure 3 depicts the lottery pairs in a Marschak-Machina probability triangle. Lottery pair 4 cannot be depicted since one of the lotteries is a three-outcome lottery. However, lotteries in pair 4 are constructed from either lottery pair 2 or lottery pair 3. To construct this lottery from pair 2, one needs to take a 25%/75% probability mixture, where the original pair 2 lotteries are received with a 25% chance and \$3.00 are received with a 75% chance. Alternatively, one can replace the common outcome of a 75% chance of \$0.00 in pair 3 with the common consequence of 75% chance of \$3.00.

Allais' so-called certainty effect pertains to the disposition towards certainty. Through his original thought experiment, he showed most people preferred a riskier alternative when two alternatives were very likely to yield nothing. Contrary to the predictions of independence, when replacing the common outcome of nothing in both lotteries such that one alternative yields a *certain* reward, then most people seem to prefer the safer, and certain, alternative.[19]

Consistent with Allais' certainty effect, pairs 2 and 5 include an option with certainty. Hence, pairs 1 and 3 represent uncertain gambles, and pairs 2 and 5 represent gambles where certainty is an alternative. Unlike Allais' classical paradox, monetary amounts are clearly lower. Like Allais, the independence axiom (parallel indifference curves in Figure 3) is broadly a choice of the same alternative (either A, the safer, or B, the riskier alternative). More explicitly, the independence axiom would imply the same choices across 2, 3, and 4 (either A or B). Note, that pair 1 is identical to pair 3, except the outcomes in pair 3 are half of what they are in pair 1. Similarly, all outcomes in pair 5 are the same as pair 2, but \$3 are added to each outcome. So differences

[19] Formally, the certain pair was (1, 100 million francs) or (0.1, 500 million francs; 0.89, 100 million francs; 0 francs) and for the uncertain pair he replaced the common consequence of (0.89, 100 million francs) in both options with (0.89, 0 francs).

Figure 3: Probability triangle for lottery choice tasks



in pairs 2 and 5 could be explained via wealth effects or extreme forms of reference dependence—ignoring standard calibration critiques (Rabin, 2013). For our main analysis, we group pairs 1 and 3 and pairs 2 and 5 to avoid the mental gymnastics of rationalizing choices using ludicrously 'calibrated' utility functions. We note a strict reading of neoclassical theory, assuming independence and preferences over final wealth level, would imply consistent choices across the board (either A or B). All experimental instructions have been deposited with the Open Science Framework: https://osf.io/2qpnw/?view_only=8152fec1eb48401283995375e6e5840d.

Table 8: Lottery pairs

	A: Safer	B: Riskier
Pair 1	A_1: (0.75, \$0.00; \$0.75)	B_1: (0.8, \$0.00; \$1.25)
Pair 3	A_3: (0.75, \$0.00; \$1.50)	B_3: (0.8, \$0.00; \$2.50)
Pair 4	A_4: (0.25, \$1.50; \$3.00)	B_4: (0.05, \$0.00; 0.2, \$2.50; \$3.00)
Pair 2	A_2: (1, \$1.50)	B_2: (0.2, \$0.00; \$2.50)
Pair 5	A_5: (1, \$4.50)	B_5: (0.2, \$3.00; \$5.50)

5.2 Experiment 3 Results

Table A4 in the Online Appendix shows standardized differences of observable characteristics between treatments as well as comparisons with subjects that started but did not complete the study.

Table 9 shows the frequency of choosing the safer option (lottery A) by groups of lottery pairs and treatments. Overall, our results are consistent with Allais' certainty predictions. A higher fraction of our subjects chose the (safe) certain alternative when available (pairs 2 and 5). Consistent with another violation of independence, we also find that pair 4 induces more risk-averse behavior than 1 and 3 but less than 2 and 5. One potential explanation for the proportion of safer choices of pair 4 is that pair 4 increases complexity by adding three-outcome lotteries (raising risk aversion compared to 1 and 3) while eliminating the certainty effect (lowering risk aversion compared to 2 and 5). We also find that real payments for all choices increase the frequency of the safer alternative in pairs 1 and 3. One potential explanation for this behavior is that by paying for all uncertain lotteries, we make those payments *more certain*, leading our subjects to be more risk averse. Since pairs 2 and 5 already feature certain payments, pay-all incentives have no effect on their behavior.

Table 9: Observed frequencies (in %) of choices of safer lotteries

	Pair 1 & 3	Pair 4	Pair 2 & 5	Total
Hypothetical & POR	29.15	43.43	53.03	41.53
Hypothetical & PACn	26.61	39.45	54.59	40.37
Hypothetical & PAIn	22.77	47.52	54.46	40.40
Real & POR	23.50	45.00	60.00	42.40
Real & PAC	35.64	54.46	55.45	47.33
Real & PAI	33.00	46.00	53.50	43.80
Total	28.42	45.90	55.16	42.61

Notes: POR, PACn, PAIn stand for pay-one-randomly, pay-all-correlated and pay-all-independently, respectively; n indicates sum of payoffs are divided by the number of tasks; 'Hyp' is short for 'Hypothetical'.

Table 10: Observed frequencies (in %) of choices of safer lottery by lottery pair

	Pair 1	Pair 2	Pair 3	Pair 4	Pair 5	Total
Hypothetical & POR	28.28	46.46	30.00	43.43	59.60	41.53
Hypothetical & PACn	22.94	46.79	30.28	39.45	62.39	40.37
Hypothetical & PAIn	19.80	48.51	25.74	47.52	60.40	40.40
Real & POR	20.00	59.00	27.00	45.00	61.00	42.40
Real & PAC	30.69	50.50	40.59	54.46	60.40	47.33
Real & PAI	32.00	46.00	34.00	46.00	61.00	43.80
Total	25.57	49.51	31.26	45.90	60.82	42.61

Notes: POR, PACn, PAIn stand for pay-one-randomly, pay-all-correlated and pay-all-independently, respectively. n indicates sum of payoffs are divided by the number of tasks. 'Hyp' is short for 'Hypothetical'.

Table 10 is similar to Table 9 but breaks down frequency of safer choice by pair. This table more clearly shows that aggregate frequencies are qualitatively similar to Cox et al. (2015). We do, however, find meaningful differences for specific pairs across the treatments. Moreover, we find less heterogeneity across our treatments. We conjecture that the discrepancy is driven by the larger number of incentive schemes evaluated in Cox et al. (2015) and the smaller sample size they had for each incentive scheme.

We also estimate logit regressions by grouping together similar lottery pairs according to the probability triangle shown in Figure 3. Table 11 shows the marginal effects from these logit regressions of choosing the safer lottery. As argued above, the pay-all mechanisms under full incentivization positively affect the probability of choosing the safer lottery but have no effect under lottery pairs 2 and 5 that involve certainty.

