Motivated Beliefs, Not Motivated Representations

Daniele Mauriello

Job Market Paper

November 10, 2026

Click here for the latest version

Abstract

This paper studies whether motivated reasoning operates through the choice of mental representation: do people adopt the representation of a problem that licenses more favourable beliefs? In the proposed framework, individuals choose among representations, trading off the motivated value of the beliefs each one licenses against its defensibility. When a heuristic provides a defensible alternative to the rational representation, the framework predicts that individuals adopt the heuristic or the rational representation depending on which one licenses the favoured conclusion. These predictions are tested across three datasets: data from Exley and Kessler (2024) and two pre-registered online experiments ($N_1 = 1{,}212$, $N_2 = 762$). In none of the three datasets does the motive move individuals towards the representation that licenses the most favourable conclusion, and in neither experiment does the joint effect of the motive and the heuristic robustly depart from the sum of their individual effects. Heuristics thus do not provide a convenient justification for holding favourable beliefs, and motives are no cure for heuristic reasoning.

JEL Codes: D83, D91, C90.

Keywords: Motivated reasoning, mental models, belief formation, heuristics, motivated errors, correlation neglect, selection neglect, online experiments.

Contact: Daniele Mauriello, University of Bonn, daniele.mauriello@uni-bonn.de. **Acknowledgements:** I thank Florian Zimmermann, Sebastian Kube, and Zwetelina Iliewa for their invaluable guidance and support. I thank Botond Kőszegi, the members of the Bonn behavioral lab meetings, and seminar participants for helpful comments and discussions. **Funding:** Support from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) through CRC TR 224 (Project A01) is gratefully acknowledged. I gratefully acknowledge funding from IZA and the Bonn Graduate School of Economics (BGSE). **Ethics approval:** Both experiments obtained ethics approval from the German Association for Experimental Economic Research (#ch8rfrGM and #xPgMJzYW). **Preregistration:** Both experiments were preregistered at AsPredicted (#284113 and #320081).

1 Introduction

People like to see themselves, and the things they care about, in a favourable light. Supporters of a political party, for instance, would rather believe that their party is successful and enjoys broad public support than that it is unpopular. Preferences of this kind reflect the fact that beliefs carry affective value: individuals derive utility from holding beliefs that are consistent with what they wish to be true (Köszegi, 2006; Bénabou and Tirole, 2016). Yet individuals cannot simply believe whatever they like. The available evidence limits which beliefs they can plausibly hold (Caplin and Leahy, 2001; Bracha and Brown, 2012): even the most devoted supporters cannot seriously believe that their party will win every vote in the next election.

How tightly the evidence constrains beliefs, however, depends on how conclusive it is. Evidence is inconclusive when it is compatible with a range of beliefs rather than pinning down a single one, and this can happen for different reasons. On the one hand, the evidence may be noisy, leaving the underlying state imprecisely determined. For example, evidence on a party's support might come from a poll which only provides a noisy estimate of the party's actual support. On the other hand, the evidence may license different conclusions depending on how individuals represent the process that generated it (Musolff and Zimmermann, 2025). For instance, evidence on a party's support might come from the social media news feed, which can be either read as representative of public opinion or as curated to match the reader's own views.

These two sources of inconclusiveness open different opportunities for motivated reasoning. Noisy evidence leaves room to form favourable beliefs *within* a given mental representation. Evidence that is compatible with competing representations may, in addition, leave room to choose *which* representation to use in the first place. The literature on motivated reasoning has documented the first channel: noisy evidence facilitates motivated belief formation by widening the range of beliefs a given representation licences (Eil and Rao, 2011; Möbius et al., 2022; Coutts, 2019; Dana et al., 2007; Haisley and Weber, 2010). Much less is known about the second. Can the motive to hold favourable beliefs shape which mental representation individuals adopt? In the example above, does the desire to see their party as popular affect whether supporters read their news feed as representative of public opinion or as curated to match their own views?

This paper addresses this question theoretically and empirically. I first develop a framework that formalises how motives can affect the choice of mental representation, and then test its predictions using new and existing experimental data.

Building on classical models of motivated reasoning, in which individuals distort beliefs *within* a mental representation, my framework also lets them choose *among* representations, in the spirit of Bolte and Fan (2024), trading off the motivated value of the beliefs each one licenses against its defensibility. Motivated reasoning can therefore produce two kinds of distortion. Besides the continuous distortion within a representation that standard models already capture, it can produce

a discrete one, by changing which representation individuals adopt. When an individual finds several representations defensible, the one they end up adopting depends on which conclusion the motive favours.

Heuristics offer a natural setting in which to empirically study this discrete channel, because many of them amount to a simplified representation of the problem at hand aimed at reducing cognitive costs (Gennaioli and Shleifer, 2009; Bordalo et al., 2023, 2026). In the example above, treating the news feed as representative provides a simple (albeit biased) way to infer public opinion, whereas accounting for selection and correlation requires a more demanding assessment of how the information was generated (Levy and Razin, 2015; Enke and Zimmermann, 2019; Enke, 2020). In environments that invite such a heuristic, individuals have at least two defensible representations available: the rational one and the heuristic. The framework then yields two predictions, depending on which conclusion the motive favours. When the heuristic delivers the desired conclusion, it provides a *convenient justification* individuals need to defend the beliefs they wish to hold: a "cognitive wiggle room". When it delivers an undesired conclusion, the motive can instead *debias* individuals, for instance by prompting them to think harder after reaching an unwelcome answer. Studying how motives shape the choice of representation thus also speaks to a question recently raised by Exley and Kessler (2024): how do motivated reasoning and cognitive biases interact?

I start from the evidence in Exley and Kessler (2024), who document *motivated errors*: mistakes that individuals make when doing so lets them justify selfish behaviour. Additionally, to examine whether such errors persist in environments conducive to cognitive biases, they run two online experiments with three main conditions: (i) *Heuristic only*, in which participants undergo some estimation tasks and face an environment conducive to correlation neglect or to anchoring;[1] (ii) *Aligned*, in which participants additionally have a motive expected to distort their beliefs in the same direction as the cognitive bias; and (iii) *Opposed*, in which the motive is expected to push them in the opposite direction as the cognitive bias.[2]

The authors find that errors are larger than in the *Heuristic only* condition when the motive and the cognitive bias are aligned, and smaller when they are opposed. A change in average errors, however, does not reveal which representation individuals adopt. The same pattern would arise if individuals adopted the heuristic representation and layered an unrelated motivated distortion on top of it, or if the bias and the motive simply affected different people.

