Psychological Mechanisms for Eliciting Preferences and Beliefs [*]

Evan Friedman[†] Suanna Oh Duncan Webb

First version: 20th November 2023
Current version: 25th November 2023

Abstract

We introduce, and experimentally test, two novel psychological mechanisms for improving answer quality in surveys. The "bonus method" offers additional payments mid-way through the survey, framed as a gift given in exchange for effort. The "recall method" involves telling subjects that they will be paid for restating a randomly selected subset of the answers they previously gave at the end of the survey. The idea is that, by answering questions carefully, subjects may be better able to reconstruct their answers at a later time (without the need for memorization) and thus earn more money. A virtue of both methods is that they can be applied to arbitrary questions. This includes elicitations over "unverifiable" objects which cannot be incentivized through conventional means, e.g. beliefs about states of the world that are unobserved by the researcher or preferences over objects that cannot be "paid out." We find that the recall method has significant promise: relative to both treatments with no incentives and conventional incentives, the method leads to greater internal consistency without reducing the rate at which objective questions are answered correctly.

Keywords: recall method; bonus method; incentives; survey
JEL Classification: C81, C83, C91

[*] We thank seminar participants at briq, NHH FAIR, Erasmus School of Economics, the Berkeley PEP meeting, the Berlin Behavioral Economics Seminar, the MIT Behavioral Lunch, and the Belief-based Utility workshop 2022. We gratefully acknowledge financial support from the Collaborative Research Center (CRC) and from Ecole Universitaire de Recherche - PGSE.

[†] Email: evan.friedman@psemail.eu, suanna.oh@psemail.eu, duncan.webb@psemail.eu

1 Introduction

A central challenge in social science research is being able to accurately elicit subjects' preferences and beliefs. These latent objects are often necessary to test theory or identify structural relationships; or they may be of direct interest, as in surveys and opinion polls. One critical issue with eliciting these objects is that, without proper incentives, subjects may be inattentive and answer questions carelessly (Curran [2016]). This could introduce noise that obscures the underlying relationships between different factors, or it could lead to biased estimates due to subjects relying on effort-saving heuristics.

A number of methods have been devised for incentivizing individuals to provide truthful reports—those that reflect their true preferences or beliefs. For preference-based questions of the form "what do you prefer?", one could simply "pay out" the chosen alternative. Similarly, when eliciting beliefs about an event that is observable to the researcher, subjects can be given higher payments based on their accuracy relative to realized states.[1] We call such questions *verifiable* since they are linked to an observable reality in a way that makes them straightforward to incentivize. However, many important questions are *unverifiable* and therefore cannot be incentivized using conventional means. These include preference-based questions where the alternatives are too costly or too abstract to pay out as well as long-term or counterfactual forecasts.

The literature offers very few mechanisms that can be applied to unverifiable questions. One prominent exception is the Bayesian truth serum (Prelec [2004]), which is based on the idea of peer prediction. If a subject is Bayesian, believes others are telling the truth, and interprets her own truthful answer as a signal about others' answers, being truthful is incentive compatible. While mounting experimental evidence challenges these assumptions (e.g. Benjamin [2019]), we do not argue that these assumptions must be satisfied for the method to be useful. However, the fact that the method requires considerable sophistication from subjects suggests that alternative, simpler methods may work even better and may be easier to implement.

In this paper, we consider two novel mechanisms, rooted in psychological principles, that we hypothesize will increase answer quality in both verifiable and unverifiable questions. The *bonus method* offers additional payments mid-way through the survey,

[1] See Schotter and Trevino [2014] and Schlag et al. [2015] for excellent review articles on belief elicitation.

framed as a gift given in exchange for effort. The *recall method* involves telling subjects that there will be a *recall stage* at the end of the survey: some of the questions will be randomly selected and subjects will be paid for correctly restating the answers they previously gave. The idea is that, for questions that concern stable underlying preferences or beliefs, answering carefully may allow subjects to reconstruct their answers at a later time *without* the need for memorization. Hence, answering carefully may increase subjects' earnings. In a large-scale online experiment ($N = 2,428$), we test these methods against *no incentive* and *conventional incentive* benchmarks.

Our survey consists of two types of questions. The first type are *objective* questions–verifiable questions with correct answers that can be arrived at with sufficient effort (as opposed to requiring outside knowledge). These are used as a direct measure of answer quality. The second type are *belief-based* questions–verifiable or unverifiable questions (depending on the treatment) concerning subjects' beliefs. These come in (obfuscated) duplicate-pairs and are used to measure internal consistency, another measure of answer quality that does not rely on verifiability.[2] We find that, while the bonus method does increase a self-reported measure of "thankfulness" for payment and an incentivized measure of reciprocity toward the experimenters, it does not improve answer quality. The recall method shows greater promise, leading to greater internal consistency in belief-based questions–both verifiable and unverifiable–without reducing the rate at which objective questions are answered correctly.

Having established the main finding–that only the recall method improves answer quality–we take a closer look at auxiliary data to better understand the mechanism. We find that subjects in all incentive treatments (bonus, recall, and conventional incentives) put in more effort as measured by both response times and self reports, but only in the recall treatment do subjects actually have better recall performance–as measured by how many of their previous answers that they correctly restate. This suggests support for the basic mechanism–that subjects are putting in greater effort in answering questions as a means to better construct their answers at a later time, thus increasing their recall bonus. We also find that the recall method is nearly as effective when, instead of being paid money, subjects are paid in "points" with no monetary value. This suggests that subjects are primarily motivated by an intrinsic incentive to perform well. Finally, we

[2] Our measure is related to test-retest reliability (Bland and Altman [1986]), which has a long tradition in clinical measurement.

find that subjects in the recall method treatments continue to have greater internal consistency even in a separate section toward the end of the survey that is explicitly *un*incentivized. Rather than "crowding out" attention once incentives are removed, the positive effects persist.

Because the rate of answering objective questions correctly does not go down with the recall method, there is little evidence that subjects are "gaming" by choosing or writing down memorable answers. One possible concern is that subjects may be answering questions normally and then simply making an effort to memorize their answers as a means to improve recall performance, which could lead to greater internal consistency (even though there is no direct incentive to be consistent). We argue that this is implausible, however, both because of our efforts to obfuscate questions as duplicates and because our results are robust to dropping subjects who claim any role for memory when asked for the strategies they use to improve recall performance.

The paper is organized as follows. We discuss *Related literature* below. Section 2 discusses what we mean by verifiable and unverifiable questions, Section 3 introduces, and offers a conceptual framework for illustrating, the bonus and recall methods, Section 4 gives the experimental design, Section 5 gives our measures of answer quality, Section 6 gives our main results–the effects of treatments on answer quality, Sections 7 and 8 explore the mechanisms, robustness, and heterogeneity underlying the recall method, and Section 9 concludes.

Related literature. The starting point for any such investigation is the long-standing debate about the importance of incentives generally, i.e. the gap between real and hypothetical decisions. In different contexts, researchers have found that the gap can be small (e.g. Beattie and Loomes [1997], Dohmen et al. [2011], Falk et al. [2019], and Enke et al. [2023]), moderate (e.g. Camerer and Hogarth [1999]), or large (e.g. Hertwig and Ortmann [2001]). Of course, if there is little affect of a given incentive scheme, it would be hasty to conclude that incentives do not matter; it may just be that the given incentive scheme is inadequate. There is thus plenty of scope for introducing new methods.

The Bayesian truth-serum (BTS) (Prelec [2004]), based on the idea of peer prediction, is perhaps the first method introduced for incentivizing unverifiable questions. Similar methods have since been introduced (Miller et al. [2005], Radanovic and Faltings [2013], Witkowski and Parkes [2012], Baillon [2017], Cvitanic et al. [2019]), all of which require strong assumptions: Bayesianism and the belief that others are truthful. BTS has been shown to increase the probability psychologists admit to questionable research practices (John et al. [2012]) and reduce overclaiming of knowledge (Weaver and Prelec [2013]). In a context more similar to ours, Baillon et al. [2022] show that BTS can reduce reliance on defaults in surveys, though they find no affect on answer quality in the absence of defaults.

Recent work by Danz et al. [2022] shows that, in some contexts, conventional incentives may actually backfire: in eliciting beliefs over simple events that occur with known probability, incentivizing with the binarized scoring rule (Danz et al. [2022]) makes beliefs *less* accurate. This suggests that simpler methods, that do not require mathematical reasoning or Bayesianism, may do even better. One candidate, more similar in spirit to the methods we introduce, is is the work of Jacquement et al. [2019] who shows that people perform better in discrete choice tasks if they first sign an oath to "faithfully and conscientiously fulfill the tasks."

The methods we introduce, namely the bonus and recall methods, are similarly rooted in psychological principles. The bonus method is based on the well-known "gift exchange" effect (e.g. Fehr and Gachter [2000]), whereby subjects reciprocate by putting in more effort. To the best of our knowledge, this has not been applied to increase truth-telling in surveys. The recall method expands upon a literature in psychology and economics that has explored the relationship between truth-telling and recall (Besken [2018]; Vieira and Lane [2012]; Benjamin et al. [1998]; Zimmermann [2020]). We operationalize this affect: by truth-telling, subjects will recall better, which will earn them more money.

2 Verifiable and unverifiable questions

We make a distinction between *verifiable* and *unverifiable* questions. We define a verifiable question as any that either (1) has an objectively correct answer, (2) concerns preference over objects that can be "paid out" to subjects, or (3) concerns beliefs over events or facts that are observable to the researcher. Being asked to perform a calculation (with an objectively correct answer), to choose a preferred lottery, and to give the probability that a ball drawn from an urn is blue are all verifiable. This is because the calculation is either correct or incorrect, the chosen lottery can simply be "paid out", and the belief report can be compared against the realized ball color. The virtue of verifiable questions is that they are straightforward to incentivize through conventional means: paying for correct calculations, giving the chosen lottery, or using one of the many "scoring rules" for the accuracy of belief reports.[3]

We refer to any question that is not verifiable as *unverifiable*. These include preference elicitations over objects that are abstract or prohibitively expensive to pay out, such as questions concerning subjective well-being. These also include belief elicitations over events or facts that are not observed by the researcher, such as forecasts over the distant future. Not only are unverifiable questions difficult to incentivize, but we also speculate that some subjects may approach them differently. By the very nature of these questions, subjects may feel like there is no correct answer, and for that reason, may not be intrinsically motivated to provide significant effort. The virtue of the methods we introduce is that they can be applied to verifiable and unverifiable questions alike.