We conclude by analyzing the consistency of behavior with neoclassical theory by counting deviations from choosing the same alternative across the five pairs. On average, subjects exhibit 1.5 violations (0.066 s.e.) under POR. Only hypothetical incentives marginally increased consistency (~ -0.2; 0.1 s.e.) with the canonical assumptions compared to the other incentive schemes. We hypothesize two reasons for this finding. First, given the smallness of the hypothetical rewards, participants were unlikely to exhibit endowment or wealth effects. Second, certainty effects are less likely to be present when the only certainty is that these choices are unincentivized. It is not by chance that Allais' original paradox required quite substantial hypothetical rewards.

Table 11: Marginal effects from logit regressions of choosing the safer lotteries

	Pairs 1, 3		Pair 4		Pair 2, 5	
	(1)		(2)		(3)	
Hypothetical & POR	0.053	(0.052)	-0.016	(0.070)	-0.070	(0.053)
Hypothetical & PACn	0.031	(0.050)	-0.056	(0.068)	-0.054	(0.053)
Hypothetical & PAIn	-0.007	(0.049)	0.025	(0.070)	-0.055	(0.053)
100% & PACn	0.121**	(0.053)	0.095	(0.070)	-0.046	(0.053)
100% & PAIn	0.095*	(0.051)	0.010	(0.070)	-0.065	(0.053)
Observations	1220		610		1220	
AIC	1673.825		1451.210		848.345	
BIC	1709.571		1486.956		874.825	

Notes: Clustered standard errors in parentheses. * p<0.1, ** p<0.05 *** p<0.01. Base category for the treatment variables is the 100% & POR treatment. POR, PACn, PAIn stand for pay-one-randomly, pay-all-correlated and pay-all-independently, respectively; n indicates sum of payoffs are divided by the number of tasks.

6 Conclusions

While most previous work related to incentive payment schemes is theoretical, this paper explored the dynamics of incentive mechanisms in economic experiments using an empirical approach that focuses on value elicitation and choice under risk across three experimental studies with a large sample. Given the abundant literature showcasing empirical deviations from theoretical expectations, we argue that an empirical approach is needed in the incentive scheme argument to voice the outcomes produced by participants in experiments. We found that while the nature of the incentive—hypothetical or real—had minimal impact on participants' bidding behavior or risk preferences, the design elements, such as the magnitude of induced values and the range of offers, significantly influenced outcomes. Specifically, larger induced values and smaller offer ranges led to more accurate bidding, aligning closer to theoretical expectations. Therefore, our results suggest design elements in the experiment environment may influence decision-making more than the incentive mechanism.

Comparing the BDM mechanism with the SPA, the latter showed an improvement in aligning bids with the induced values, indicating that SPA produce less missbidding than the BDM. Decision-making noise and misconceptions about payoff functions were minimal across both auction mechanisms.

An older literature tried to argue that risk preferences may be able to explain differences in bidding behavior. Harrison's (1989) critique was that incentives in auctions were mostly flat, and thus, behavior deviating from theoretical predictions may be driven by insufficient incentives. Our results suggest several considerations to this explanation. First, the source of uncertainty matters, behavioral deviations are different in the BDM and second price auction. So whether a randomization device or other people drive uncertainty matters. Second, certainty effects may be created by the choice of payment mechanism and can lead to subjects behaving in a more risk-averse manner. These certainty effects are more likely driven by the framing of the problem, e.g., paying for two different choices, than the increase produced over the potential rewards. Therefore, *non-expected* risk preferences are likely to affect elicited choices; however, their non-expected nature makes it tricky to pin down the overall effect over the resulting (non-expected) preferences. Because of their elusive nature, careful theoretic modeling and empirical richness are both necessities.

Our findings suggest that the effectiveness of incentive mechanisms in eliciting true preferences in economic experiments is complex. While certain design elements like the magnitude of rewards and range of offers play a crucial role, the choice of elicitation mechanism (BDM vs. SPA) also significantly impacts the accuracy of outcomes. We conjecture that based on our results of Experiment 3, where we document significant certainty effects, the discrepancy between BDM and SPA may be driven by the source of the uncertainty (i.e., a random mechanism versus the bids of other participants). This highlights the need for careful consideration of these factors in experimental design to ensure the reliability and validity of results in economic research.

We conclude by asserting the growing need to understand the complex interplay between cognitive effort and improved (or more revealing) choices. The mounting number of perplexing null results on hypothetical bias can only be explained by securing a tighter grasp on this relationship. We call these results perplexing because even theoretically improper incentives yield identical responses while more opportunities for mistakes and more complex decision objects can exacerbate differences both between and within methods. It is imperative that we hear more about this discussion from empirical studies, to balance the predominantly theoretical nature of the existing literature.

References

Agranov, M. and P. Ortoleva (2017). Stochastic choice and preferences for randomization. *Journal of Political Economy 125*(1), 40–68.

Ahler, D. J., C. E. Roush, and G. Sood (2021). The micro-task market for lemons: data quality on amazon's mechanical turk. *Political Science Research and Methods*, 1–20.

Ahles, A., M. A. Palma, and A. C. Drichoutis (2024). Testing the effectiveness of lottery incentives in online experiments. *American Journal of Agricultural Economicsc (forthcoming)*.

Allais, M. (1953). Le comportement de l'homme rationnel devant le risque: critique des postulats et axiomes de l'école américaine. *Econometrica: journal of the Econometric Society*, 503–546.

Azrieli, Y., C. P. Chambers, and P. J. Healy (2018). Incentives in experiments: A theoretical analysis. *Journal of Political Economy 126*(4), 1472–1503.

Azrieli, Y., C. P. Chambers, and P. J. Healy (2020). Incentives in experiments with objective lotteries. *Experimental Economics 23*(1), 1–29.

Baltussen, G., G. T. Post, M. J. van den Assem, and P. P. Wakker. Random incentive systems in a dynamic choice experiment. *15*(3), 418–443.

Beattie, J. and G. Loomes (1997). The impact of incentives upon risky choice experiments. *Journal of Risk and Uncertainty 14*(2), 155–168.

Benjamin, D. J., M. A. Fontana, and M. S. Kimball (2020). Reconsidering risk aversion. Technical report, National Bureau of Economic Research.

Blavatskyy, P., A. Ortmann, and V. Panchenko (2022, February). On the experimental robustness of the allais paradox. *American Economic Journal: Microeconomics 14*(1), 143–63.

Bolle, F. High reward experiments without high expenditure for the experimenter? *11*(2), 157–167.

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

Breig, Z. and P. Feldman (2023). Revealing risky mistakes through revisions. *Available at SSRN 3975829*.

Briz, T., A. C. Drichoutis, and R. M. Nayga Jr (2017). Randomization to treatment failure in experimental auctions: The value of data from training rounds. *Journal of Behavioral and Experimental Economics 71*, 56–66.

Brown, A. L. and P. J. Healy (2018). Separated decisions. *European Economic Review 101*, 20–34.