I therefore re-analyse the correlation-neglect experiment of Exley and Kessler (2024), in which it is possible to distinguish between two representations: a *rational* one, which accounts for the correlation between signals, and a *naive* one, which treats them as independent. Following Enke and

[1] Correlation neglect is the tendency to overlook the double-counting problem that arises when updating beliefs from correlated signals (Enke and Zimmermann, 2019). Anchoring is the tendency of estimates to be pulled towards an initial, possibly irrelevant, reference value (Tversky and Kahneman, 1974).

[2] Exley and Kessler (2024) call these conditions *Cognitive bias only*, *Exacerbate bias* and *Mitigate bias*. I use the labels of my own experiments throughout.

Zimmermann (2019), I classify respondents as rational, naive, or making other errors, depending on whether their estimates match the answer implied by the rational representation, the one implied by the naive representation, or neither, and estimate a multinomial logit of type membership on treatment. If the motive shifts individuals between the rational and the naive representation, respondents should be more likely to be classified as naive in *Aligned*, and as rational in *Opposed*, than in the *Heuristic only* condition.

The data support neither prediction. In *Aligned*, respondents are less likely to be rational, but not more likely to be naive; instead, they make other errors and bunch at one extreme of the response range. In *Opposed*, respondents are less likely to be naive, but not more likely to be rational; again, they shift towards other errors at the extremes.

Absolute errors, which measure the size of errors regardless of their direction, tell the same story: if the motive debiased individuals, absolute errors should be smaller in *Opposed* than in the *Heuristic only* condition, but they are larger.

These results suggest that the motive does not move individuals between the rational and the naive representation. They do not, however, establish whether the motive interacts with the heuristic at all. Because the design of Exley and Kessler (2024) has no motive-only condition, the size of the motivated distortion absent the heuristic is unknown. One can therefore not tell whether the observed exacerbation and mitigation reflect motivated reasoning alone, or an interaction with the heuristic that operates through a channel other than switching between the rational and the naive representation.

My experiments close this gap with a full 2×2 factorial design that crosses whether the problem admits a heuristic shortcut with the presence of a directional motive. Figure 1 summarises the resulting conditions: *Control*, *Heuristic only*, *Motive only*, and *Motive*×*Heuristic*. Since the theory makes different predictions depending on whether the motive favours the naive or the rational conclusion, the *Motive*×*Heuristic* cell is split into two conditions, *Aligned* and *Opposed*.

I implement this design in two pre-registered online experiments on Prolific.[3] The first (n = 1,212) replicates the correlation-neglect (CN) task and the motivated-error (ME) setting of Exley and Kessler (2024); I refer to it as *CN*×*ME*. The second ($n = 762$) uses selection neglect (SN) as the heuristic and a motive tied to participants' own score on an IQ test, which keeps the conditions closer in complexity; I refer to it as *SN*×*IQ*.[4]

In both experiments, the factorial design lets me test whether the motive distorts beliefs more when the heuristic is available. The benchmark is additivity: if the motive simply operated on top of the heuristic, errors in *Aligned* would equal the sum of the errors in *Motive only* and *Heuristic only*, and errors in *Opposed* would equal their difference. Errors larger than the sum in *Aligned*, or smaller than the difference in *Opposed*, would instead indicate that the option to rely on a heuristic

[3] Prolific is a survey platform widely used in social-science research (Peer et al., 2022).

[4] Selection neglect is the tendency to treat a selected sample of information as if it were representative, overlooking the process that determines which observations are seen (Enke, 2020).

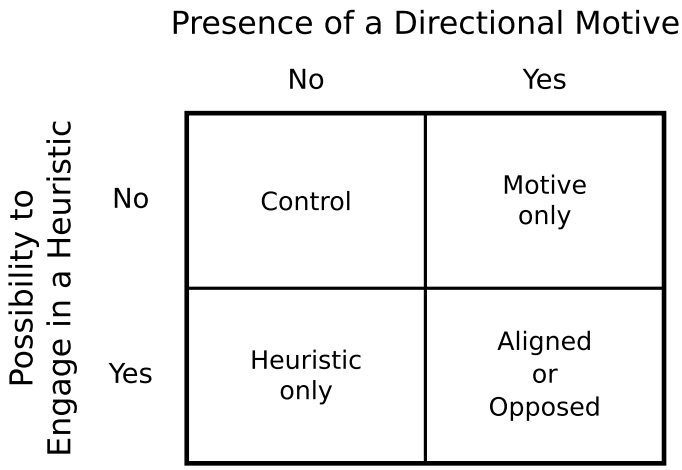


Figure 1: The 2×2 factorial design.

makes individuals more susceptible to motivated reasoning. Such an interaction could arise through either channel of the framework—individuals switching to a representation the motive favours, or distorting beliefs more within a given representation. I separate the two by classifying respondents as rational, naive, or making other errors.

The two experiments paint a consistent picture. I find no robust evidence that the option to rely on a heuristic makes individuals more susceptible to motivated reasoning, and no evidence that the motive moves individuals between the rational and the naive representation.

Consider first the additivity benchmark. When the motive and the heuristic point in the same direction, errors in *Aligned* do not robustly exceed the sum of the two individual effects. In *CN*×*ME*, errors in *Aligned* exceed those in *Heuristic only*, replicating Exley and Kessler (2024), but the excess over the additive benchmark is not significant in the full sample ($p = 0.17$).[5] In *SN*×*IQ*, the interaction is close to zero. When the motive and the heuristic point in opposite directions, neither experiment shows errors in *Opposed* that are smaller than the additive benchmark. In both directions, then, I find no robust evidence that the errors due to the motive change when the environment is conducive to the heuristic.

The classification of respondents points to the same conclusion. In neither experiment are respondents in *Aligned* more likely to be classified as naive than those in *Heuristic only*, and in neither are respondents in *Opposed* more likely to be classified as rational—the same pattern I find

[5]It becomes significant only in a sample that excludes participants who may not have been exposed to the motive. These are participants who prefer even the smallest bonus to the charity donation. Because the calibration cannot make them indifferent between the two payments, they face not just a motive but a proper, albeit small, incentive to err in the direction that secures the personal payment.

in the data of Exley and Kessler (2024). Across three datasets, then, the motive does not shift individuals towards whichever of the rational and the naive representation it favours.

Taken together, these results suggest that motives do not operate by selecting among mental representations. The heuristic does not serve as a convenient justification for desired conclusions, and an inconvenient conclusion reached via a heuristic does not prompt individuals to adopt the rational representation. In the language of the framework, motivated reasoning operates within a representation, as standard models assume, but does not tip the choice between them. This is what the framework predicts when individuals do not find the two representations comparably defensible: if most of them either see the feature the heuristic ignores or do not, the motive has no close call to tip.