3 The bonus and recall methods

In this section, we provide a simple conceptual framework to illustrate the mechanisms through which we posit that the bonus and recall methods operate.

Framework. Suppose a subject is answering a question which either has an objectively correct answer (or an answer that is truthful for that subject, e.g. a preference or belief). There are $n + 1$ possible answers. The subject may have some intrinsic motivation for answering the question well and so receives a payoff of $\gamma_1 \geq 0$ if the answer is correct (or truthful) and a payoff of 0 otherwise ($\gamma_1 = 0$ if the subject only cares about monetary payoffs). The subject puts some amount of costly effort e_1 into answering the question well. Effort $e_1 \in \{L, H\}$ can either be low or high and costs c_{e_1}, where $c_H > c_L = 0$ so that high effort is more costly and the cost of low effort is normalized to 0. Exerting effort e_1 gives the correct answer with probability p_{e_1} where $p_H > p_L > \frac{1}{n+1}$. Hence, higher effort increases the probability of getting the correct answer, but even low effort yields the correct answer more often than would uniformly randomizing. The

[3] See Schotter and Trevino [2014] and Schlag et al. [2015] for excellent review articles on belief elicitation.

overall expected payoff from choosing e_1 is thus $\gamma_1 p_{e_1} - c_{e_1}$, so that it is optimal to exert high effort if and only if $\gamma_1 p_H - c_H \geq \gamma_1 p_L \iff c_H \leq \gamma_1(p_H - p_L)$.

The bonus method. The bonus method is implemented as follows. At the start of the survey (or section of the survey), subjects are told that they will receive additional payments to encourage them to answer questions carefully and with additional effort. In our implementation, the additional payments are in addition to the completion fee and come as a surprise toward the middle of the survey. These payments are explicitly framed as a bonus given in exchange for effort. We posit that this increases the intrinsic incentive to answer questions well, which in terms of our framework would manifest as in an increase in γ_1. Clearly, holding other parameters fixed, subjects will put in high effort for sufficiently large γ_1, and so this is predicted to increase effort for some subjects.

The recall method. The recall method is implemented as follows. At the start of the survey (or section of the survey), subjects are told that there will be a *recall stage* at the end of the survey: some questions will be randomly selected and subjects will be paid for correctly restating the answers they previously gave.

The idea behind the method is simple. If answers are given randomly or inattentively, then the only way to increase the probability of restating them is to commit them to short-term memory, which is cognitively costly. Suppose, on the other hand, that questions concern stable underlying preferences, long-term memories, or reproducible cognitive processes (such as a calculation). In such case, careful or truthful reports can always be reconstructed through introspection or re-engagement with a familiar process. Hence, by being careful or truthful, one can restate their answers, without having to memorize them.

We now illustrate how the recall method can incentivize increased effort in answering questions through a mechanism that does not depend on directly memorizing answers. In the recall stage, we assume that subjects remember how much effort e_1 they exerted when answering questions, but do not remember their answers. Hence, the only way for subjects to restate a previous answer is to answer the question again. We therefore assume that it is as if subjects face a similar problem of having to answer the question. They exert effort $e_2 \in \{L, H\}$ at cost c_{e_2}, which yields the correct answer with probability p_{e_2} and any one of the n incorrect answers with probability $(1 - p_{e_2})/n$. Note, however, that the incentive here is to restate the same answer previously given, not the correct answer, so there is no intrinsic incentive to being correct ($\gamma_2 = 0$). The probability of receiving the bonus with effort profile $(e_1, e_2) \in \{H, L\}^2$ is $p_{e_1}p_{e_2} + (1-p_{e_1})(1-p_{e_2})/n^2$, where the first term is the probability of correctly restating the correct answer and the second term is the probability of correctly restating one of the incorrect answers. Letting the bonus for correctly restating be X and incorporating both effort costs and intrinsic incentive, the expected payoff is

$$U(e_1, e_2) = \gamma_1 p_{e_1} - c_{e_1} + X(p_{e_1}p_{e_2} + (1-p_{e_1})(1-p_{e_2})/n^2) - c_{e_2}.$$

It is easy to show that, if recall bonus X is sufficiently high, then it is optimal to exert high effort in answering the question.

Proposition 1. *It is optimal to exert high effort in answering the question if and only if the recall bonus is sufficiently high: $e_1^* = H$ if and only if*[4]

$$X \geq min\left\{\frac{2c_H - \gamma_1(p_H - p_L)}{(p_H^2 + (1-p_H)^2/n^2 - p_L^2(1-p_L)^2/n^2)}, \frac{c_H - \gamma_1(p_H - p_L)}{(p_Hp_L + (1-p_H)(1-p_L)/n^2 - p_L^2 - (1-p_L)^2/n^2)}\right\}.$$

Proof. See Appendix 10.1. □

We emphasize again that this simplified framework precludes any role for memory: agents are assumed unable to memorize their answers. Of course, if it is not costly to commit answers to memory, then this is likely part of the strategy that real subjects use. However, it is likely that putting effort into answering questions well makes it easier to remember answers, in which case allowing for a role for memory could reinforce the proposed mechanism.

Alternatively, one may be concerned that the recall method pushes people toward using strategies to exploit or "game" the incentives in a way that does not improve, or even reduces, answer quality. Examples of such strategies include choosing easy-to-remember answers, writing down answers, searching for answers online, or choosing answers randomly and then committing those answers to memory.[5] In Section 7, we argue that some subjects do use these strategies to an extent, but that this does not drive our results. Most subjects appear to be putting in more effort toward answering

[4] If this holds with equality, then the subject is indifferent between high and low effort.

[5] As we explain to subjects, we randomize answer options whenever questions are selected for recall, and hence always choosing the first option, say, amounts to randomization.

questions as a means to improve their recall performance.

We also note that, in any survey with diverse question-types such as ours, the strategy required to game a particular question may be specific to special features of that question. On the other hand, the strategy of "simply answering well", which is not question-specific, may be less cognitively demanding to implement over the course of the survey.

4 Experimental design

Overview. We run treatments spanning five incentive schemes: *Recall Method* (RM), *Bonus Method* (B), *Conventional Incentives* (CI), *Recall Method-Points* (RMP), and *No Incentives* (NI).

We use a between-subject design with eight treatments. Each treatment is defined by the incentive scheme and the composition between verifiable and unverifiable questions. For each incentive scheme, we run a *Verifiable* (V) treatment involving only verifiable questions. For each of RM, B, and NI conditions, we also have *Unverifiable* (U) treatments in which some questions are replaced with unverifiable analogues–meant to be as similar as possible except for verifiability (see "V and U treatments" below for more detail). This gives the following eight treatments: RM-V, RM-U, B-V, B-U, NI-V, NI-U, CI-V, and RMP-V. The treatments are summarized in Table 1 along with sample sizes, which range from 299 to 307 subjects.[6]

Note that we include both V and U versions of treatments for all incentive schemes except for CI and RMP, for which we only have V versions. For CI, this is necessary as conventional incentives require verifiability. For RMP, we opted to only include a V treatment to save on costs; its primary purpose is to determine whether it is the monetary incentives in RM that matter or if it is just the intrinsic desire to recall.

All eight treatments have the same five-stage structure, summarized in Table 2. The stages are *Background*, *Main*, *Spillover*, *Recall*, and *Debrief*. Excepting the difference across V and U versions, the questions in the Background, Main, and Spillover stages are identical across all treatments. The Recall stage has the same structure across all treatments and is based on 3 randomly selected questions from the Main stage.

[6] Our pre-registered target sample size was 300 subjects per treatment. We simultaneously invited the same number of subjects to each treatment until we approximately met the targets. The small range in sample sizes across treatments suggests negligible treatment-specific attrition.

Treatment	Short-hand	Number of subjects
Recall Method-Verifiable	RM-V	300
Recall Method-Unverifiable	RM-U	299
Bonus-Verifiable	B-V	306
Bonus-Unverifiable	B-U	307
No Incentives-Verifiable	NI-V	302
No Incentives-Unverifiable	NI-U	307
Conventional Incentives-Verifiable	CI-V	307
Recall Method-Points-Verifiable	RMP-V	300
Total		2,428

Table 1: *Eight between-subject treatments.*

The Debrief stage has questions common to all treatments and some treatment-specific questions.

Nomenclature. Many of our analyses involve using data pooled across V and U treatments. Whenever referring to pooled data, we simply drop "-V" or "-U" from the treatment names. For example, "RM" refers to RM-V and RM-U pooled together, and similarly for the other treatments.

Details of the five stages. The *Background stage* has 30 unincentivized questions, including standard demographics questions, questions about household expenditures and habits, and a single attention check question. The rationale for this section is as follows. We think of incentives as being particularly important in longer experiments where fatigue sets in, so this stage is meant to slightly fatigue subjects in a way that obscures the purpose of the experiment.

The *Main stage* has 24 questions, incentivized according to the treatment. These include (1) *objective* questions–verifiable questions with correct answers that can be arrived at with sufficient effort (as opposed to requiring outside knowledge), and (2) *belief-based* questions–verifiable or unverifiable questions (depending on the treatment) concerning subjects' beliefs over economic and racial inequality, education, and immigration. We also include two questions testing Bayesian reasoning, but these were exploratory and not included in our pre-registered tests.[7] All questions come in pairs. For objective questions, the questions in the pair are distinct, but of a similar type (e.g. two different

[7] Enke et al. [2023] show that failures of Bayesian updating occur at similar rates even under very high incentives.

Raven's Matrices). For belief-based questions, the questions in the pair are (reworded or obfuscated) duplicates. All primary analyses concern the questions in the Main stage, which we use to measure answer quality and thus the effectiveness of different incentive schemes. Because each analysis we conduct uses only a (preregistered) subset of these questions, we write a brief description for all 24 questions in Table 3 with enough detail to uniquely identify each question pair (both questions in the pair are treated symmetrically in the analysis). We refer to these questions as Q1-Q24.