Brown, A. L., J. Liu, and M. Tsoi (2023, June 12). Is there a better way to elicit valuations than the BDM?

Buschena, D. and D. Zilberman (2000). Generalized expected utility, heteroscedastic error, and path dependence in risky choice. *Journal of Risk and Uncertainty 20*, 67–88.

Cason, T. N. and C. R. Plott (2014). Misconceptions and game form recognition: Challenges to theories of revealed preference and framing. *Journal of Political Economy 122*(6), 1235–1270.

Chandler, J., C. Rosenzweig, A. J. Moss, J. Robinson, and L. Litman (2019). Online panels in social science research: Expanding sampling methods beyond mechanical Turk. *Behavior Research Methods 51*, 2022–2038.

Charness, G., C. Eckel, U. Gneezy, and A. Kajackaite (2018). Complexity in risk elicitation may affect the conclusions: A demonstration using gender differences. *Journal of risk and uncertainty 56*, 1–17.

Charness, G., U. Gneezy, and B. Halladay (2016). Experimental methods: Pay one or pay all. *Journal of Economic Behavior & Organization 131*, 141–150.

Chmielewski, M. and S. C. Kucker (2020). An MTurk crisis? shifts in data quality and the impact on study results. *Social Psychological and Personality Science 11*(4), 464–473.

Clot, S., G. Grolleau, and L. Ibanez. Shall we pay all? an experimental test of random incentivized systems. *73*, 93–98.

Cochran, W. G. and D. B. Rubin (1973). Controlling bias in observational studies: A review. *Sankhyā: The Indian Journal of Statistics, Series A 35*(4), 417–446.

Cox, J. C., V. Sadiraj, and U. Schmidt (2015). Paradoxes and mechanisms for choice under risk. *Experimental Economics 18*, 215–250.

Cubitt, R. P., C. Starmer, and R. Sugden (1998). On the validity of the random lottery incentive system. *Experimental Economics 1*, 115–131.

Danz, D., L. Vesterlund, and A. J. Wilson (2022). Belief elicitation and behavioral incentive compatibility. *American Economic Review 112*(9), 2851–2883.

Deaton, A. and N. Cartwright (2018). Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine 210*, 2–21.

Diggle, P. J., P. Heagerty, K.-Y. Liang, and S. L. Zeger (2002). *Analysis of Longitudinal Data* (2nd ed.). New York, USA: Oxford University Press Inc.

Drichoutis, A. C. and R. M. Nayga Jr (2022). Game form recognition in preference elicitation, cognitive abilities, and cognitive load. *Journal of Economic Behavior & Organization 193*, 49–65.

Dwenger, N., D. Kübler, and G. Weizsäcker (2018). Flipping a coin: Evidence from university applications. *Journal of Public Economics 167*, 240–250.

Enke, B., U. Gneezy, B. Hall, D. Martin, V. Nelidov, T. Offerman, and J. van de Ven (2023, 07). Cognitive Biases: Mistakes or Missing Stakes? *The Review of Economics and Statistics 105*(4), 818–832.

Feldman, P. and J. Rehbeck (2022). Revealing a preference for mixtures: An experimental study of risk. *Quantitative Economics 13*(2), 761–786.

Feldman, P. J. and P. J. Ferraro (2023). A certainty effect for preference reversals under risk: Experiment and theory. *Working paper*.

Freeman, D. J. and G. Mayraz (2019). Why choice lists increase risk taking. *Experimental Economics 22*, 131–154.

Gneezy, U., A. Imas, and J. List (2015, February). Estimating individual ambiguity aversion: A simple approach. Working Paper 20982, National Bureau of Economic Research.

Grether, D. M. and C. R. Plott (1979). Economic theory of choice and the preference reversal phenomenon. *The american economic review 69*(4), 623–638.

Haaland, I., C. Roth, and J. Wohlfart (2023, March). Designing information provision experiments. *Journal of Economic Literature 61*(1), 3–40.

Hackethal, A., M. Kirchler, C. Laudenbach, M. Razen, and A. Weber (2023). On the role of monetary incentives in risk preference elicitation experiments. *Journal of Risk and Uncertainty 66*, 189–213.

Ham, J. C., J. H. Kagel, and S. F. Lehrer (2005). Randomization, endogeneity and laboratory experiments: the role of cash balances in private value auctions. *Journal of Econometrics 125*(1), 175–205. Experimental and non-experimental evaluation of economic policy and models.

Hansen, R. G. and J. R. Lott (1991). The winner's curse and public information in common value auctions: Comment. *The American Economic Review 81*(1), 347–361.

Harrison, G. W. (1989). Theory and misbehavior of first-price auctions. *The American Economic Review*, 749–762.

Harrison, G. W. (1992). Theory and misbehavior of first-price auctions: Reply. *The American Economic Review 82*(5), 1426–1443.

Harrison, G. W. and J. T. Swarthout (2014). Experimental payment protocols and the bipolar behaviorist. *Theory and Decision 77*, 423–438.

Ho, D. E., K. Imai, G. King, and E. A. Stuart (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis 15*(3), 199–236.

Holt, C. A. (1986). Preference reversals and the independence axiom. *The American Economic Review 76*(3), 508–515.

Holt, C. A. and S. K. Laury (2002). Risk aversion and incentive effects. *American economic review 92*(5), 1644–1655.

Imbens, G. W. and D. B. Rubin (2016). *Causal Inference for Statistics, Social, and Biomedical Sciences, An introduction*. Cambridge and New York: Cambridge University Press.

Imbens, G. W. and J. M. Wooldridge (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature 47*(1), 5–86.

Irwin, J. R., G. H. McClelland, and W. D. Schulze (1992). Hypothetical and real consequences in experimental auctions for insurance against low-probability risks. *Journal of Behavioral Decision Making 5*(2), 107–116.

Johnson, C., A. Baillon, H. Bleichrodt, Z. Li, D. van Dolder, and P. P. Wakker (2021). Prince: An improved method for measuring incentivized preferences. *Journal of Risk and Uncertainty 62*, 1–28.

Kagel, J. H. and D. Levin (1986). The winner's curse and public information in common value auctions. *The American Economic Review 76*(5), 894–920.

Kagel, J. H. and D. Levin (1991). The winner's curse and public information in common value auctions: Reply. *The American Economic Review 81*(1), 362–369.

Karni, E. and Z. Safra (1987). "preference reversal" and the observability of preferences by experimental methods. *Econometrica 55*(3), 675–685.

Kendall, C. and A. Chakraborty (2022). Future self-proof elicitation mechanisms. *Available at SSRN: https://ssrn.com/abstract=4032946.*

Kupper, L. L. and K. B. Hafner (1989). How appropriate are popular sample size formulas? *The American Statistician 43*(2), 101–105.

Li, Y. (2021). The ABC mechanism: an incentive compatible payoff mechanism for elicitation of outcome and probability transformations. *Experimental Economics 24*, 1019–1046.