One reading of this pattern is that it reflects how representations are adopted. Following Bordalo et al. (2026), my framework ties a representation's defensibility to the salience of the features it directs attention to. If salience alone settles the representation as individuals first read the problem, the representation is less chosen than *seen*, and by the time the motive comes into play it can act only on the belief that reading delivers. A motivated choice between representations would, moreover, require individuals to know which conclusion each representation leads to, and that knowledge is self-defeating: whoever has worked out the rational answer has noticed what the heuristic ignores, and can no longer adopt it in good faith. Motives should then shape which representation individuals adopt only when competing representations are explicitly put to them and each stays defensible in light of the other, as with the competing narratives of Eliaz and Spiegler (2020).

The implication is a simplifying one. Motivated reasoning and heuristic reasoning largely add up, so the two can be studied, and corrected, separately. It also follows that one cannot be enlisted to correct the other: a motive that pushes against the heuristic does not lead individuals to the rational representation.

Related literature

This paper builds on and contributes to several strands of literature. First, it contributes to the literature on motivated reasoning (see Bénabou and Tirole, 2016, for a review), which has proposed and documented several mechanisms through which directional motives can distort beliefs. Strategies to uphold favourable beliefs include selective exposure to information (Oster et al., 2013; Ganguly and Tasoff, 2017), selective scrutiny of information (Sanitioso and Kunda, 1991; Ditto and Lopez, 1992; Taber and Lodge, 2006), selective memory (Amelio and Zimmermann, 2023; Zimmermann, 2020), and selective aggregation of information (Eil and Rao, 2011; Möbius et al., 2022; Mayraz, 2011; Thaler, 2021).

Papers belonging in this last category show that individuals exploit the uncertainty in noisy information structures to justify favourable beliefs. Eil and Rao (2011) give participants noisy signals about their rank in intelligence and physical attractiveness, and find that they update close to Bayes' rule after good news but respond less, and less predictably, to bad news. Möbius et al.

(2022) find a similar asymmetry, together with conservatism, when participants learn about their rank on an IQ quiz. Mayraz (2011) shows that forecasts are biased towards the outcome that pays the forecaster, and that the bias shrinks when the data leave less room for subjective uncertainty.

While the papers mentioned above focus on how motives distort beliefs within a given mental representation, this paper joins a limited strand of literature that studies how motives can shape which mental representation individuals adopt and the interaction between motivated reasoning and cognitive biases. Ginossar and Trope (1987) pioneered this strand by showing that individuals asked to role-play as a defence lawyer are more likely to attend to base-rate information in the classic taxi-cab problem (Tversky et al., 1982) than individuals with a pure accuracy motive — that is, a directional motive can pull people towards the normatively correct answer when it happens to align with it. Kahan et al. (2017) provide related evidence: when interpreting a contingency table, numerate participants are more likely to compare the relevant ratios, rather than raw counts, when doing so yields a conclusion that agrees with their political views. Similarly, Hagenbach and Saucet (2023) find that receivers in disclosure games are less skeptical of strategically selected information when skepticism would lead to an unwelcome conclusion, a pattern they show can arise either from distorting beliefs after a skeptical inference or from stopping the inference short. More recently, Exley and Kessler (2024) and Bolte and Fan (2024) showed that directional motive can either exacerbate or mitigate errors driven by correlation neglect and anchoring, depending on whether the motive and the cognitive bias are aligned.

The studies described above, however, can only show the robustness of motivated reasoning to settings conducive to heuristics. Their designs and analyses do not allow the identification of a possible interaction effect. The pattern of exacerbation and mitigation they document is equally consistent with individuals experiencing a cognitive bias and layering a separate, unrelated motivated distortion on top of it, or with the two biases simply affecting different people.[6] I contribute to this literature by providing a design which is explicitly built to directly assess the difference between the independent and joint effects of the two types of bias on individuals.

Second, this paper bridges the literature on motivated reasoning and on mental models. The literature on mental models is based on the idea that individuals do not form beliefs from a unified reading of how the world works, but rather hold subjective mental models of the world which often depart from how it actually works (Esponda and Pouzo, 2016; Bohren and Hauser, 2025; Spiegler, 2020; Bordalo et al., 2026; Gagnon-Bartsch et al., 2018). Formalisations of this idea differ in where they locate the source of the mental models. It could be, for example, a misspecified causal model of how the relevant variables relate (see, e.g., Ambuehl et al., 2026; Spiegler, 2020, 2016; Schwartzstein

[6]Additionally, in the Exley and Kessler (2024) design, assignment to *Aligned* or *Opposed* is varied within individuals. The same participant is exposed to a motive pointing in the same direction as the bias in some rounds and in the opposite direction in others. Taking the observed pattern of exacerbation and mitigation as evidence of a genuine biasing or debiasing mechanism would require the same individual to be "biased" in a round, and to shake it off in the next one, only to fall for it again afterwards. My between-subject design avoids this by holding each individual's alignment fixed throughout.

and Sunderam, 2021), or attention to only some of the problem's features (see, e.g., Bordalo et al., 2026; Gagnon-Bartsch et al., 2018). In this paper, in the spirit of Kunda (1990) I contribute to this strand of literature by introducing a top-down mechanism for how individuals select which mental model to use: motivated reasoning. In this regard, my work is closest in spirit to Eliaz and Spiegler (2020), who also model individuals who "hedonically" select which narrative to adopt. In their model, individuals' choice over different narratives is constrained by objective long-run data. In the spirit of Bordalo et al. (2026), I deviate from Eliaz and Spiegler (2020) by detaching the defensibility of a mental model from objective data: in my framework, an individual can adopt a defensible mental model even if there is no data to support it. Because Eliaz and Spiegler (2020)'s narratives must fit long-run frequency data, their framework is inherently backward-looking: it speaks to recurring problems on which a track record has accumulated. Detaching defensibility from data lets my model apply equally to new situations, where no such record yet exists but a range of mental models can still be credible.

The remainder of the paper proceeds as follows. Section 2 lays out the theoretical model and derives testable predictions. Section 3 re-analyses the correlation-neglect experiment of Exley and Kessler (2024) in light of these predictions. Section 4 sets out the design common to the two new experiments, and Sections 5 and 6 present each experiment and its results. Section 7 discusses what the results imply for the model and concludes.

2 A Two-Stage Model of Motivated Reasoning

This section formalises the theoretical framework sketched in the introduction. Faced with a body of evidence, a decision maker (DM) forms a belief in two steps: they first select a defensible *mental representation*, and only then settle on a belief within it. Both steps are open to motivated reasoning. I model the first in the spirit of Bolte and Fan (2024), and the second in the spirit of Caplin and Leahy (2001). Across representations, the motive tips the first-stage choice toward whichever defensible reading of the evidence licenses the more congenial conclusion. What disciplines that choice is not the representation's fit to the truth (of which the DM may be unaware) but how defensible the representation is. Within a representation, beliefs are then pulled toward what the DM would like to be true, at a cost that grows with the distance from the rational belief the representation implies. The rest of this section builds these pieces in turn: it describes mental representations in Section 2.1, it sets up the mental representation choice problem and characterises its solution in Section 2.2, and then it presents the predictions the two stages jointly deliver in Section 2.3.