The *Spillover stage* has 10 unincentivized questions. The first 5 are similar to those that appear in the Main stage—3 objective questions and a pair of belief-based questions that are (obfuscated) duplicates that appear at positions 1 and 5. The second 5 questions include a second attention check question and some questions about how subjects answered in the Main stage. The purpose of this stage is to determine whether the effect of different incentive schemes is confined to the incentivized Main stage or if there is a "spillover" effect when the incentives are removed. A priori, we think both positive and negative spillovers are plausible. If incentives increase effort and attention, subjects may be slow to adjust once incentives are removed in which case spillovers may be positive. Or, it may be that, having put in more effort previously due to incentives, removing incentives gives subjects license to thereafter pay less attention in which case spillovers may be negative.

In the *Recall stage*, 3 questions are randomly selected from the Main stage (randomized at the subject level), and subjects are paid for correctly restating their previously given answers. This stage differs slightly across treatments (see "Details of the Recall stage" below for more detail).

The *Debrief stage* has some unincentivized questions about how the incentives affected how questions were answered during the survey. There are questions common to all treatments as well as treatment-specific questions, reflecting differences in incentive structures; there are no more than 13 such questions in a treatment.

V and U treatments. Questions Q1-Q8 are the questions that have verifiable and unverifiable versions in V and U treatments, respectively. To take an example, in V treatments, Q5-Q6 ask for the percentage of income that goes people in the top 20th percentile of income. In U treatments, Q5-Q6 ask the same question, but about people in the year 2050.

Stage	Number of questions
Background	30
Main	24
Spillover	10
Recall	3
Debrief	Up to 13, depending on treatment

Table 2: *Five-stage structure common to all treatments.*

	Type	Short name	Description	Duplicate/Version	Outcome
Q1	Belief-based (verifiable in V, unverifiable in U)	Stop/frisk A	% stopped who are Black/Latinx	Duplicate A	consist.
Q2		Stop/frisk B	% stopped who are not Black/latinx	Duplicate B	consist.
Q3		Immigrants A	State with most illegal immigrants	Duplicate A	consist.
Q4		Immigrants B	State with most illegal immigrants	Duplicate B	consist.
Q5		Top 20% A	% income going to the top 20%	Duplicate A	consist.
Q6		Top 20% B	% income going to the bottom 80%	Duplicate B	consist.
Q7		Education A	Countries' educational spending	Duplicate A	consist.
Q8		Education B	Countries' educational spending	Duplicate B	consist.
Q9	Objective (verifiable)	FOSD A	FOSD-ranked lotteries	Version A	correct.
Q10		FOSD B	FOSD-ranked lotteries	Version B	correct.
Q11		Average A	Word problem: average of 5 numbers	Version A	correct.
Q12		Average B	Word problem: average of 5 numbers	Version B	correct.
Q13		Raven's A	Raven's Matrix	Version A	correct.
Q14		Raven's B	Raven's Matrix	Version B	correct.
Q15		CRT A	Cognitive Reflection Test	Version A	correct.
Q16		CRT B	Cognitive Reflection Test	Version B	correct.
Q17		Dots A	Dot counting	Version A	correct.
Q18		Dots B	Dot counting	Version B	correct.
Q19		Word A	Word problem: sneakers	Version A	correct.
Q20		Word B	Word problem: house	Version B	correct.
Q21		Reading A	Reading comprehension	Version A	correct.
Q22		Reading B	Reading comprehension	Version B	correct.
Q23		Urns A	Bayesian updating urns	Version A	exploratory
Q24		Urns B	Bayesian updating urns	Version B	exploratory

Table 3: *Questions of the Main stage.*

Details of the Recall stage. Subjects are made aware of the Recall stage at the start of the Main stage in RM and RMP-V treatments. In all other treatments, the Recall stage comes as a surprise. The Recall stage payments differ slightly across treatments. In the RM treatments, subjects were paid £1 (1 British Pound) for each of the 3 questions that they correctly restated their answers to (and so received up to £3 total, on top of the completion fee). In RMP-V, they received 1 "point" for each of the 3 questions correctly restated, where each point has no monetary value. To save on costs, most subjects in other treatments received only £0.10 for each of the 3 correctly restated questions. In the NI treatments, subjects were cross-randomized so that 50% of subjects received £1 per correctly restated answer, and the other 50% received only £0.10 per correctly restated answer. This allows us to test whether recall performance is affected by incentives conditional on the Recall stage coming as a surprise. We find no such effect.

Details of incentive schemes. Different incentive schemes applied only to the Main stage. Hence, with one minor exception (see "Payment" below), the survey was identical across treatments until the start of the Main stage.

In the RM treatments, subjects were told about the Recall stage, i.e. that they would receive £1 for each of up to 3 correctly restated questions from the Main stage. In RMP-V, subjects were told about the Recall stage, i.e. that they would receive 1 point for each of up to 3 correctly restated questions. In the other conditions, the Recall stage came as a surprise.

In the B treatments, subjects were told that they would receive £7 (instead of £4) for completing the survey. This was framed as a bonus meant to increase effort.

In CI-V, subjects were told they would be paid based on their answers for 3 randomly selected questions from the Main stage: if the question had an objectively correct answer, they would receive £1 for giving the correct answer; if the question asked subjects to choose their preferred object (e.g. a lottery), they would receive their chosen object.

In the NI treatments, subjects were simply asked to put effort into answering the questions well.

Instructions. It is only at the start of the Main stage that instructions differ across treatments, reflecting the different incentives. We think of these instructions as being particularly important for our purposes as we are testing psychological mechanisms. For instance, it is plausible that the B treatments are most effective if the purpose for giving the bonus–i.e. to incentivize effort–is stated explicitly. Hence, in all treatments, we give each incentive scheme its "best chance" by stating that the purpose of the incentive scheme is to encourage effort and giving explanation for why putting in effort may increase rewards. We also ask subjects outright to answer questions carefully, using a similar language across all treatments. Appendix 10.4 gives the Main stage instructions for each of the treatments, and the surveys in their entirety are available upon request.

Payment. All subjects received at least £4, which was the advertised completion fee. At the start of the experiment, before giving consent to participate, all subjects were told again that they would receive £4, except for subjects in the NI and RMP-V treatments who were told that they would receive £7. The language was identical across the treatments, except with "7" replacing "4" in the NI and RMP-V treatments. In particular, it was not emphasized that these subjects would receive more than advertised. The purpose of this was to ensure that subjects in these unincentivized treatments earned as much money as in the incentivized treatments, so we do not bias our results in favor of the novel incentive schemes. In particular, the only difference between NI and B treatments is the timing of the announcement of the additional £3 and the framing of this as a bonus to encourage effort in the latter. On average, subjects received £7 for an experiment lasting 40 minutes, with some difference in length across treatments.

Prolific. The experiment took place on the Prolific online platform. Subjects were adults based in the US. Prolific was used because it is known to have high data quality relative to similar platforms (Peer et al. [2022]). Hence, we suspect that the effects of incentives we report are likely to be "lower bounds" for those on other platforms.

Multiple choice questions; randomization. All questions in the Main stage were multiple choices questions, except for Q15-Q16 which were standard cognitive reflection test (CRT) questions (Frederick [2005]). The order of questions in the Background and Main stages was randomized at the subject level. The order of questions in the Spillover and Debrief stages was fixed.

5 Measures of answer quality

Our primary goal in this paper is to determine the effectiveness of different incentive schemes. Hence, the Main stage is comprised of questions that lead to measures of answer quality. We consider two measures. The first is *correctness*–whether an objective question is answered correctly. The second is *consistency*–whether the same answer is given across a pair of duplicate belief-based questions. Correctness requires that the questions be verifiable as we–the researchers–must know if the question is answered correctly. Consistency does not require verifiability, and so we use it to measure answer quality in both verifiable questions (in the V treatments) and unverifiable questions (in the U treatments). Hence, by using both measures, we can determine the effectiveness of the incentive schemes in both verifiable and unverifiable questions. We give greater detail below.

Correctness. For each of objective questions Q9-Q22, we create a correctness indicator for if the question is answered correctly. All 14 questions are distinct, but they come in 7 pairs of similar question types: Q9-Q10 ask to choose a preferred lottery from a set with one that stochastically dominates, Q11-Q12 are word problems where one has to average five numbers, Q13-Q14 are Raven's matrices, Q15-Q16 are cognitive reflection test questions, Q17-Q18 are dot-counting tasks,[8] Q19-Q20 are word problems which require both reading carefully and a simple computation, and Q21-Q22 are reading comprehension questions. These questions are summarized in Table 3. Our preregistered analysis is based on pooling all 14 questions.

Remark 1. All correctness questions Q9-Q22 can be answered correctly by simply putting in more effort, e.g. the dot counting tasks Q17-Q18. They do not require knowledge of facts that subjects may, or may not, have been exposed to, nor do they require Bayesian reasoning.[9]

Consistency. Another measure we consider is internal consistency, which refer to simply as "consistency." If preferences or beliefs are stable over the survey horizon, then any variation across multiple elicitations is due to measurement error. Hence, greater

[8] The dot-counting tasks are simplified versions of those that appear in Dewan and Neligh [2020].

[9] We did include a pair of Bayesian updating questions (Q23-Q24) for exploratory purposes, but these were not part of our preregistered comparisons as Enke et al. [2023] show that rates of failure of Bayesian updating are little affected by incentives.

consistency implies lower measurement error and thus greater answer quality. Similar ideas have a long tradition in clinical measurement (Bland and Altman [1986]) and, more recently, in economics (Gillen et al. [2019]).

Using the belief-based questions Q1-Q8, which come in 4 duplicate pairs, we create an indicator for if the same responses are given to both questions in the pair. The questions in the pair are reworded or obfuscated duplicates. Q1-Q2 ask for the percentage of people stopped as part of New York City's "stop and frisk" policy that are Black or Latinx. Q1 asks for the percentage that are Black/Latinx; Q2 asks for the percentage that are neither black nor Latinx; and to be consistent, answers must sum to 100%. Q3-Q4 ask for the U.S. State with the largest number of illegal immigrants, worded in different ways. Q5-Q6 ask for the share of income going to the top 20% of earners. Q5 asks for the share going to the top 20%; Q6 asks for the share going to the bottom 80%; and to be consistent, answers must sum to 100%. Q7-Q8 ask for the Country that spend the most on education, worded in different ways. These questions are summarized in Table 3. Our preregistered analysis is based on pooling all 4 question pairs.