Li, Z., J. Müller, P. P. Wakker, and T. V. Wang (2017). The rich domain of ambiguity explored. *Management Science 64*(7), 3227–3240.

Lichtenstein, S. and P. Slovic (1971). Reversals of preference between bids and choices in gambling decisions. *Journal of experimental psychology 89*(1), 46.

Litman, L. and J. Robinson (2021). Beyond mechanical Turk: Using online market research platforms. In *Conducting Online Research on Amazon Mechanical Turk and Beyond.*

Liu, H. and T. Wu (2005). Sample size calculation and power analysis of time-averaged difference. *Journal of Modern Applied Statistical Methods 4*(2), 434–445.

Machina, M. J. (1989). Dynamic consistency and non-expected utility models of choice under uncertainty. *Journal of Economic Literature 27*(4), 1622–1668.

Mamadehussene, S. and F. Sguera (2022). On the reliability of the bdm mechanism. *Management Science 69*(2), 1166–1179.

Martínez-Marquina, A., M. Niederle, and E. Vespa (2019). Failures in contingent reasoning: The role of uncertainty. *American Economic Review 109*(10), 3437–3474.

McGranaghan, C., K. Nielsen, T. O'Donoghue, J. Somerville, and C. D. Sprenger (2024). Distinguishing common ratio preferences from common ratio effects using paired valuation tasks. *American Economic Review 114*(2), 307–347.

Moher, D., S. Hopewell, K. F. Schulz, V. Montori, P. C. Gotzsche, P. J. Devereaux, D. Elbourne, M. Egger, and D. G. Altman (2010). CONSORT 2010 explanation and elaboration: Updated guidelines for reporting parallel group randomised trials. *BMJ 340*.

Molnar, A. (2019). SMARTRIQS: A simple method allowing real-time respondent interaction in qualtrics surveys. *Journal of Behavioral and Experimental Finance 22*, 161–169.

Mutz, D. C. and R. Pemantle (2015). Standards for experimental research: Encouraging a better understanding of experimental methods. *Journal of Experimental Political Science 2*(2), 192–215.

Peer, E., L. Brandimarte, S. Samat, and A. Acquisti (2017). Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology 70*, 153–163.

Peer, E., D. Rothschild, A. Gordon, Z. Evernden, and E. Damer (2022). Data quality of platforms and panels for online behavioral research. *Behavior Research Methods 54*, 1643–1662.

Rabin, M. (2013). Risk aversion and expected-utility theory: A calibration theorem. In *Handbook of the fundamentals of financial decision making: Part I*, pp. 241–252. World Scientific.

Samuelson, P. A. (1938). A note on the pure theory of consumer's behaviour. *Economica 5*(17), 61–71.

Segal, U. (1988). Does the preference reversal phenomenon necessarily contradict the independence axiom? *The American Economic Review 78*(1), 233–236.

Serizawa, S., N. Shimada, and T. T. K. Tse (2024, February). Toward an understanding of dominated bidding in a vickrey auction experiment. Discussion Paper 1229, The Institute of Social and Economic Research, Osaka University, 6-1 Mihogaoka, Ibaraki, Osaka 567-0047, Japan. Revised April 2024.

Smith, V. L. (1976). Experimental economics: Induced value theory. *The American Economic Review 66*(2), 274–279.

Stagnaro, M. N., J. Druckman, a. berinsky, A. A. Arechar, R. Willer, and D. G. Rand (2024, Feb). Representativeness versus attentiveness: A comparison across nine online survey samples.

Starmer, C. and R. Sugden (1991). Does the random-lottery incentive system elicit true preferences? an experimental investigation. *The American Economic Review 81*(4), 971–978.

Vassilopoulos, A., A. C. Drichoutis, and R. M. Nayga Jr (2018). Loss aversion, expectations and anchoring in the BDM mechanism. *Munich Personal RePEc Archive No. 85635*.

Online Appendix

Additional Tables

Table A1: Descriptive statistics of $|Bid - IV|/IV$ by induced value and upper limit of the support

	Upper support limit			
	\$4	\$5	\$6	Total
	0.635	0.809	1.028	0.824
IV=1	(0.689)	(0.930)	(1.211)	(0.981)
	[0.500]	[0.500]	[0.500]	[0.500]
	0.197	0.208	0.237	0.214
IV=3	(0.195)	(0.199)	(0.239)	(0.213)
	[0.167]	[0.167]	[0.167]	[0.167]
	0.416	0.508	0.632	**0.519**
Total	(0.552)	(0.737)	(0.958)	**(0.772)**
	[0.223]	[0.237]	[0.282]	**[0.250]**

Notes: Table shows means, standard deviations in parenthesis and medians in brackets.

Table A2: Study 1: Pairwise normalized differences between the Incentives treatments for observable characteristics

	Hypothetical vs.				0.2% vs.			1% vs.		50% vs.	All vs. dropouts
	0.2%	1%	50%	100%	1%	50%	100%	50%	100%	100%	
Gender	0.115	0.003	0.116	0.061	0.117	0.028	0.090	0.118	0.064	0.075	0.064
Age	0.054	0.022	-0.038	-0.075	-0.030	-0.089	-0.124	-0.057	-0.092	-0.035	-0.121
Children	0.117	0.048	0.094	0.063	0.068	0.023	0.054	0.045	0.014	0.031	0.071
Income	0.206	0.191	0.209	0.203	0.182	0.206	0.177	0.140	0.168	0.201	0.175
Marital	0.081	0.098	0.140	0.110	0.114	0.166	0.170	0.069	0.127	0.104	0.033
Education	0.175	0.211	0.103	0.144	0.139	0.115	0.179	0.172	0.207	0.117	0.173
Hispanic	0.042	0.009	0.025	0.022	0.051	0.017	0.021	0.034	0.031	0.003	0.006
Region	0.267	0.312	0.252	0.241	0.121	0.083	0.102	0.106	0.097	0.120	0.108

Table A3: Study 2: Pairwise normalized differences between the Incentives treatments for observable characteristics

	Hyp-PAn vs.			Hyp-POR vs.		Real-PAn vs.	All vs.
	Hyp-POR	Real-PAn	Real-POR	Real-PAn	Real-POR	Real-POR	dropouts
Gender	0.283	0.348	0.285	0.160	0.074	0.231	0.024
Age	0.070	-0.075	-0.190	-0.140	-0.255	-0.106	-0.076
Children	0.050	0.075	0.051	0.025	0.001	0.024	0.057
Income	0.494	0.461	0.366	0.453	0.247	0.399	0.174
Marital	0.185	0.166	0.327	0.225	0.319	0.262	0.081
Education	0.299	0.348	0.408	0.226	0.347	0.276	0.239
Hispanic	0.228	0.162	0.040	0.066	0.188	0.122	0.007
Region	0.336	0.239	0.346	0.097	0.193	0.184	0.123

Table A4: Study 3: Pairwise normalized differences between treatments for observable characteristics