2.1 Mental representations

Let $\theta \in \Theta$ denote the underlying object the DM is forming beliefs on and has evidence about.[7] I define a mental representation $m \in \mathcal{M}$ as a mapping, $m : \mathcal{E} \to \Delta(\Theta)$, from the evidence, $e \in \mathcal{E}$, to a belief about θ, $g_{m,e}(\theta) \in \Delta(\Theta)$. Each m embodies a set of assumptions about the structure of the problem, and hence about how the evidence e is generated from θ.[8] $g_{m,e}(\theta) \in \Delta(\Theta)$ is the posterior obtained by updating a fixed prior on θ through the likelihood those assumptions imply. The belief $g_{m,e}$ is therefore Bayesian only *internally*, relative to the structure the representation assumes: it is the correct posterior given m, but coincides with the true Bayesian posterior only when m's assumptions are themselves correct.

For example, given four news reports about the performance of a political party ($e = \{\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3, \hat{\theta}_4\}$), one mental representation, m_{iid}, might assume the four reports are iid unbiased signals of θ. Updating under this assumption weights each report equally. Another mental representation, $m_{correlated}$, might assume the reports are all inflated by a common amount. Updating under this mental representation subtracts a constant from each report before aggregating. Each mental representation is thus a different assumed data-generating process, and g_m is the belief a Bayesian would hold if that process were the true one.

While several micro-foundations are possible, in this paper I adopt the one of Bordalo et al. (2026) and treat each mental representation as an *attention profile*. In their framework, a statistical problem is described by a vector of F features, and a decision maker forms beliefs by encoding only some of them. An attention profile is a binary vector $\alpha \in \{0,1\}^F$, with $\alpha_j = 1$ if feature j is attended to and $\alpha_j = 0$ if it is neglected. Given α, the DM forms beliefs about θ from the attended features alone. The resulting belief is internally coherent given α, but it coincides with the Bayesian posterior only when the attended features are the ones actually relevant to θ. Under this micro-foundation, a mental representation m can be thought of as an attention profile α_m, with g_m the belief it induces. In the running example, m_{iid} neglects the feature recording whether the four reports share a common source of inflation and so treats them as independent unbiased signals, whereas $m_{correlated}$ attends to that feature and nets it out before aggregating.

In what follows, the evidence e is held fixed throughout the DM's problem, so I suppress it from the notation and write g_m for $g_{m,e}$.

[7] θ can be any object the DM might form a belief about: a point estimate, a probability distribution, or richer structure still. I assume throughout that Θ is finite or countable; under this assumption, nothing below depends on what kind of object θ is. The proofs rely on a coordinate-wise first-order condition, and extending them to a continuum of states would require a calculus-of-variations argument that I do not pursue here.

[8] This formalisation runs in the opposite direction to the usual definition of mental models in the literature, where a mental model maps the state of the world to a distribution over outcomes, that is, a likelihood. Given a prior, the two are linked by Bayes' rule: m is the Bayesian inverse of the likelihood the representation assumes.

2.2 Mental representation choice

Having defined a mental representation m and the Bayesian belief g_m it implies, I now turn to the first stage of the decision problem: the choice of the mental representation. I formalise this by having the DM trade off the motivated value of the belief adopted given the mental representation, $V(m)$, against the defensibility of the mental representation, $D(m)$. In this stage, the DM chooses a mental representation m from the set of available mental representations $\mathcal{M}$ to maximize the sum of these two objects, weighted by a malleability parameter λ.

$$m^* \in \arg\max_{m \in \mathcal{M}} \; V(m) + \frac{1}{\lambda} D(m) \tag{1}$$

Adopting a less defensible mental representation is worthwhile only when the belief it licenses delivers enough additional motivated value to offset the defensibility conceded.

The remainder of this subsection unpacks the two objects that (1) trades off: the indirect motivated value of a mental representation, $V(m)$, and its defensibility, $D(m)$.

2.2.1 The indirect value of a mental representation: $V(m)$

In my framework, the DM does not gain utility directly from holding a mental representation, m, but rather from the belief $\gamma \in \Delta(\Theta)$ it licenses. While the rational belief implied by a mental representation, g_m, is fixed, in line with the motivated-reasoning literature, I allow the DM to still manipulate their belief *within* a mental representation.

Following Caplin and Leahy (2001), I assume that the DM gains utility from holding a belief γ under the mental representation m according to:

$$U(\gamma, m) = v(\gamma) - \frac{1}{\lambda} \mathrm{KL}(\gamma | g_m), \qquad \mathrm{KL}(\gamma | g_m) = \sum_{\theta} \gamma(\theta) \ln \frac{\gamma(\theta)}{g_m(\theta)} \tag{2}$$

In equation (2), $v(\gamma)$ represents the psychological payoff the DM gains from holding beliefs that assign higher probabilities to outcomes that are favourable to them. In a sense, $v(\gamma)$ ranks beliefs by how gratifying they are to hold (rather than by their accuracy). For instance, if θ is the performance of the DM's favourite political party, v is higher for beliefs γ that place more probability mass on higher values of θ: the DM enjoys thinking of their party as a successful one. The second term in equation (2) captures the mental cost for departing from the rational belief, g_m, implied by the mental representation m. This cost is given by the Kullback–Leibler divergence between γ and g_m,[9] scaled by a malleability parameter λ that governs how costly it is to believe something other than g_m.

Because the DM can still spin their belief within the model, the value of holding m is the utility

[9] An integral against θ in place of the sum when Θ is a continuum.

of that optimally spun belief, not the utility of g_m itself:

$$\gamma_m^* \in \arg\max_{\gamma \in \Delta(\Theta)} U(\gamma, m), \qquad V(m) = U\big(\gamma_m^*, m\big). \tag{3}$$

Evaluating $V(m)$ therefore requires solving the within-representation belief choice first, the *second stage* of the problem. I solve it in Appendix A under two approximations: the motive's payoff v is locally linear around g_m, with state-by-state marginal reward $\psi(\theta)$, and the divergence is replaced by its second-order (χ^2) expansion around g_m.[10] The solution is

$$\gamma_m^*(\theta) = g_m(\theta) + \lambda\, g_m(\theta)\Big(\psi(\theta) - \mathbb{E}_{g_m}[\psi]\Big) = g_m(\theta) + \lambda\, \mathrm{Cov}_{g_m}\big(\mathbb{1}_{\{\theta\}},\, \psi\big), \tag{4}$$

$$V(m) = v(g_m) + \frac{\lambda}{2}\, \mathrm{Var}_{g_m}(\psi). \tag{5}$$

Three takeaways follow. First, the value of a representation has two parts: the motive's assessment of its honest belief, $v(g_m)$, and a *spin bonus*. At the optimum, the DM keeps half of the gross gain from spinning, as the divergence cost absorbs the other half. Second, the spin bonus increases in $\mathrm{Var}_{g_m}(\psi)$, so two representations the motive rates equally on their honest merits can still differ in appeal: the one that spreads mass over states the motive values differently leaves more room for motivated spin. Motives therefore act on the first-stage choice not only through where a representation's belief sits, but also through how much freedom it leaves the DM to move it. Third, the DM shifts probability toward the states the motive rewards more than the average state under g_m, in proportion to $g_m(\theta)$ itself: a representation that puts little weight on a state leaves little room to inflate it, and one that is certain of a state leaves no room at all.