6 Main results: correctness and consistency

In this subsection, we report our main results concerning correctness and consistency. These analyses (specification and questions used) exactly follow our preregistration.

For correctness, we use the questions Q9-Q22 that do not vary across V and U versions of treatments. For this reason, we pool across V and U versions and therefore consider 5 pooled treatments: NI, B, CI, RM, and RMP.[10] We regress the correctness indicators on 4 treatment dummies (NI being the excluded category) with question fixed effects, clustering standard errors at the subject level. The coefficients, interpretable as the difference in average correctness between each treatment and NI, are plotted in Figure 1.

We find that there is *no effect* of any treatments on correctness relative to NI. The

[10] Q9-Q22 are exactly the same across V and U versions, however, one might be concerned that having verifiable or unverifiable versions of Q1-Q8 may affect answers for Q9-Q22, in which case it may not be appropriate to compare pooled results from NI, B, and RM to pooled results from CI and RMP, which only have V versions. This turns out not to be a concern: results are nearly identical, albeit with larger standard errors, if we only consider V versions.

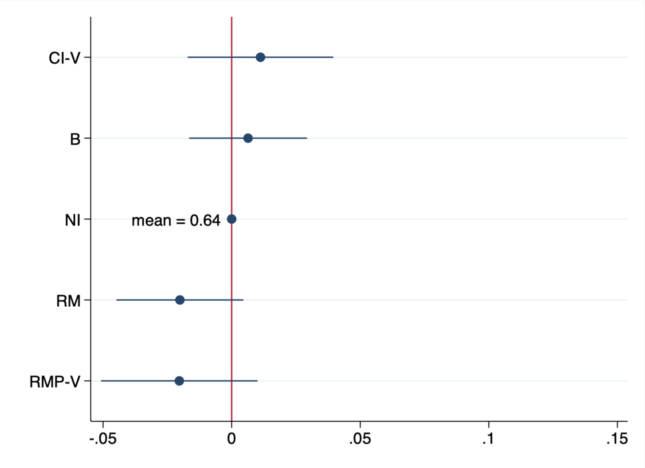


Figure 1: *Correctness.*

estimates range from -0.02 to 0.01, none of which are significant. Furthermore, these are precise nulls, with 95% confidence intervals ranging from 0.05 to 0.06 in width.

Result 1. There is no effect of any treatment on correctness, and these are precise nulls.

For consistency, we use the questions Q1-Q8. Since these questions differ across V and U versions of treatments, we present consistency results separately for each version. Hence, we compare B-V, CI-V, RM-V, and RMP-V to NI-V; and separately compare B-U and RM-U to NI-U. We regress the consistency indicators on treatment dummies (NI being the excluded category) with question fixed effects, clustering standard errors at the subject level. The coefficients, interpretable as the difference in average consistency between each treatment and NI, are plotted in Figure 2.

We find that there is a *large effect* of the Recall method on correctness, relative to NI. RM-V and RM-U have estimates of 0.08 and 0.09 (12.9% and 14.5%), which are highly significant. RMP-V, i.e. without monetary incentive, also has a large effect of 0.07 (11%). No other treatments have a significant effect.

Result 2. There is a large effect of the recall method on consistency. This is true for verifiable and unverifiable questions, with and without monetary incentives.

These results suggest that only the recall method has a significant effect on subject responses. However, it may be that other treatments do affect the distribution of an-

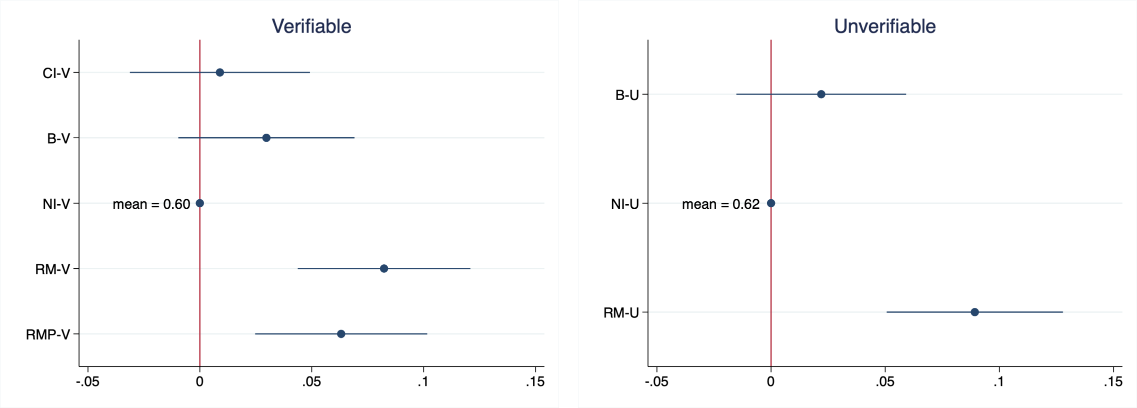


Figure 2: *Consistency.*

swers, but this is not reflected in these aggregates. Hence, we visualize the data in a more informative way. To this end, we calculate for each subject: a (1) *correctness index* ranging from 0 to 14 that gives the total number of correct answers and a (2) *consistency index* ranging from 0 to 4 that gives the total number of question pairs answered consistently. We thus associate two indices to each subject derived from our pre-registered measures. In Figure 3, we plot the correctness index versus the consistency index for each incentive condition superimposed with the same for the matched NI treatment: RM in the top-left panel, B in the top-right panel, RMP-V in the bottom-left panel, and CI-V in the bottom-right panel.

Comparing B and CI-V to the matched NI treatments in the two right panels of the figure, we see that the incentives have very little affect on the joint distribution of correctness and consistency indices. This suggests that, not only do these incentive schemes have no affect on aggregate measures, they also do little to affect the entire distribution of answers. However, when comparing RM and RMP-V to the matched NI treatments in the two left panels, we see that, indeed, the joint distribution of correctness and consistency indices changes considerably.

Result 3. Only the recall treatments have an effect on the joint distribution of subject-level measures of correctness and consistency.

Remark 2. While the bonus method does not affect the distribution of choices, we show in Appendix 10.3 that it does lead to an increase in both "thankfulness" for payment and (incentivized) willingness to answer additional questions. This suggests that the bonus

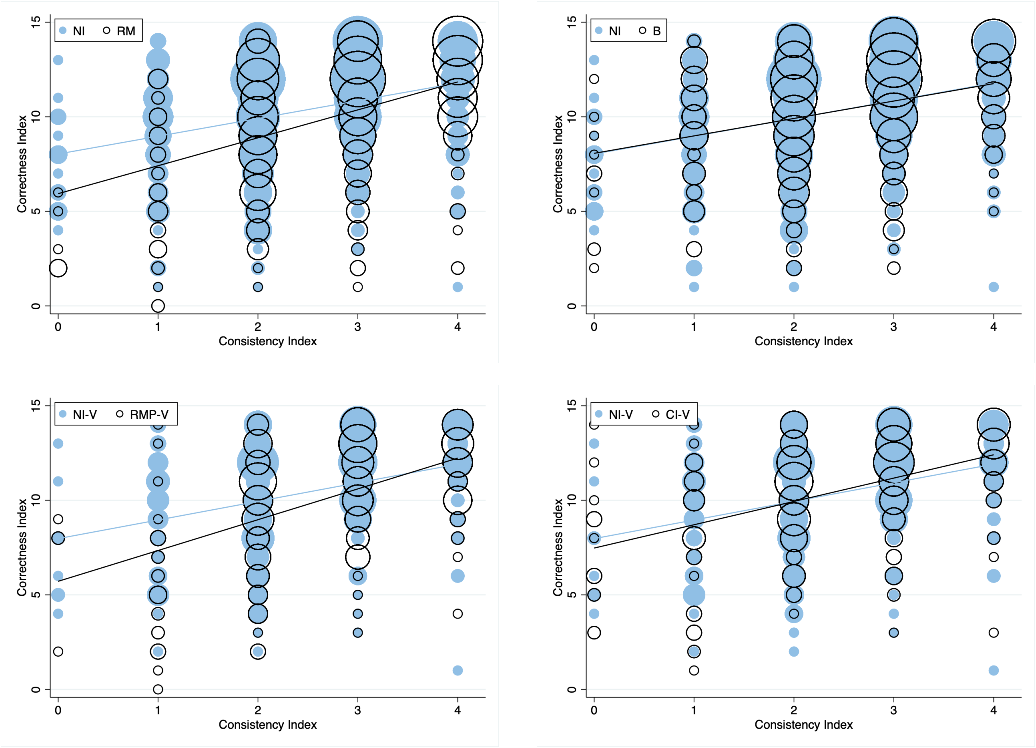


Figure 3: *Correctness vs. consistency.*

method may be a costless way of increasing survey length without a reduction in data quality.

Another salient feature emerges from Figure 3: a strong correlation between correctness and consistency in all treatments. This is our first indication that consistency, like correctness, captures a sensible measure of answer quality.

Result 4. In all treatments, correctness and consistency are strongly correlated: subjects who tend to be more correct also tend to be more consistent.

On the basis of Result 3, we focus the rest of the paper on understanding the recall method using auxiliary data. Section 7 focuses on the important question of whether answer quality is actually improving under the recall method–or whether the increased internal consistency we observe is the result of subjects responding to incentives in unexpected ways (e.g. gaming). Section 8 explores mechanisms in order to make practical recommendations for survey design.

7 The recall method: is answer quality *actually* improving?

We observe that the recall method leads to increased consistency without a reduction in correctness. We interpret this as an increase in answer quality. However, one concern is that the increase in consistency may not reflect improved answer quality, but is simply an artifact of strategies that subjects use to improve recall performance. In this section, we argue that this is not the case: the increase in consistency *does* reflect higher quality responses.

7.1 Gaming does not drive the results

We use the term "gaming" to refer to any strategy that modifies the way in which questions are answered–other than through increased attention or effort–to increase the probability of getting the recall bonus. In particular, we are concerned with strategies that may bias answers so that they are a worse reflection of underlying preferences or beliefs, or are more likely to be incorrect in the case of objective questions.

Our first indication that we should not be very concerned about gaming is that the recall method does not reduce correctness (Section 6). Hence, while there still may be some gaming, whatever form it takes must be relatively benign.