	Hyp-POR vs.					Hyp-PACn vs.				Hyp-PAIn vs.			Real-POR vs.		Real-PACn vs.	All vs.
	Hyp-PACn	Hyp-PAIn	Real-POR	Real-PACn	Real-PAIn	Hyp-PAIn	Real-POR	Real-PACn	Real-PAIn	Real-POR	Real-PACn	Real-PAIn	Real-PACn	Real-PAIn	Real-PAIn	dropouts
Gender	0.025	0.146	0.161	0.121	0.113	0.132	0.157	0.140	0.136	0.070	0.188	0.243	0.151	0.237	0.125	0.174
Age	-0.114	-0.257	-0.148	-0.147	-0.196	-0.149	-0.036	-0.040	-0.090	0.111	0.101	0.052	-0.005	-0.054	-0.047	-0.251
Children	0.219	0.317	0.219	0.081	0.197	0.096	0.001	0.138	0.022	0.097	0.235	0.118	0.137	0.021	0.116	0.166
Income	0.625	0.491	0.446	0.462	0.445	0.497	0.470	0.522	0.484	0.600	0.520	0.519	0.521	0.367	0.476	0.195
Marital	0.279	0.325	0.340	0.379	0.460	0.160	0.223	0.163	0.209	0.070	0.103	0.186	0.136	0.241	0.190	0.131
Education	0.403	0.454	0.522	0.396	0.510	0.595	0.553	0.343	0.417	0.326	0.370	0.393	0.382	0.442	0.311	0.208
Hispanic	0.380	0.114	0.225	0.169	0.030	0.268	0.158	0.213	0.350	0.111	0.056	0.083	0.055	0.194	0.139	0.149
Region	0.365	0.397	0.200	0.192	0.306	0.554	0.435	0.298	0.145	0.508	0.473	0.407	0.369	0.411	0.269	0.160

Table A5: Regressions of bid deviations on treatment variables for the BDM mechanism (with demographics)

	$Bid - IV$		$\lvert Bid - IV\rvert/IV$	
	(3)		(4)	
Constant	0.518***	(0.132)	0.648***	(0.077)
IV = 1 & Support = 5	0.185***	(0.017)	0.167***	(0.015)
IV = 1 & Support = 6	0.410***	(0.022)	0.387***	(0.021)
IV = 3 & Support = 4	-0.768***	(0.019)	-0.440***	(0.014)
IV = 3 & Support = 5	-0.558***	(0.019)	-0.432***	(0.013)
IV = 3 & Support = 6	-0.317***	(0.019)	-0.402***	(0.013)
Hypothetical & PAC	0.001	(0.120)	0.091	(0.073)
Hypothetical & PACn	-0.119	(0.109)	-0.002	(0.066)
Hypothetical & PAI	0.082	(0.110)	0.081	(0.069)
Hypothetical & PAIn	-0.068	(0.105)	-0.013	(0.064)
Hypothetical & POR	-0.037	(0.101)	0.006	(0.065)
0.2% & PAC	-0.098	(0.098)	-0.058	(0.062)
0.2% & PACn	-0.137	(0.107)	-0.020	(0.063)
0.2% & PAI	-0.063	(0.100)	-0.034	(0.063)
0.2% & PAIn	0.036	(0.102)	0.044	(0.062)
0.2% & POR	0.041	(0.107)	0.059	(0.067)
1% & PAC	0.091	(0.105)	0.092	(0.067)
1% & PACn	-0.010	(0.096)	-0.055	(0.062)
1% & PAI	-0.091	(0.094)	-0.046	(0.058)
1% & PAIn	0.070	(0.113)	0.056	(0.070)
1% & POR	-0.047	(0.098)	-0.006	(0.061)
50% & PAC	-0.137	(0.105)	-0.003	(0.062)
50% & PACn	0.047	(0.098)	0.036	(0.060)
50% & PAI	-0.038	(0.100)	-0.007	(0.063)
50% & PAIn	0.077	(0.105)	0.063	(0.064)
50% & POR	-0.059	(0.100)	0.017	(0.060)
100% & PAC	0.034	(0.097)	-0.002	(0.062)
100% & PACn	0.088	(0.103)	0.034	(0.062)
100% & PAI	-0.102	(0.091)	-0.048	(0.057)
100% & PAIn	0.039	(0.099)	0.016	(0.063)
Observations	14256		14256	
R^2	0.164		0.188	
Adj. R^2	0.161		0.185	
F-stat. (p-value)	40.847	(< 0.001)	28.309	(< 0.001)

Notes: Clustered standard errors in parentheses. * p<0.1, ** p<0.05 *** p<0.01. Base categories for the treatment variables are: IV = 1 & Support = 4, 100% & POR.

Table A6: Ordered logit estimates of underbidding, bidding (close to) the IV and overbidding