2.2.2 The defensibility of a mental representation: $D(m)$

The second object in (1), $D(m)$, measures how defensible a mental representation is as a reading of the evidence. This is the constraint that disciplines the choice of model, the first-stage counterpart of the role the reference belief g_m plays in disciplining second-stage spin. Engaging a mental representation the DM finds natural is almost free; talking oneself into a far-fetched one is costly, and (1) makes the DM bear that cost. A less defensible model is therefore adopted only when the belief it licenses delivers enough additional indirect motivated value $V(m)$ to offset the defensibility conceded.

Crucially, $D(m)$ is not disciplined by the mental representation's fit to the truth. The DM's choice of mental representation is disciplined by how defensible it is, not by how accurate it is. A mental representation can thus be both cheap to adopt and inaccurate, so biased reasoning survives in my framework even when the motive is absent or when there are no observed long-run frequencies

[10] The first is exact when v is linear in γ, as in Caplin and Leahy (2001); both are accurate whenever the optimal spin is small.

to learn from. Detaching defensibility from accuracy sets my framework apart from the narratives of Eliaz and Spiegler (2020), whose "credibility" is pinned down by consistency with observed long-run frequencies.

As for mental representations, several micro-foundations for D are possible. Consistent with Section 2.1, I adopt the one implied by Bordalo et al. (2026), which also connects the first stage to an existing account of why people read the same evidence differently. Recall that, under this micro-foundation, a mental representation can be thought of as an attention profile $\alpha \in \{0,1\}^F$ over the problem's features, with g_α the belief that profile induces. In their framework, bottom-up attention selects the profile that maximises total salience:

$$\alpha^* \in \arg\max_{\alpha} \; C(\alpha) + P(\alpha) + \epsilon_\alpha, \tag{6}$$

where $C(\alpha)$ is the *contrast* of the resulting representation — how sharply it favours one hypothesis over the other — $P(\alpha)$ is its *prominence* — how top-of-mind the attended features are, given how the problem is described — and ϵ_α is an individual-level salience shock.

I therefore measure the defensibility of a mental representation by the salience of its attention profile, $D(\alpha) := C(\alpha) + P(\alpha) + \epsilon_\alpha$. This turns the first stage into a *motivated choice of attention profile*: the DM picks α to maximise $V(\alpha) + \frac{1}{\lambda}\big(C(\alpha) + P(\alpha) + \epsilon_\alpha\big)$, with bottom-up salience setting the default mental representation and the motive acting as a top-down override. Absent a motive, the DM simply picks the most salient profile.

2.3 Predictions

The two stages together generate three testable predictions.

1. **Within-model, motivated distortion.** Holding the mental representation fixed, beliefs move away from the model-implied belief g_m toward the states the motive rewards (as described in equation (4)). This is the Caplin and Leahy (2001) channel, recovered as my second stage.

2. **Unmotivated model choice.** Absent a motive ($V(m) = 0, \forall m \in \mathcal{M}$), the DM chooses the most defensible mental representation (which might not be the rational one). This is the Bordalo et al. (2026) channel, recovered as my first stage absent a motive.

3. **Motivated choice of model.** When the DM finds two mental representations comparably defensible ($D(m) \sim D(m')$), the motive tips the first-stage choice (described in equation (1)) toward whichever model licenses the desired belief. Beliefs then switches discretely between the accurate and the biased model with the *direction* of the motive.

Prediction 3 applies directly to heuristics, many of which amount to a simplified mental representation of the problem at hand. Suppose the DM has two representations available: a rational one, m_r, and a heuristic one, m_h. In the running example, these are $m_{correlated}$ and m_{iid}. By (1), the DM adopts the heuristic if and only if

$$V(m_h) - V(m_r) \;>\; \frac{1}{\lambda}\big(D(m_r) - D(m_h)\big) \tag{7}$$

which implies that the share of DMs who adopt the heuristic is increasing in $V(m_h) - V(m_r)$.

- **Prediction 3a (Aligned).** When the heuristic licenses the more congenial belief, $V(m_h) > V(m_r)$, the motive *lowers* the share of DMs who adopt the rational representation and *raises* the share of DMs who adopt the heuristic: the heuristic provides the justification the DM needs to hold the belief they wish to hold, a "cognitive wiggle room".
- **Prediction 3b (Opposed).** When the rational representation licenses the more congenial belief, $V(m_r) > V(m_h)$, the motive *lowers* the share of DMs who adopt the heuristic and *raises* the share of DMs who adopt the rational representation: the motive can debiases the DM.

3 Motivated Errors Revisited

To test Predictions 3a and 3b, I start by re-analysing the data of Exley and Kessler (2024), who document *motivated errors*: mistakes that individuals make when doing so lets them justify selfish behaviour. They show that, in an environment conducive to correlation neglect, a directional motive enlarges errors when it pushes in the same direction as the heuristic and shrinks them when it pushes in the opposite direction. This is the pattern the two-stage framework predicts, but average errors cannot reveal how it arises: it is equally consistent with individuals switching between representations and with a motivated distortion within a representation. Their correlation-neglect task, however, has a feature that makes it possible to tell these apart: the rational and the naive representation each imply a distinct answer, so one can infer which representation a respondent's answers reflect.

In this section, I first describe their experiment. I then classify respondents by the representation their answers reflect, and test whether the motive shifts them between the rational and the naive one. Next, I turn to absolute errors, which speak to whether the motive debiases respondents. I close by taking stock of what these data can and cannot tell us.

3.1 Experimental design (Exley and Kessler (2024))

The data come from the correlation-neglect study of Exley and Kessler (2024). In July 2021, the authors recruited 1,200 participants on Amazon Mechanical Turk, who received $3.50 for completing the study plus a bonus from one randomly selected decision.