We consider three plausible gaming strategies: (1) choosing answers that are easier to remember, (2) searching for answers online, and (3) writing answers down. For each of these strategies, we ask RM- and RMP-subjects in the Debrief stage whether they engage in the strategy at all, and if they do, for what percentage of questions did they employ the strategy. In eliciting percentages, subjects were asked to report intervals of five percentage points: 0-5%, 5-10%, ... , 95-100%. For each of the three gaming strategies, Table 4 reports the share of subjects who state using the strategy at least once and the average percentage of questions affected. For the latter, we report the upper bound (e.g. 0-5% is coded as 5%) to be conservative.

Gaming strategy	Treatment	Percent. of subjects	Percent of questions
Choosing easy-to-remember answers	RM	14.5	4.0
	RMP	11.0	2.7
Searching for answers online	RM	4.3	0.9
	RMP	3.3	0.6
Writing answers down	RM	5.8	2.5
	RMP	3.3	1.4

Table 4: *Self-reported gaming.*

We do find that some subjects report choosing easy-to-remember answers at least once, but that this affects no more than 4% of questions in RM and 2.7% in RMP. The use of the other strategies are negligible. In any case, we show in Appendix 10.2 that our results are nearly unaffected by dropping all subjects who report using any of these gaming strategies at least once.

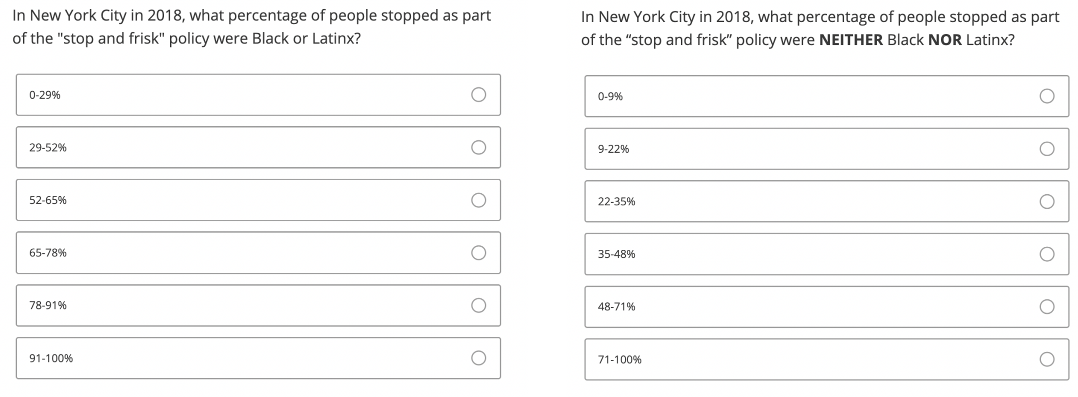
Result 5. Based on self-reports, there is some evidence of gaming, but the results are unaffected by dropping subjects who report any gaming whatsoever. We conclude that gaming does not drive our results.

7.2 Subjects are not simply memorizing their answers

Related to gaming, another concern is that subjects are answering questions normally and then simply memorizing their answers *and* repeating them when they recognize a question as being a duplicate. This would generate increased consistency, without affecting correctness, and represents–in our view–the greatest threat to interpreting consistency as a measure of answer quality.

A priori, there are reasons to doubt that this is happening to a significant degree. The first reason is that the recall method does not give any explicit incentive to be consistent. It is of course possible that subjects memorize their answers in order to improve their recall performance, increasing the probability they recognize duplicates; and then try to repeat their answers as a means to make recall easier. However, we show in Section 8.2 that recall performance is only slightly better in RM than in B, so we think subjects are likely able to recognize duplicates at similar rates across incentive treatments. The second reason we doubt memorization of answers is driving consistency is that some duplicates are heavily obscured. In such cases, we believe few subjects will recognize that the questions are duplicates and, even if they do, memorization could backfire. As an example, consider the pair of duplicate questions Q1 and Q2, whose screenshots are given in Figure 4. Q1 asks for the percentage of people stopped as part of New York City's "stop and frisk" policy who were Black or Latinx, with answer options being intervals of possible percentages. Q2 asks about the percentage stopped who were neither Black nor Latinx. In this example, being consistent requires not giving the same answer, but the complementary answer. Hence, one would need to (1) remember their answer (which is not a simple number), (2) recognize the duplicate, and (3) subtract their previously chosen interval from 100.

We also asked RM subjects, in the Debrief stage, the following open-ended question: "What was your main strategy for making sure you could correctly restate your answers?" There were, indeed, some subjects who said they made an effort to memorize their answers. This does not imply that they gave the same answers on duplicate questions, nor even recognized the existence of duplicates. In any case, we think of these subjects as the ones most likely to be using this strategy. Some open-ended answers are difficult to interpret, but when in doubt, we take a very broad view and assume some role for memorizing answers. In all, we identified about 34% of subjects across RM and RMP



In New York City in 2018, what percentage of people stopped as part of the "stop and frisk" policy were Black or Latinx?

- 0-29% ○
- 29-52% ○
- 52-65% ○
- 65-78% ○
- 78-91% ○
- 91-100% ○

In New York City in 2018, what percentage of people stopped as part of the “stop and frisk” policy were **NEITHER** Black **NOR** Latinx?

- 0-9% ○
- 9-22% ○
- 22-35% ○
- 35-48% ○
- 48-71% ○
- 71-100% ○

Figure 4: *Obscured duplicate questions.* For this question pair (Q1-Q2), consistency goes from 38% in NI to 49% in RM.

treatments as using or possibly using memorizing answers as part of their strategy. We show in Appendix 10.2 that our main results are little affected by dropping all such subjects. In fact, consistency slightly *increases* after dropping these subjects.

Result 6. Based on self-reports, some subjects claim to be memorizing answers as part of their strategy to improve recall performance, but the results are robust to dropping these subjects. We conclude that an increase in consistency is not driven by subjects who simply memorize their answers.

8 The recall method: mechanisms

In this section, we consider auxiliary data to better understand the mechanisms underlying the recall method.

8.1 Subjects are putting in more effort

We hypothesize that the channel through which incentives may improve answer quality is through increased attention and/or effort. Response times are commonly used as measures of both, and we find that, indeed, all incentive schemes we consider do in fact lead to increased response times. Figure 5 plots average response times in Background and Main stages across different treatments after dropping subjects whose average response times in either stage exceed the 99th percentile for that stage.[11] This suggests that subjects are responding to incentives through increased attention/effort; however, as previously shown, this is only manifested in choice in the RM treatments.

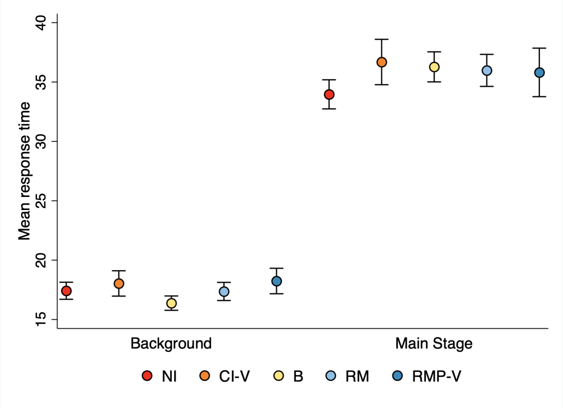


Figure 5: *Effect of incentives on response times.*

Another measure of attention/effort is based on self-reports from the Debrief stage. We asked subjects outright how the particular incentive that they were exposed to affected attention and effort. In RM treatments, for example, we asked: "You were told during the survey that you would be paid for correctly restating some of your answers from Section 2. How did this affect how much **attention** you paid when answering the questions in that section?" Answer options ranged from "Greatly decreased attention", coded as 1, to "Greatly increased attention", coded as 5. We also asked the same question replacing "attention" with "effort," and there were analogous questions for all the treatments.[12] To reduce measurement error, we average answers from the attention and effort questions, yielding an index ranging from 1 to 5 for each subject. Figure 5 plots averages of the index for each treatment, revealing that all incentive treatments significantly increase self-reported attention/effort relative to NI. In particular, RM gives the highest value, which is significantly, or marginally significantly, higher than in the other

[11] As with all online surveys, there are a few subjects with extremely high response times, most likely indicating that they left their computers for extended periods. The result requires removing such extreme subjects, but is robust to many variations. The result also holds if we drop subjects who have extremely low response times.

[12] In NI treatments, subjects were asked: "You were told at the beginning of the survey that your payment for completing the survey would be £7 instead of the £4 which was advertised on Prolific. How did this affect how much **attention** you paid when answering the questions in Section 2?"

incentive treatments.

Result 7. All incentive treatments increase attention and effort (relative to NI), as measured by response times and self-reports.

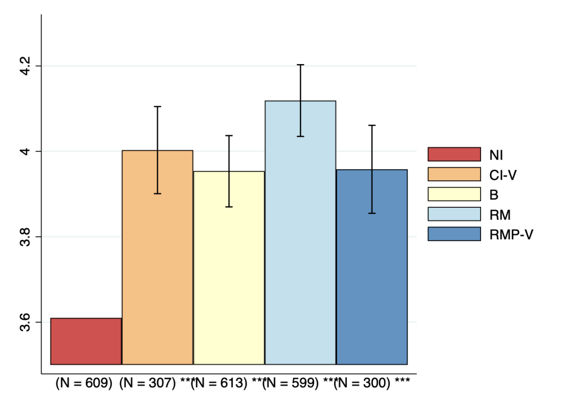


Figure 6: *Effect of incentives on attention/effort.*

8.2 Subjects are recalling more

The explicit incentive under the recall method is to restate answers correctly during the recall stage. Hence, it ought to be that recall performance improves under the recall method. Otherwise, one might be concerned that the recall method would be ineffective as subjects would expect to find it easy to recall without putting additional effort into answering questions.

Remember that, while the recall stage comes as a surprise in the non-RM treatments, all subjects participate in the recall stage. Hence, we can directly compare recall performance across treatments. As shown in Figure 7, we see that there is a small, but highly statistically significant effect of the recall method on recall performance, relative to NI. In NI and CI-V, subjects correctly recall approximately 2.6 (out of 3) questions. In RM, RMP-V, and B, subjects correctly recall approximately 2.7 questions, with the performance being slightly higher in RM. Hence, while recall performance is high even when the recall stage comes as a surprise, the recall method does improve recall performance, consistent with the proposed mechanism.