	(1)		(2)		(3)		(4)	
	1: $Bid < IV$		1: $Bid < 0.99 \cdot IV$		1: $Bid < 0.95 \cdot IV$		1: $Bid < 0.90 \cdot IV$	
	2: $Bid = IV$		2: $0.99 \cdot IV \leq Bid \leq 1.01 \cdot IV$		2: $0.95 \cdot IV \leq Bid \leq 1.05 \cdot IV$		2: $0.90 \cdot IV \leq Bid \leq 1.10 \cdot IV$	
	3: $Bid > IV$		3: $Bid > 1.01 \cdot IV$		3: $Bid > 1.05 \cdot IV$		3: $Bid > 1.10 \cdot IV$	
IV = 1, Support = 5	0.073**	(0.034)	0.073**	(0.034)	0.076**	(0.034)	0.086**	(0.035)
IV = 1, Support = 6	0.171***	(0.036)	0.181***	(0.036)	0.179***	(0.036)	0.193***	(0.037)
IV = 3, Support = 4	-1.140***	(0.043)	-1.201***	(0.043)	-1.287***	(0.043)	-1.340***	(0.044)
IV = 3, Support = 5	-0.878***	(0.041)	-0.946***	(0.041)	-1.028***	(0.041)	-1.053***	(0.042)
IV = 3, Support = 6	-0.544***	(0.039)	-0.582***	(0.038)	-0.627***	(0.038)	-0.650***	(0.039)
Hypothetical, PAC	-0.127	(0.204)	-0.115	(0.201)	-0.096	(0.198)	-0.097	(0.195)
Hypothetical, PACn	-0.357*	(0.206)	-0.310	(0.207)	-0.250	(0.206)	-0.281	(0.201)
Hypothetical, PAI	0.041	(0.204)	0.047	(0.204)	0.027	(0.201)	0.059	(0.199)
Hypothetical, PAIn	-0.156	(0.201)	-0.130	(0.201)	-0.079	(0.202)	-0.066	(0.200)
Hypothetical, POR	-0.050	(0.193)	-0.062	(0.189)	-0.088	(0.182)	-0.087	(0.178)
0.2%, PAC	-0.212	(0.200)	-0.182	(0.197)	-0.170	(0.195)	-0.232	(0.190)
0.2%, PACn	-0.397*	(0.210)	-0.373*	(0.209)	-0.335	(0.207)	-0.351*	(0.203)
0.2%, PAI	-0.030	(0.206)	-0.002	(0.203)	0.012	(0.202)	-0.027	(0.196)
0.2%, PAIn	-0.036	(0.212)	-0.005	(0.209)	-0.019	(0.203)	-0.039	(0.200)
0.2%, POR	0.027	(0.203)	0.058	(0.198)	0.106	(0.198)	0.083	(0.196)
1%, PAC	-0.134	(0.196)	-0.117	(0.194)	-0.113	(0.190)	-0.083	(0.187)
1%, PACn	0.062	(0.198)	0.067	(0.195)	0.078	(0.191)	0.111	(0.188)
1%, PAI	-0.137	(0.198)	-0.097	(0.198)	-0.091	(0.196)	-0.100	(0.193)
1%, PAIn	0.084	(0.202)	0.078	(0.201)	0.082	(0.200)	0.064	(0.197)
1%, POR	-0.038	(0.206)	-0.024	(0.204)	-0.013	(0.201)	-0.044	(0.193)
50%, PAC	-0.209	(0.199)	-0.161	(0.197)	-0.159	(0.197)	-0.143	(0.191)
50%, PACn	0.028	(0.194)	0.034	(0.192)	0.040	(0.189)	0.070	(0.182)
50%, PAI	-0.199	(0.217)	-0.185	(0.214)	-0.170	(0.211)	-0.189	(0.203)
50%, PAIn	0.088	(0.198)	0.102	(0.198)	0.165	(0.198)	0.148	(0.195)
50%, POR	-0.099	(0.200)	-0.073	(0.197)	-0.063	(0.196)	-0.077	(0.197)
100%, PAC	0.054	(0.194)	0.047	(0.193)	0.076	(0.192)	0.057	(0.186)
100%, PACn	0.121	(0.200)	0.161	(0.199)	0.190	(0.198)	0.164	(0.191)
100%, PAI	-0.244	(0.198)	-0.201	(0.192)	-0.207	(0.184)	-0.146	(0.177)
100%, PAIn	-0.017	(0.200)	-0.004	(0.198)	0.003	(0.197)	0.043	(0.194)
Demographics	No		No		No		No	
τ_1	-1.118***	(0.147)	-1.205***	(0.145)	-1.374***	(0.144)	-1.582***	(0.141)
τ_2	-0.437***	(0.147)	-0.359**	(0.144)	-0.257*	(0.143)	-0.156	(0.139)
Observations	15450		15450		15450		15450	
Log-Likelihood	-14961.520		-15379.327		-15775.988		-15926.422	
AIC	29985.041		30820.654		31613.975		31914.843	
BIC	30222.047		31057.660		31850.981		32151.850	

Notes: Clustered standard errors in parentheses. * p<0.1, ** p<0.05 *** p<0.01. Base categories for the treatment variables are: IV = 1 & Support = 4, 100% & POR.

Table A7: Ordered logit estimates of underbidding, bidding (close to) the IV and overbidding (with demographics)

	(1)		(2)		(3)		(4)	
	$Bid \lesseqgtr IV$		$Bid \lesseqgtr 1.01 \cdot IV$		$Bid \lesseqgtr 1.05 \cdot IV$		$Bid \lesseqgtr 1.1 \cdot IV$	
IV = 1, Support = 5	0.087^{**}	(0.036)	0.081^{**}	(0.036)	0.083^{**}	(0.036)	0.086^{**}	(0.037)
IV = 1, Support = 6	0.174^{***}	(0.037)	0.178^{***}	(0.037)	0.176^{***}	(0.037)	0.188^{***}	(0.038)
IV = 3, Support = 4	-1.146^{***}	(0.045)	-1.218^{***}	(0.045)	-1.303^{***}	(0.045)	-1.361^{***}	(0.046)
IV = 3, Support = 5	-0.867^{***}	(0.043)	-0.944^{***}	(0.043)	-1.029^{***}	(0.043)	-1.058^{***}	(0.044)
IV = 3, Support = 6	-0.561^{***}	(0.041)	-0.606^{***}	(0.040)	-0.648^{***}	(0.040)	-0.672^{***}	(0.041)
Hypothetical, PAC	-0.185	(0.208)	-0.177	(0.205)	-0.159	(0.202)	-0.152	(0.200)
Hypothetical, PACn	-0.386^{*}	(0.217)	-0.337	(0.217)	-0.273	(0.215)	-0.310	(0.209)
Hypothetical, PAI	0.074	(0.211)	0.080	(0.210)	0.062	(0.207)	0.096	(0.205)
Hypothetical, PAIn	-0.219	(0.204)	-0.197	(0.205)	-0.139	(0.206)	-0.129	(0.205)
Hypothetical, POR	-0.078	(0.202)	-0.093	(0.198)	-0.105	(0.191)	-0.105	(0.186)
0.2%, PAC	-0.236	(0.210)	-0.206	(0.206)	-0.187	(0.203)	-0.236	(0.199)
0.2%, PACn	-0.337	(0.218)	-0.314	(0.216)	-0.267	(0.215)	-0.275	(0.209)
0.2%, PAI	-0.089	(0.208)	-0.055	(0.205)	-0.032	(0.203)	-0.070	(0.197)
0.2%, PAIn	0.058	(0.215)	0.089	(0.213)	0.064	(0.206)	0.037	(0.203)
0.2%, POR	0.055	(0.209)	0.080	(0.203)	0.136	(0.202)	0.114	(0.200)
1%, PAC	-0.098	(0.201)	-0.089	(0.198)	-0.080	(0.195)	-0.046	(0.192)
1%, PACn	0.062	(0.200)	0.061	(0.197)	0.073	(0.192)	0.110	(0.190)
1%, PAI	-0.179	(0.202)	-0.135	(0.201)	-0.126	(0.199)	-0.134	(0.196)
1%, PAIn	0.168	(0.209)	0.158	(0.207)	0.161	(0.205)	0.161	(0.202)
1%, POR	-0.064	(0.216)	-0.048	(0.215)	-0.037	(0.211)	-0.063	(0.201)
50%, PAC	-0.257	(0.203)	-0.204	(0.201)	-0.199	(0.201)	-0.193	(0.195)
50%, PACn	0.029	(0.196)	0.037	(0.195)	0.044	(0.190)	0.073	(0.184)
50%, PAI	-0.183	(0.216)	-0.173	(0.212)	-0.152	(0.208)	-0.159	(0.200)
50%, PAIn	0.130	(0.206)	0.148	(0.206)	0.212	(0.206)	0.201	(0.203)
50%, POR	-0.147	(0.204)	-0.115	(0.202)	-0.104	(0.201)	-0.107	(0.201)
100%, PAC	0.087	(0.197)	0.076	(0.195)	0.098	(0.194)	0.075	(0.187)
100%, PACn	0.167	(0.205)	0.217	(0.205)	0.258	(0.204)	0.245	(0.197)
100%, PAI	-0.208	(0.205)	-0.165	(0.198)	-0.179	(0.190)	-0.110	(0.181)
100%, PAIn	-0.019	(0.205)	-0.010	(0.203)	-0.0005	(0.202)	0.036	(0.199)
Demographics	Yes		Yes		Yes		Yes	
τ_1	-1.269^{***}	(0.281)	-1.344^{***}	(0.278)	-1.516^{***}	(0.269)	-1.642^{***}	(0.260)
τ_2	-0.575^{**}	(0.281)	-0.476^{*}	(0.277)	-0.375	(0.268)	-0.185	(0.259)
Observations	14256		14256		14256		14256	
Log-Likelihood	-13724.212		-14117.529		-14473.208		-14594.454	
AIC	27568.424		28355.057		29066.416		29308.908	
BIC	28022.320		28808.953		29520.312		29762.804	