The study has two parts, one of which is randomly selected at the end to determine payment. The first is a calibration task. In a multiple price list, each participant chooses between a donation of 150 cents to a charity (the Make-A-Wish Foundation) and a bonus payment for themselves that rises from 0 to 150 cents in increments of 5 cents. The row at which they switch identifies an interval of bonus payments within which the participant is close to indifferent between the bonus and 150 cents for the charity, and X is set to the lower bound of this interval.[11] This amount is used in the second part to make bonus payments and charitable donations equally valuable to each participant.

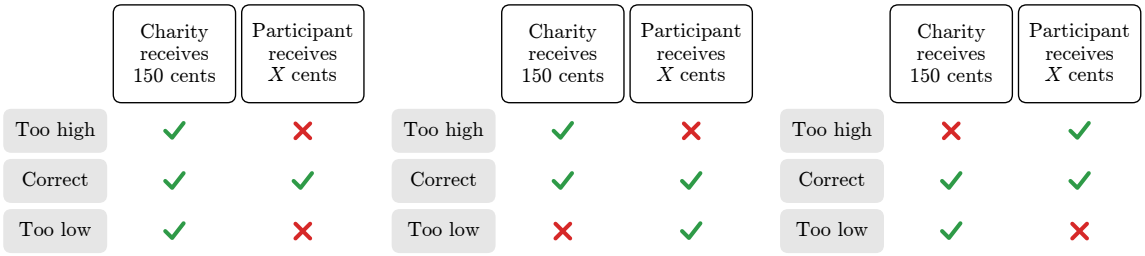
In the main task, participants estimate a number from information presented in a correlated way. Following Enke and Zimmermann (2019), the target is the average of four "estimates". Participants observe Estimate 1 directly, but learn about the other three only through three "news channels", each reporting the average of Estimate 1 and one of the other estimates. A respondent who accounts for this structure can back out all four estimates and average them. A respondent who neglects it and treats the four reports as independent signals instead overweights Estimate 1, so their answer is pulled towards it. Participants answer ten such questions on a 0–100 scale. The questions come in five pairs that share the same four estimates, and hence the same correct answer, but differ in whether Estimate 1 is the lowest estimate (*Low* questions), so that correlation neglect pulls the answer down, or the highest (*High* questions), so that it pulls the answer up. Table 7 in Appendix B reports the estimates and the signals respondents observe for each question.

Three between-subject treatments, shown in Figure 2, vary who benefits from an error in a given direction. In the *No motive* treatment,[12] the charity receives 150 cents regardless of the answer, and the participant additionally receives X cents if their answer is correct, which gives them an accuracy incentive. The *Underestimate* treatment replaces the charity's payment after an underestimate with a payment of X cents to the participant; the *Overestimate* treatment does the same for overestimates. Because X is calibrated to be worth as much to the participant as 150 cents for charity, the two treatments keep stakes constant while giving participants a self-serving reason to err in one direction.

The analyses below do not compare the three treatments directly. Instead, they compare three conditions, defined by whether a participant has a motive to err in a given direction and, if so,

[11] This means that, if anything, participants slightly prefer the donation to a bonus of X cents for themselves. If the lower bound is 0, X is set to 5 cents. These participants seem to dislike the donation, so the calibration cannot make them indifferent between the two payments.

[12] Exley and Kessler (2024) call this treatment *Control*; I rename it to avoid confusion with the *Control* condition of my experiments.



(a) *No motive* (b) *Underestimate* (c) *Overestimate*

Figure 2: Payments by treatment in Exley and Kessler (2024)

Notes: Each panel shows, for an answer that is too high, correct or too low, whether the charity receives 150 cents and whether the participant receives X cents. X is the participant's charity equivalent, elicited in the calibration task: the bonus that makes the participant indifferent between receiving it and having the charity receive 150 cents.

whether that direction coincides with the one in which correlation neglect pulls the answer. Participants in *No motive* answer all ten questions in the *Heuristic only* condition. Participants in the two motive treatments face different conditions, depending on the question. They answer in *Aligned* when the motive and correlation neglect push the same way: on Low questions in *Underestimate*, where they benefit from an underestimate and correlation neglect pulls the answer down, and on High questions in *Overestimate*, where they benefit from an overestimate and correlation neglect pulls the answer up. They answer in *Opposed* when the two push in opposite directions: on High questions in *Underestimate* and on Low questions in *Overestimate*. The comparison with *Heuristic only* is therefore between participants, whereas the comparison between *Aligned* and *Opposed* is within participants, each of whom answers five questions in each. To keep the terminology consistent with my experiments, I refer to these conditions as *Heuristic only*, *Aligned* and *Opposed*; Exley and Kessler (2024) call them *Cognitive bias only*, *Exacerbate bias* and *Mitigate bias*. The final sample comprises 12,000 answers (1,200 participants $\times$ 10 questions).

3.2 Rational or naive?

The experiment maps directly onto the two-representation case of Section 2.3. Participants receive correlated signals, which they can read through a *rational* representation, m_r, that accounts for the correlation, or through a *naive* one, m_h, that treats the signals as independent. The *Heuristic only* condition provides the no-motive benchmark. In *Aligned*, the motive favours the conclusion the naive representation licenses, $V(m_h) > V(m_r)$, as in Prediction 3a. In *Opposed*, it favours the conclusion the rational representation licenses, $V(m_r) > V(m_h)$, as in Prediction 3b. If the motive operates through the choice between the rational and the naive representation, respondents should therefore be more likely to adopt the naive representation in *Aligned*, and the rational one

in *Opposed*, than in the *Heuristic only* condition.

To measure whether an answer reflects the rational or the naive representation, I use the naiveté parameter of Enke and Zimmermann (2019). For an answer a to a question with correct answer c and naive answer a^N (the answer a respondent would give by treating the four reports as independent), the naiveté parameter is

$$\chi = \frac{a - c}{a^N - c}. \tag{8}$$

It equals 0 when the answer coincides with the rational one and 1 when it coincides with the naive one. Values between 0 and 1 indicate partial neglect of the correlation, negative values an error in the direction opposite to correlation neglect, and values above 1 an error that overshoots even the naive answer. Because χ divides the error by the signed gap between the naive and the correct answer, it measures every error in units of that gap and in the direction of correlation neglect. This makes it comparable across questions, including those in which correlation neglect pulls the answer in opposite directions.