Result 8. The recall and bonus methods increase recall performance relative to NI.

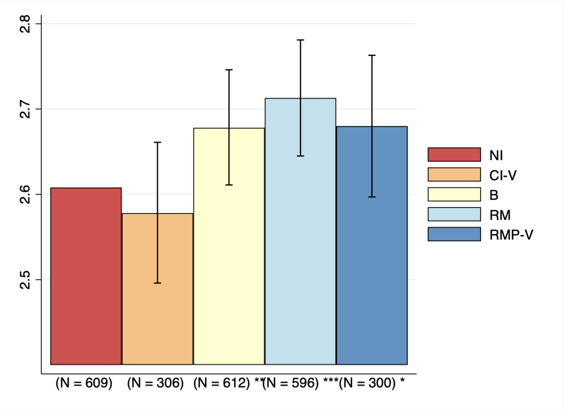


Figure 7: *Recall performance across treatments.*

8.3 Effects of recall persist even after the incentive is removed

Taking place between the Main and Recall stages, the Spillover stage consists of 10 verifiable questions that are explicitly *un*incentivized, as emphasized to subjects. The Spillover stage contains 3 questions for which we can measure correctness and 2 questions–a pair–that allow us to measure consistency. Figure 8 replicates Figures 1 and 2 for these questions, and we find very similar results as in the main stage: the recall method leads to increased consistency without a decrease in correctness.

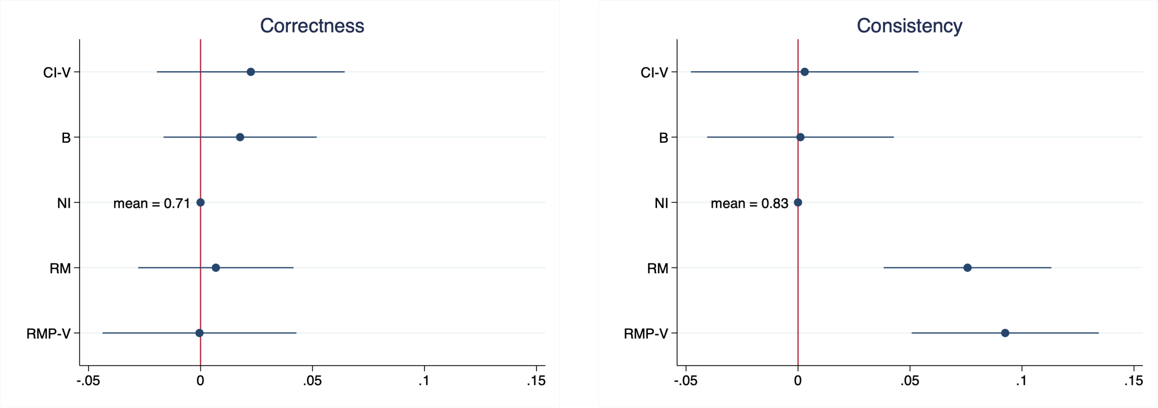


Figure 8: *Effects of recall persist even after the incentive is removed.*

Result 9. The effects of the recall method persist even after incentives are removed: consistency remains high without a reduction in correctness.

A priori, the result could have gone the other way: having put in extra effort when the incentives were in place throughout the survey, subjects may have put in even less effort once the incentive was removed. That we do not find this suggests that subjects do not want to change their "mode" of answering questions, perhaps because the very act of changing modes is cognitively demanding. That the recall method does not "crowd out" attention is encouraging, suggesting the potential for longer experiments with mixed incentive schemes.

8.4 The recall method works without money

The only difference between RM and RMP treatments is that subjects in RM are paid £1 for each correctly restated answer whereas subjects in RMP are paid 1 "point" with no monetary value. As previously mentioned, it is emphasized to subjects that points are of no monetary value, but represent high performance (see Appendix 10.4 for instructions). Interestingly, we find that subjects in both treatments act similarly in terms of correctness, consistency, response times, and recall performance, which can be seen in Figures 1, 2, 5, and 7 respectively. It is true that, relative to RMP, RM yields slightly higher (insignificant) point estimates in terms consistency, response times, and recall performance, but the results are substantially the same. Hence, while it may be that some subjects are more motivated by financial incentives, we conclude that subjects are predominantly intrinsically motivated.

Result 10. The effects of the recall method are similar with and without financial incentives: RM and RMP yield similar results in terms of correctness, consistency, response times, and recall performance.

9 Conclusion

In this paper, we introduce, and experimentally test two novel psychological mechanisms for incentivizing truth-telling in surveys. The bonus method offers additional payments mid-way through the survey, framed as a gift given in exchange for effort. The recall method involves telling subjects that they will be paid for restating a randomly selected subset of the answers they previously gave at the end of the survey. The idea is that, by answering questions carefully, subjects may be better able to reconstruct their answers at a later time (without the need for memorization) and thus earn more money.

We find that, while the bonus method does increase both self-reported "thankfulness" for payment and an incentivized measure of reciprocity toward the experimenters, it does not improve answer quality. The recall method shows greater promise, leading to greater internal consistency in belief-based questions without reducing the rate at which objective questions are answered correctly. This is true for both verifiable and unverifable questions.

We find support for the basic mechanism of the recall method. Subjects are putting in more effort as a means to increase their recall performance, which leads to higher answer quality, and greater recall performance. A primary concern is that subjects are "gaming" the recall incentive, but we find little evidence for this: the recall method does not reduce the rate at which subjects correctly answer objective questions; and dropping subjects who self-report gaming strategies has no effect on the results. We also find that the recall method is similarly effective when subjects are paid in non-monetary "points," suggesting that the method can be used to improve answer quality *costlessly*. Finally, we find that the positive effects persist even after the incentives are removed, suggesting potential application in complex surveys with multiple sections and mixed incentive schemes.

Many experiments find that conventional incentives have little affect on behavior. Indeed, we also replicate this in our setting: conventional incentives increase measures of attention, but have no affect on choices. And yet, we find large and statistically significant effects of the novel recall method. The fact that we find an effect *at all* suggests the recall method's potential as well as the potential of other psychological mechanisms. We conjecture that conventional incentives, which reward performance directly, have little affect because subjects already feel an intrinsic incentive to perform well. By contrast, psychological mechanisms such as the recall method give subjects another dimension on which to perform well. This raises effort, which can be channeled toward improving answer quality.

10 Appendix

10.1 Omitted proofs

Proof of Proposition 1. Writing out the expressions for the expected utility from $(e_1, e_2) \in \{L, H\}^2$ yields:

$$\begin{aligned}
U(L, L) &= \gamma_1 p_L + X(p_L^2 + (1 - p_L)^2/n^2) \\
U(L, H) &= \gamma_1 p_L + X(p_L p_H + (1 - p_L)(1 - p_H)/n^2) - c_H \\
U(H, L) &= \gamma_1 p_H + X(p_H p_L + (1 - p_H)(1 - p_L)/n^2) - c_H \\
U(H, H) &= \gamma_1 p_H + X(p_H^2 + (1 - p_H)^2/n^2) - 2c_H
\end{aligned}$$

Inspection reveals that $U(H, L) \geq U(L, H)$ and strictly so if $\gamma_1 > 0$, the intuition being that both $(e_1, e_2) = (H, L)$ and $(e_1, e_2) = (L, H)$ give the same probability of receiving the recall bonus at the same total effort cost, but the former increases the probability of earning the intrinsic benefit of being correct. It is easy to show that each of the other strategies may be optimal, depending on parameters. In order for $e_1 = H$ to be optimal, it must be that $min\{U(H, L), U(H, H)\} \geq U(L, L)$, which holds if and only if the inequality in the proposition holds. □

10.2 Main results not driven by gaming or memorizing answers

Figures 9 and 10 replicate Figures 1 and 2 after dropping subjects who self-report gaming at least once, as described in Section 7.1. Figures 11 and 12 replicate Figures 1 and 2 after dropping subjects who mention that memorizing answers was part of their strategy to increase recall performance, as described in Section 7.2. In all cases, the results are substantially unchanged.

10.3 Bonus method increases thankfulness and reciprocity

In the Debrief stage, we ask all subjects how thankful they are for the payment they will receive, on a scale of 1 to 6. As shown in Figure 13, subjects in the Bonus (B) treatment are significantly more thankful relative to all other treatments. We also ask subjects if they are willing to answer three additional questions, which we implement if they answer