Notes: Clustered standard errors in parentheses. * p<0.1, ** p<0.05 *** p<0.01. Base categories for the treatment variables are: IV = 1 & Support = 4, 100% & POR.

Table A8: Regressions of bid deviations on treatment variables for the SPA (with demographics)

	$Bid - IV$		$\lvert Bid - IV\rvert /IV$	
	(3)		(4)	
Constant	0.026	(0.148)	0.415***	(0.055)
IV = 1.7	-0.150***	(0.022)	-0.123***	(0.018)
IV = 2.4	-0.267***	(0.023)	-0.136***	(0.018)
IV = 3	-0.331***	(0.030)	-0.125***	(0.018)
Hypothetical & PAn	-0.040	(0.050)	0.012	(0.020)
100% & PAn	-0.024	(0.043)	0.031*	(0.018)
100% & POR	-0.041	(0.041)	0.006	(0.017)
Observations	1640		1640	
R^2	0.132		0.087	
Adj. R^2	0.114		0.067	
F-stat. (p-value)	11.217	(< 0.001)	6.363	(< 0.001)

Notes: Clustered standard errors in parentheses. * p<0.1, ** p<0.05 *** p<0.01. Base categories for the treatment variables are: IV = 1, Hypothetical and POR.

Table A9: Regressions of bid deviations on treatment variables for the SPA (with subjects matched with bots)

	$Bid - IV$		$\lvert Bid - IV\rvert /IV$		$Bid - IV$		$\lvert Bid - IV\rvert /IV$	
	(1)		(2)		(3)		(4)	
Constant	0.058*	(0.032)	0.262***	(0.017)	-0.052	(0.135)	0.377***	(0.052)
IV = 1.7	-0.155***	(0.018)	-0.119***	(0.015)	-0.154***	(0.018)	-0.114***	(0.015)
IV = 2.4	-0.262***	(0.019)	-0.136***	(0.015)	-0.258***	(0.019)	-0.131***	(0.015)
IV = 3	-0.337***	(0.023)	-0.124***	(0.015)	-0.332***	(0.024)	-0.120***	(0.015)
Hypothetical & PAn	-0.028	(0.042)	0.007	(0.016)	-0.028	(0.044)	0.003	(0.016)
100% & PAn	-0.052	(0.039)	0.016	(0.014)	-0.043	(0.041)	0.016	(0.015)
100% & POR	-0.019	(0.039)	0.008	(0.016)	-0.007	(0.040)	0.002	(0.016)
N of bots=1	0.030	(0.050)	0.002	(0.025)	0.010	(0.051)	0.002	(0.024)
N of bots=2	0.034	(0.045)	-0.025	(0.015)	0.040	(0.047)	-0.030*	(0.016)
N of bots=3	-0.048	(0.045)	-0.001	(0.018)	-0.061	(0.049)	-0.018	(0.019)
Demographics	No		No		Yes		Yes	
Observations	2548		2548		2444		2444	
R^2	0.082		0.054		0.108		0.076	
Adj. R^2	0.079		0.051		0.094		0.062	
F-stat. (p-value)	32.184	(< 0.001)	12.034	(< 0.001)	10.334	(< 0.001)	5.422	(< 0.001)

Clustered standard errors in parentheses. * p<0.1, ** p<0.05 *** p<0.01. Base categories for the treatment variables are: IV = 1, Hypothetical and POR.

Table A10: Maximum likelihood estimates of Decision-making noise and misperception models (with demographics)

	Noise-only model		CP mixture model		γ mixture model			
	(1)		(2)		(3)			
	λ		λ		λ		γ	
Constant	2.869***	(0.687)	3.334***	(0.911)	3.828***	(0.860)	0.874***	(0.057)
IV = 1, Support = 5	-0.766***	(0.129)	-0.835***	(0.187)	-1.225***	(0.197)	0.029***	(0.009)
IV = 1, Support = 6	-1.320***	(0.134)	-1.460***	(0.203)	-2.060***	(0.212)	0.064***	(0.013)
IV = 3, Support = 4	0.966***	(0.235)	0.345	(0.328)	0.723**	(0.368)	0.205***	(0.023)
IV = 3, Support = 5	3.134***	(0.274)	2.684***	(0.358)	2.631***	(0.297)	0.165***	(0.009)
IV = 3, Support = 6	2.709***	(0.221)	3.018***	(0.304)	1.625***	(0.246)	0.075***	(0.008)
Hypothetical, PAC	-1.104**	(0.528)	-1.515**	(0.746)	-1.422**	(0.629)	0.058	(0.051)
Hypothetical, PACn	-0.090	(0.618)	-0.321	(0.836)	-0.202	(0.743)	0.080*	(0.045)
Hypothetical, PAI	-0.671	(0.562)	-0.936	(0.758)	-0.689	(0.723)	-0.026	(0.050)
Hypothetical, PAIn	-0.450	(0.605)	-0.715	(0.843)	-0.759	(0.730)	0.020	(0.049)
Hypothetical, POR	-0.143	(0.622)	-0.382	(0.864)	-0.169	(0.781)	0.032	(0.043)
0.2%, PAC	0.690	(0.668)	1.124	(1.251)	0.852	(0.821)	0.028	(0.042)
0.2%, PACn	-0.105	(0.613)	-0.350	(0.844)	-0.302	(0.717)	0.098**	(0.045)
0.2%, PAI	0.196	(0.699)	0.078	(1.014)	0.274	(0.908)	0.018	(0.042)
0.2%, PAIn	-0.350	(0.529)	-0.558	(0.737)	-0.337	(0.659)	0.002	(0.045)
0.2%, POR	-0.551	(0.555)	-0.869	(0.766)	-0.673	(0.681)	0.038	(0.047)
1%, PAC	-0.828	(0.516)	-1.156	(0.727)	-0.956	(0.636)	0.001	(0.044)
1%, PACn	1.056	(0.881)	1.769	(1.797)	1.563	(1.203)	-0.025	(0.040)
1%, PAI	0.531	(0.592)	0.570	(0.876)	0.730	(0.764)	0.027	(0.042)
1%, PAIn	-0.620	(0.571)	-0.909	(0.782)	-0.732	(0.715)	-0.002	(0.049)
1%, POR	0.076	(0.599)	-0.009	(0.850)	0.103	(0.757)	0.023	(0.044)
50%, PAC	-0.143	(0.576)	-0.422	(0.780)	-0.280	(0.706)	0.070	(0.048)
50%, PACn	-0.435	(0.545)	-0.705	(0.764)	-0.394	(0.700)	-0.014	(0.043)
50%, PAI	0.044	(0.607)	-0.024	(0.888)	0.302	(0.781)	0.022	(0.042)
50%, PAIn	-0.477	(0.536)	-0.695	(0.754)	-0.501	(0.669)	-0.028	(0.049)
50%, POR	-0.120	(0.574)	-0.332	(0.810)	-0.270	(0.722)	0.038	(0.043)
100%, PAC	-0.239	(0.554)	-0.437	(0.787)	-0.130	(0.701)	-0.015	(0.042)
100%, PACn	-0.048	(0.570)	-0.092	(0.835)	0.116	(0.708)	-0.051	(0.046)
100%, PAI	0.716	(0.648)	0.871	(1.049)	0.827	(0.834)	0.025	(0.039)
100%, PAIn	-0.271	(0.541)	-0.457	(0.744)	-0.261	(0.668)	-0.006	(0.043)
			M					
Constant			0.046***	(0.014)				
Demographics	Yes		Yes		Yes			
Observations	14766		14766		14766			
Log-Likelihood	-53251.953		-47245.936		-47020.369			
AIC	106623.907		94613.873		94280.737			
BIC	107079.912		95077.478		95192.747			