Following Enke and Zimmermann (2019), I summarise each respondent's answers by the median of χ across the questions they answer in a given condition.[13] Taking the median makes the classification robust to an occasional outlying answer. Enke and Zimmermann (2019) show that, under correlated information, this median is bimodal, with peaks at the rational ($\chi = 0$) and the naive ($\chi = 1$) benchmark. Figure 3 shows the distribution of respondents' median χ in each condition. In the *Heuristic only* condition, it reproduces the pattern documented by Enke and Zimmermann (2019): one peak at the rational benchmark ($\chi = 0$) and one around the naive benchmark ($\chi = 1$). In both motive conditions, both peaks are lower and mass shifts towards the tails. In *Aligned*, it accumulates above the naive benchmark, with a spike around $\chi = 2$; in *Opposed*, it accumulates below the rational benchmark, with a spike around $\chi = -2$. Since answers at the ends of the 0–100 scale correspond to values of χ between roughly 2.1 and 2.4 in absolute value, these spikes reflect respondents who bunch at the end of the response range that the motive favours.

I classify a respondent as *rational* if the median χ lies within $[-0.33, 0.33]$, that is, close to the correct answer, as *naive* if it lies within $(0.33, 1.66)$, that is, close to the naive answer, and as making *other errors* otherwise. The last category collects errors that are inconsistent with both representations: errors in the direction opposite to correlation neglect, and errors that overshoot the naive answer.

In terms of these types, Predictions 3a and 3b are statements about shares. If the motive operates through the choice between the rational and the naive representation, the share of naive respondents should be higher in *Aligned* than in the *Heuristic only* condition (Prediction 3a), and the share of rational respondents should be higher in *Opposed* (Prediction 3b). I test both with a

[13] Each of the 408 *No motive* participants thus contributes one median over all ten answers, and each of the 792 motive-treatment participants one median per motive condition, each over five answers. Classifying each *No motive* participant separately on each half of their answers leaves the main conclusions unchanged (Table 8, Appendix C).

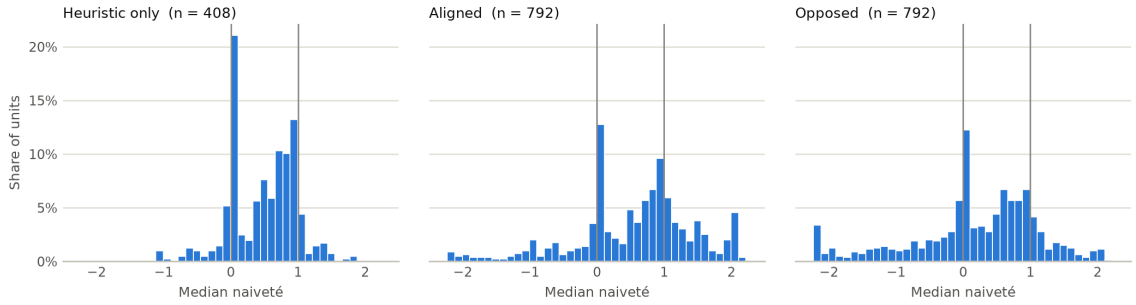


Figure 3: Distribution of respondents' median naiveté parameter by condition

Notes:$\chi = 0$ corresponds to the rational answer and $\chi = 1$ to the naive answer. Bars show the share of observations in each bin. Data from Exley and Kessler (2024).

multinomial logit of type membership on indicators for *Aligned* and *Opposed*, with the *Heuristic only* condition as the reference, and report average marginal effects in Table 1.

Table 1: Representation adopted by condition

	Rational	Naive	Other errors
Aligned	-0.125***	-0.141***	0.266***
	(0.027)	(0.034)	(0.037)
Opposed	-0.065**	-0.225***	0.290***
	(0.027)	(0.033)	(0.036)
Share in *Heuristic only*	0.328	0.618	0.054
Units		1,992	

Notes: Average marginal effects from a multinomial logit of type membership on indicators for *Aligned* and *Opposed*; the reference condition is *Heuristic only*. Types are based on the median naïveté parameter χ: rational if $\chi \in [-0.33, 0.33]$, naive if $\chi \in (0.33, 1.66)$, other errors otherwise. Standard errors in parentheses. Data from Exley and Kessler (2024). $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

The data support neither prediction. In the *Heuristic only* condition, 32.8% of respondents are classified as rational, 61.8% as naive, and 5.4% as making other errors. In *Aligned*, respondents are 12.5 percentage points less likely to be rational than in the *Heuristic only* condition ($p < 0.01$), but they are not more likely to be naive: the share of naive respondents falls by 14.1 percentage points ($p < 0.01$). The probability mass they lose goes to other errors, whose share rises by 26.6 percentage points ($p < 0.01$). In *Opposed*, the share of naive respondents falls by 22.5 percentage points ($p < 0.01$), but the share of rational respondents does not rise; it falls by 6.5 percentage points ($p < 0.05$). Again, the mass moves towards other errors, whose share rises by 29.0 percentage points ($p < 0.01$). In *Aligned*, the motive thus moves respondents away from the rational representation but not towards the naive one; in *Opposed*, it moves them away from the naive representation but

not towards the rational one. In both conditions, respondents leave the rational and the naive representation alike, for answers that neither of them implies.

The main conclusions are robust to the choice of dependent variable, classification cutoffs and sample. Appendix C repeats the analyses with the respondent-level mean of χ and with χ for each individual answer, with alternative cutoffs for the types, and without participants whose calibration failed.

3.3 Absolute errors

Absolute errors provide additional evidence on whether the motive debiases respondents. If the motive debiased respondents in *Opposed*, moving them from the naive towards the rational representation, their answers should end up closer to the correct one, and their absolute errors should fall relative to the *Heuristic only* condition.

For each answer, the absolute error is the distance between the answer and the correct one, $|a-c|$, regardless of its direction. I regress it on indicators for *Aligned* and *Opposed*, with the *Heuristic only* condition as the reference. The regression is estimated at the level of individual answers (12,000 observations), includes fixed effects for the five question pairs, and clusters standard errors by participant. Table 2 reports the results.

Table 2: Absolute errors by condition

	Absolute error
Heuristic only (constant)	15.628***
	(0.443)
Aligned	4.343***
	(0.627)
Opposed	2.174***
	(0.610)
Question-pair fixed effects	Yes
Observations	12,000
R^2	0.018

Notes: OLS of the absolute error, $|a-c|$, on indicators for *Aligned* and *Opposed*, with question-pair fixed effects; the constant is the predicted absolute error in the *Heuristic only* condition. One observation per participant and question. Standard errors clustered by participant in parentheses. Data from Exley and Kessler (2024). $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Absolute errors rise in both motive conditions. In the *Heuristic only* condition, answers are on average 15.6 points away from the correct one. This distance grows by 4.3 points in *Aligned* and by 2.2 points in *Opposed* (both $p < 0.01$). In *Opposed*, despite the motive points toward the correct answer, it does not bring answers closer to the truth. On the contrary, it increases absolute errors.

This is consistent with the classification in Section 3.2. Respondents in the motive conditions leave the naive and the rational answer alike, for other errors, and many bunch at the end of the response range that the motive favours (Figure 3). Answers at the end of the response range lie roughly twice as far from the correct one as the naive answer does, which raises absolute errors in both conditions.