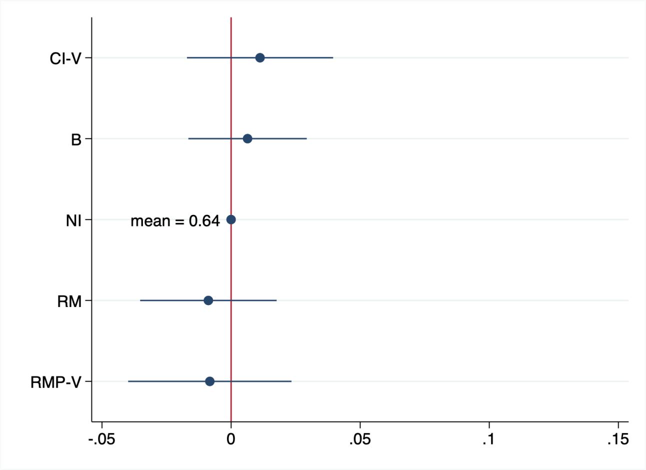


Figure 9: *Correctness after dropping subjects who self-report gaming.*

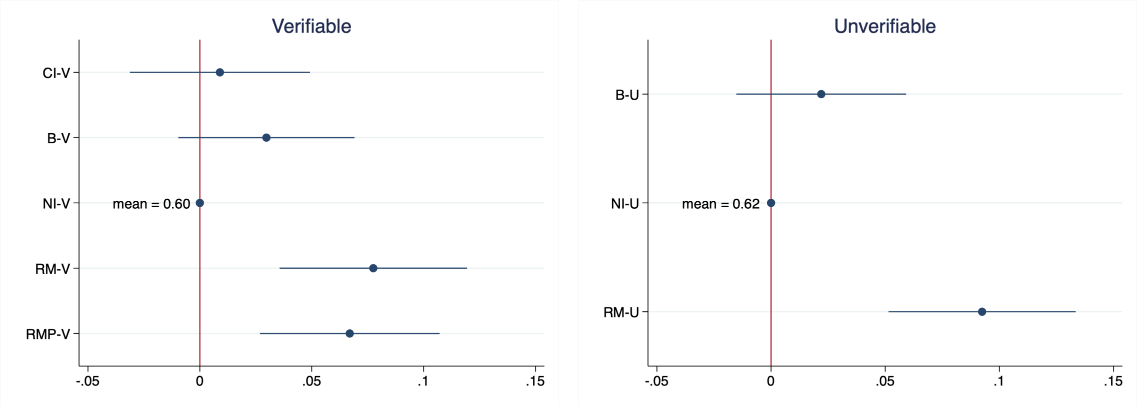


Figure 10: *Consistency after dropping subjects who self-report gaming.*

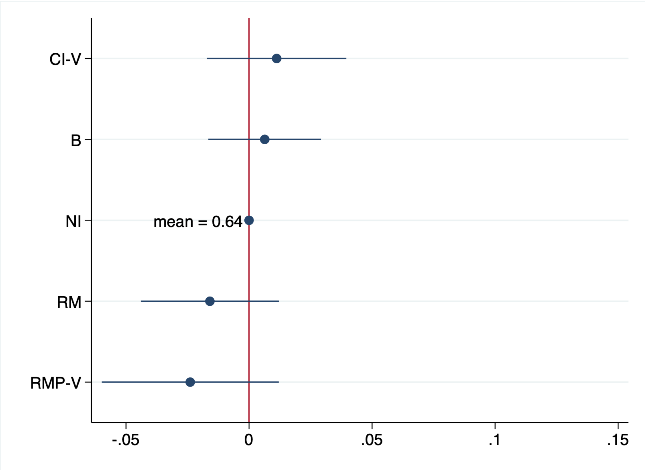


Figure 11: *Correctness after dropping subjects who mention memorizing answers.*

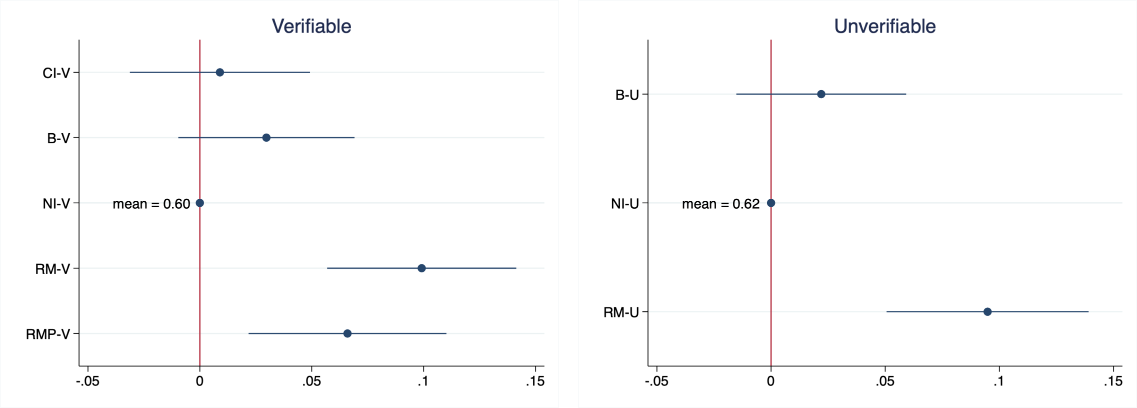


Figure 12: *Consistency after dropping subjects who mention memorizing answers.*

"yes," which we take as a measure of reciprocity toward the researchers. We find that subjects in B and NI are more likely to reciprocate than in the other treatments, as shown in Figure 14. We conjecture that the elevated willingness to reciprocate in NI is due to the fact the subjects, who were significantly quicker throughout the survey as shown in Section 8.1, are less fatigued.

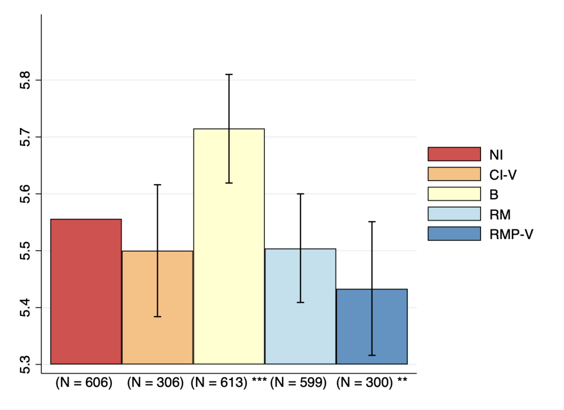


Figure 13: *Bonus method increases thankfulness for payment.*

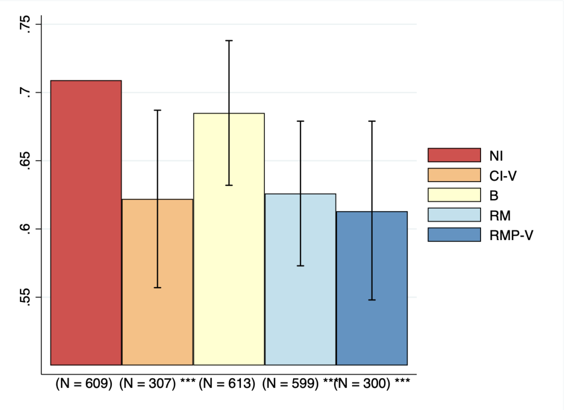
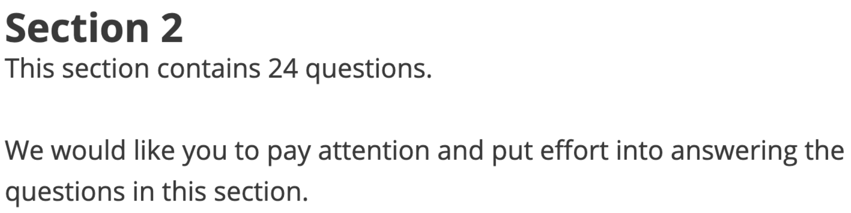


Figure 14: *Bonus method increases willingness to answer additional questions.*

10.4 Main stage instructions

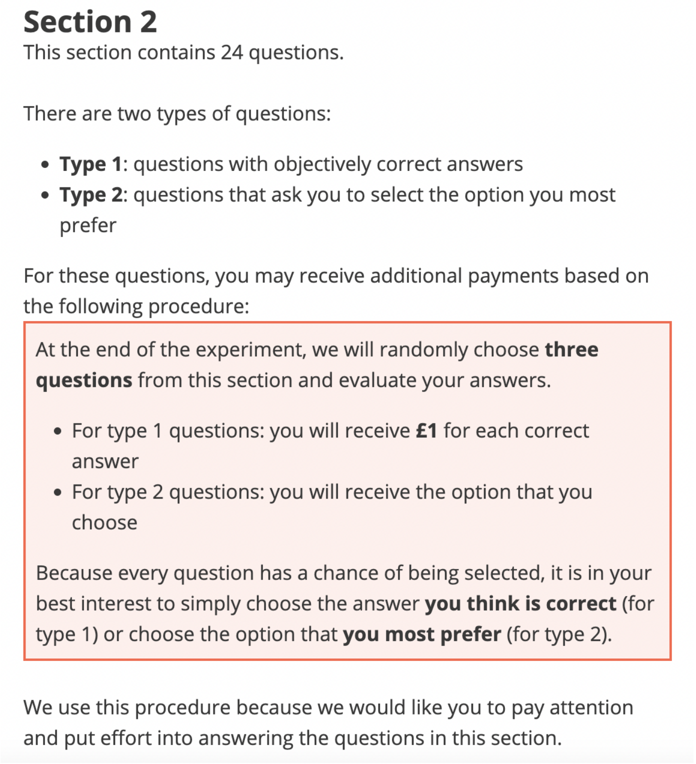


Section 2

This section contains 24 questions.

We would like you to pay attention and put effort into answering the questions in this section.

Figure 15: *NI instructions.*



Section 2

This section contains 24 questions.

There are two types of questions:

- **Type 1**: questions with objectively correct answers
- **Type 2**: questions that ask you to select the option you most prefer

For these questions, you may receive additional payments based on the following procedure:

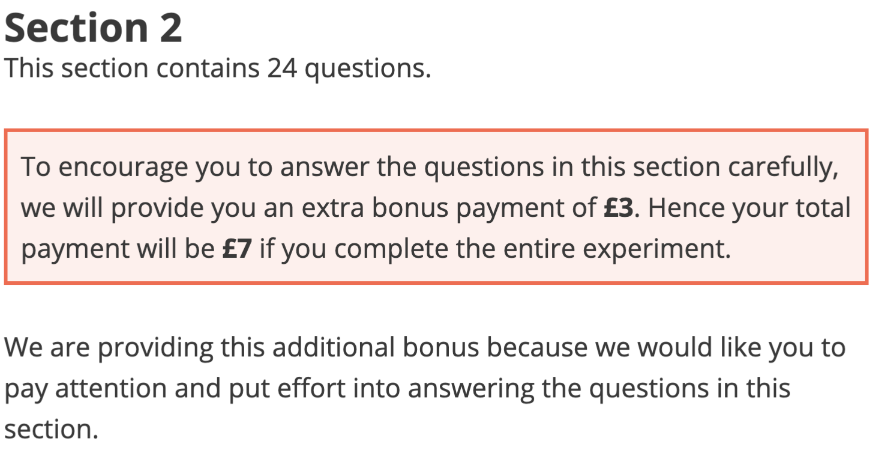
At the end of the experiment, we will randomly choose **three questions** from this section and evaluate your answers.

- For type 1 questions: you will receive **£1** for each correct answer
- For type 2 questions: you will receive the option that you choose

Because every question has a chance of being selected, it is in your best interest to simply choose the answer **you think is correct** (for type 1) or choose the option that **you most prefer** (for type 2).

We use this procedure because we would like you to pay attention and put effort into answering the questions in this section.