Notes: Clustered standard errors in parentheses. * p<0.1, ** p<0.05 *** p<0.01. Base categories for the treatment variables are: IV = 1 & Support = 4, 100% & POR.

Table A11: Regressions of bid deviations (SPA vs. BDM)

	(1) $Bid - IV$		(2) $\lvert Bid - IV\rvert/IV$	
Constant	0.643***	(0.018)	0.797***	(0.015)
SPA	-0.455***	(0.022)	-0.345***	(0.011)
IV High	-0.695***	(0.015)	-0.556***	(0.014)
Observations	17162		17162	
R^2	0.131		0.159	
Adj. R^2	0.131		0.159	
F-stat. (p-value)	1139.662 (< 0.001)		773.956 (< 0.001)	

Clustered standard errors in parentheses. * p<0.1, ** p<0.05 *** p<0.01. Base categories are the BDM mechanism and IV = 1 or IV = 1.7.

Table A12: Regressions of bid deviations (SPA vs. BDM)

	(1) $Bid - IV$		(2) $\lvert Bid - IV\rvert/IV$	
Constant	0.543***	(0.120)	0.817***	(0.067)
SPA	-0.406***	(0.058)	-0.349***	(0.028)
IV high	-0.691***	(0.015)	-0.549***	(0.015)
Observations	16406		16406	
R^2	0.140		0.164	
Adj. R^2	0.138		0.162	
F-stat. (p-value)	68.167 (< 0.001)		45.736 (< 0.001)	

Clustered standard errors in parentheses. * p<0.1, ** p<0.05 *** p<0.01. Base categories are the BDM mechanism and IV = 1 or IV = 1.7.

Table A13: Marginal effects from logit regressions of choosing the safer lotteries (with demographic controls)

	Pairs 1, 3 (1)		Pair 4 (2)		Pair 2, 5 (3)	
Hypothetical & POR	0.039	(0.055)	0.006	(0.071)	-0.050	(0.055)
Hypothetical & PACn	0.020	(0.051)	-0.085	(0.068)	-0.054	(0.053)
Hypothetical & PAIn	-0.006	(0.052)	0.041	(0.071)	-0.042	(0.054)
100% & PACn	0.107*	(0.056)	0.080	(0.071)	-0.032	(0.054)
100% & PAIn	0.093*	(0.055)	0.019	(0.075)	-0.059	(0.054)
Observations	1174		592		1184	
AIC	1647.994		1428.621		850.397	
BIC	1830.754		1606.007		1003.8197	

Clustered standard errors in parentheses. * p<0.1, ** p<0.05 *** p<0.01. Base category for the treatment variables is the 100% & POR treatment.

Sample size calculations

Our per treatment sample size was decided based on sample size calculations and served as a stopping rule for this experiment when we achieved the minimum necessary per treatment sample. Assuming $\alpha = 0.05$ (Type I error) and $\beta = 0.20$ (Type II error), the per group (treatment) minimum sample size required to compare two means μ_0 and μ_1, with common variance of σ^2 in order to achieve a power of at least $1 - \beta$ is given by Diggle et al. (2002) pp. 30; Liu and Wu (2005); Kupper and Hafner (1989):

$$n = \frac{2(z_{1-\alpha/2} + z_{1-\beta})^2(1 + (M-1)\rho)}{M(\frac{\mu_0 - \mu_1}{\sigma})^2} \tag{3}$$

To take into account the repeated measurement, the formula includes the number of repeated measurements M (in our case it is $M = 6$) as well as a value for the correlation ρ between observations for the same subject. For $\alpha = 0.05$ and $\beta = 0.20$ the values of $z_{1-\alpha/2}$ and $z_{1-\beta}$ are 1.96 and 0.84, respectively. To calculate a minimum sample size, one needs to feed the above formula with values for σ and the minimum meaningful difference $d = \mu_0 - \mu_1$. To specify the necessary parameters to feed the above formula, we extracted information from the study of Kendall and Chakraborty (2022), which is as similar as possible to our study in the sense that they also elicit preferences using the BDM mechanism over an online sample in Prolific. We asked the authors to calculate descriptive statistics of $|bid - IV|/IV$ from their BDM treatment to use in our calculations (mean = 0.20, sd = 0.16). In addition, we used values for ρ spanning the range from 0 to 0.9. With roughly 100 subjects per treatment, the minimum effect size we can detect is a 0.05 difference for a correlation of 0.6.

Table A14 shows the result of equation 3 for various values of ρ and d. It is evident that the lower the minimum meaningful difference d and the higher the correlation between periods ρ, a larger sample size is needed to detect the desired effect size with 80% power. We can also detect smaller differences than 0.05 of relative absolute deviations from the IV but one would need to restrict the range of assumed values for ρ.

Table A14: Per treatment sample size calculations for different values of ρ and d

	$\rho = 0$	$\rho = 0.3$	$\rho = 0.6$	$\rho = 0.9$
$d = 0.05$	27	67	107	147
$d = 0.07$	14	34	55	75
$d = 0.1$	7	17	27	37