Figure 16: *CI-V instructions.*



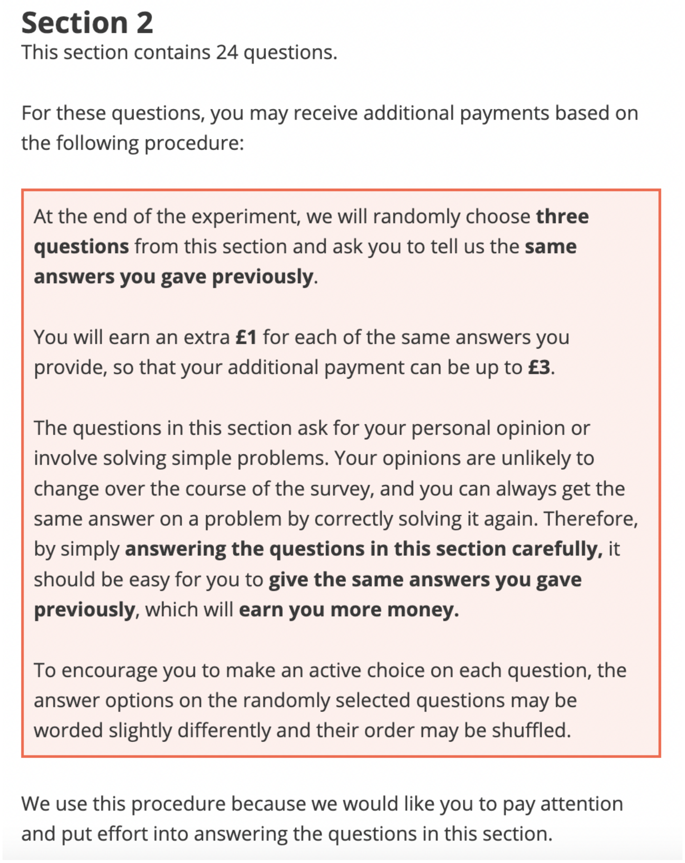
Section 2

This section contains 24 questions.

To encourage you to answer the questions in this section carefully, we will provide you an extra bonus payment of **£3**. Hence your total payment will be **£7** if you complete the entire experiment.

We are providing this additional bonus because we would like you to pay attention and put effort into answering the questions in this section.

Figure 17: *B instructions.*



Section 2

This section contains 24 questions.

For these questions, you may receive additional payments based on the following procedure:

At the end of the experiment, we will randomly choose **three questions** from this section and ask you to tell us the **same answers you gave previously**.

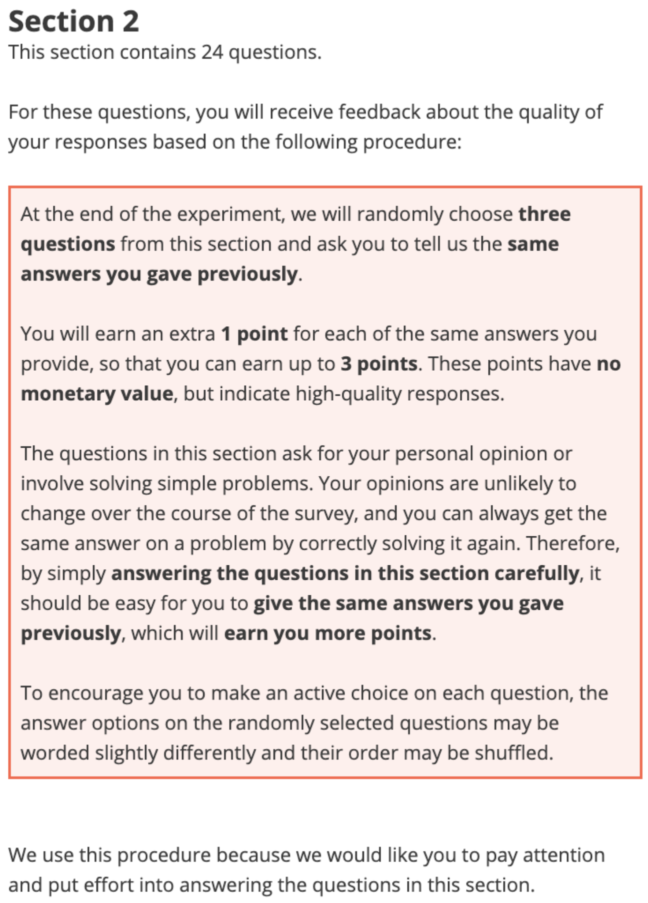
You will earn an extra **£1** for each of the same answers you provide, so that your additional payment can be up to **£3**.

The questions in this section ask for your personal opinion or involve solving simple problems. Your opinions are unlikely to change over the course of the survey, and you can always get the same answer on a problem by correctly solving it again. Therefore, by simply **answering the questions in this section carefully,** it should be easy for you to **give the same answers you gave previously**, which will **earn you more money.**

To encourage you to make an active choice on each question, the answer options on the randomly selected questions may be worded slightly differently and their order may be shuffled.

We use this procedure because we would like you to pay attention and put effort into answering the questions in this section.

Figure 18: *RM instructions.*



Section 2

This section contains 24 questions.

For these questions, you will receive feedback about the quality of your responses based on the following procedure:

> At the end of the experiment, we will randomly choose **three questions** from this section and ask you to tell us the **same answers you gave previously**.
>
> You will earn an extra **1 point** for each of the same answers you provide, so that you can earn up to **3 points**. These points have **no monetary value**, but indicate high-quality responses.
>
> The questions in this section ask for your personal opinion or involve solving simple problems. Your opinions are unlikely to change over the course of the survey, and you can always get the same answer on a problem by correctly solving it again. Therefore, by simply **answering the questions in this section carefully**, it should be easy for you to **give the same answers you gave previously**, which will **earn you more points**.
>
> To encourage you to make an active choice on each question, the answer options on the randomly selected questions may be worded slightly differently and their order may be shuffled.

We use this procedure because we would like you to pay attention and put effort into answering the questions in this section.

Figure 19: *RMP-V instructions.*

References

Aurelien Baillon. Bayesian markets to elicit private information. *Proceedings of the National Academy of Sciences*, 2017. 1

Aurelien Baillon, Han Bleichrodt, and George Granic. Incentives in surveys. *Journal of Economic Psychology*, 2022. 1

J. Beattie and G. Loomes. The impact of incentives upon risky choice. *Journal of Risk and Uncertainty*, 1997. 1

A. S. Benjamin, R. A. Bjork, and B. L. Schwartz. The mismeasure of memory: When retrieval fluency is misleading as a metamnemonic index. *Journal of Experimental Psychology. General*, 1998. 1

Daniel Benjamin. Errors in probabilistic reasoning and judgment biases. *Handbook of Behavioral Economics: Applications and Foundations 1*, 2019. 1

Miri Besken. Generating lies produces lower memory predictions and higher memory performance than telling the truth: Evidence for a metacognitive illusion. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, 2018. 1

J.M. Bland and D.G. Altman. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*, 1986. 2, 5

C. F. Camerer and R. M. Hogarth. The effects of financial incentives in experiments: A review and a capital-labor-production framework. *Journal of Risk and Uncertainty*, 1999. 1

Paul Curran. Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology*, 2016. 1

J. Cvitanic, D. Prelec, B. Riley, and B. Tereick. Honesty via choice matching. *American Economic Review: Insights*, 2019. 1

David Danz, Lise Vesterlund, and Alistair Wilson. Belief elicitation and behavioral incentive compatibility. *American Economic Review*, 2022. 1

Ambuj Dewan and Nathaniel Neligh. Estimating information cost functions in models of rational inattention. *Journal of Economic Theory*, 2020. 8

T. Dohmen, A. Falk, D. Huffman, U. Sunde, J. Schupp, and G. Wagner. Individual risk attitudes: Measurement, determinants, and behavioral consequences. *Journal of the European Economic Association*, 2011. 1

Benjamin Enke, Uri Gneezy, Brian Hall, David Martin, Vadim Nelidov, Theo Offerman, and Jeroen van de Ven. Cognitive biases: Mistakes or missing stakes? *Review of Economics and Statistics*, 2023. 1, 7, 9

A. Falk, A. Becker, T. Dohmen, B. Enke, D. Huffman, and U. Sunde. Global evidence on economic preferences. *The Quarterly Journal of Economics*, 2019. 1

Ernst Fehr and Simon Gachter. Fairness and retaliation: The economics of reciprocity. *Journal of Economic Perspectives*, 2000. 1

Shane Frederick. Cognitive reflection and decision making. *Journal of Economic Perspectives*, 2005. 4

Ben Gillen, Erik Snowberg, and Leeat Yariv. Experimenting with measurement error: Techniques with applications to the caltech cohort study. *Journal of Political Economy*, 2019. 5

R. Hertwig and A. Ortmann. Experimental practices in economics: A methodological challenge for psychologists? *Behavioral and Brain Sciences*, 2001. 1

Nicolas Jacquement, Stephane Luchini, Jason Shogren, and Verity Watson. Discrete choice under oaths. *Working paper*, 2019. 1

L. K. John, G. Loewenstein, and D. Prelec. Measuring the prevalence of questionable research practices with incentives for truth-telling. *Psychological Science*, 2012. 1

N. Miller, P. Resnick, and R. Zeckhauser. Eliciting informative feedback: the peer prediction method. *Management Science*, 2005. 1

Eyal Peer, David Rothschild, Andrew Gordon, Zak Evernden, and Ekaterina Damer. Data quality of platforms and panels for online behavioral research. *Behavior Research Methods*, 2022. 4

Drazen Prelec. A bayesian truth serum for subjective data. *Science*, 2004. 1

Goran Radanovic and Boi Faltings. A robust bayesian truth serum for binary signals. *Proceedings of the 27th AAAI Conference on Artificial Intelligence*, 2013. 1

K.H. Schlag, J. Tremewan, and J.J. Van Der Weele. A penny for your thoughts: A survey of methods for eliciting beliefs. *Experimental Economics*, 2015. 1, 3

Andrew Schotter and Isabel Trevino. Belief elicitation in the laboratory. *Annual Review of Economics*, 2014. 1, 3

Kathleen Vieira and Sean Lane. How you lie affects what you remember. *PsycEXTRA Dataset*, 2012. 1

R. Weaver and D. Prelec. Creating truth-telling incentives with the bayesian truth serum. *Journal of Marketing Research*, 2013. 1

Jens Witkowski and David Parkes. A robust bayesian truth serum for small populations. *Proceedings of the 26th AAAI Conference on Artificial Intelligence*, 2012. 1

Florian Zimmermann. The dynamics of motivated beliefs. *American Economic Review*, 2020. 1