3.4 Discussion

The two analyses point in the same direction. If the motive operated through the choice between the rational and the naive representation, it would move respondents from one to the other: towards the naive representation in *Aligned* and towards the rational one in *Opposed*. It does neither. In *Aligned*, respondents are less likely to be rational but not more likely to be naive; in *Opposed*, they are less likely to be naive but not more likely to be rational. In both conditions, they turn instead to answers that neither representation implies, and their answers end up further from the correct one. The exacerbation and mitigation of average errors documented by Exley and Kessler (2024) therefore do not reflect respondents switching between the rational and the naive representation.

These data leave one question open: whether the motive interacts with correlation neglect at all. The design of Exley and Kessler (2024) has no condition in which participants have a motive but face uncorrelated information, so the size of the motivated distortion in the absence of the heuristic is unknown. One can therefore not tell whether the patterns above reflect motivated reasoning that would arise regardless of the heuristic, or an interaction with the heuristic that operates through a channel other than the choice between the rational and the naive representation. Answering this question requires measuring the effect of the motive both with and without the heuristic, which is what the factorial design of the experiments described in the next section does.

4 Design

Section 3 left one question open: whether the motive interacts with the heuristic at all. Answering it requires a difference-in-differences: the effect of the motive when the heuristic is available, minus its effect when it is not. Identifying this contrast places two requirements on the design. First, the task must come in two versions that are as similar as possible, differing mainly in whether the heuristic is available: in one, the heuristic leads to a naive answer that differs from the rational one; in the other, only the rational representation is available. Both answers must be known to the experimenter, so that each response can be located relative to them. Second, the motive must vary exogenously, while the incentive for accuracy stays fixed.

To meet these requirements, I use a 2×2 factorial design that crosses the availability of the heuristic with the presence of a directional motive (Figure 1). In *Control*, participants face neither.

In *Heuristic only*, they face an environment in which the naive representation is available but have no motive. In *Motive only*, they have a motive but the naive representation is not available. In *Motive*×*Heuristic*, they face both. Since the theory makes different predictions depending on whether the motive favours the naive or the rational conclusion, this last cell is split into *Aligned*, in which the motive and the heuristic push beliefs in the same direction, and *Opposed*, in which they push in opposite directions.

Each cell maps onto a part of the framework of Section 2. *Motive only* measures the within-representation distortion of Prediction 1, since only one representation is available. *Heuristic only* measures which representation individuals adopt absent a motive (Prediction 2). Predictions 3a and 3b concern the interaction itself. Comparing *Aligned* and *Opposed* with *Heuristic only* is not enough, since the difference also contains the within-representation distortion that *Motive only* measures. The predictions are therefore tested against an additive benchmark: under Prediction 3a, the error in the direction of the heuristic is larger in *Aligned* than the sum of the effects in *Heuristic only* and *Motive only*; under Prediction 3b, it is smaller in *Opposed* than their difference.

I implement this design in two pre-registered experiments on Prolific. The first, *CN*×*ME*, keeps the correlation-neglect task and the motivated-error payment of Exley and Kessler (2024). The second, *SN*×*IQ*, changes both ingredients: the heuristic is selection neglect, and the motive is to think well of one's own performance on a logical reasoning test rather than to secure money for oneself. Sections 5 and 6 describe each experiment and report its results; screenshots of the experimental instructions can be found in Appendix J.

In both experiments, participants were assigned to conditions by block randomisation and remained in the same condition throughout the study. Instructions are tailored to each condition, so participants learn their payment rule and information structure but are not told that other conditions exist.

5 Experiment 1: CN×ME

5.1 Design

The *CN*×*ME* experiment builds on the design of Exley and Kessler (2024) described in Section 3.1 and departs from it in two main respects. First, it adds the *Control* and *Motive only* cells, which measure the motivated distortion absent the heuristic. Second, the direction of the heuristic is fixed for each participant, so that *Aligned* and *Opposed* are between-subject conditions.

Participants first undergo the same calibration task as Exley and Kessler (2024) with Emergency USA as the charity.[14] In this task, I elicit participants' charity equivalent X. This is the amount of

[14]Participants are told that Emergency USA is a 501(c)(3) charitable organisation that provides healthcare assistance to individuals who would otherwise have little or no access to it, and that it describes its activities as follows: "EMERGENCY provides free, high-quality care to people affected by conflict and poverty."

money that makes them indifferent between receiving it themselves as a bonus payment and having the charity receive \$1.50. In the second part, participants complete six estimation rounds presented in a random order.[15] In each round, they have to compute the average of four estimates. How these estimates are shown and how the answer is paid differ across conditions.

The heuristic comes from how the four estimates are shown. In *No correlation*, participants observe them directly, so averaging them gives the correct answer. In *Correlation*, the information is correlated, as in Enke and Zimmermann (2019) and Exley and Kessler (2024): participants observe one estimate directly and learn about the other three only through three news channels, each of which reports the average of the observed estimate and one of the others. The structure is explained to participants and checked in comprehension questions. In *Correlation+*, the observed estimate is the highest of the four, so that correlation neglect pulls answers up; in *Correlation−*, it is the lowest, so that it pulls answers down.

The motive comes from the payment rule. In *No motive*, participants receive X if their guess is correct and nothing otherwise.[16] In *Motive*, as in Exley and Kessler (2024), if participants' estimate is correct, they receive X as a bonus payment and the charity receives \$1.50. If the guess is wrong, only one of the two payments is made, depending on the direction of the error: for participants with an upward motive, a guess that is too high pays X to the participant and a guess that is too low pays \$1.50 to the charity; for participants with a downward motive, the reverse. Because X is calibrated to be worth as much to the participant as \$1.50 for the charity, the direction of the error does not change the participant's payoff, and a correct guess remains strictly better. What the direction does change is who receives the payment, which gives participants a self-serving reason to err upward or downward.

Crossing the payment rule with the information structure yields the cells of Figure 1. *No motive* × *No correlation* is the *Control* cell, and *No motive* × *Correlation* forms the *Heuristic only* cell. *Motive* × *No correlation* is the *Motive only* cell. *Motive* × *Correlation* is *Aligned* when the motive and correlation neglect push in the same direction (an upward motive with *Correlation+*, a downward one with *Correlation−*), and *Opposed* otherwise, as in the re-analysis of Section 3. Both directions are randomised across participants and kept fixed across the six rounds.

Only one decision is paid. At the end of the study, either Part 1 or Part 2 is selected, each with probability one half. If Part 1 is selected, one row of the calibration price list is drawn at random and the option the participant chose in that row is implemented. If Part 2 is selected, one of the six rounds is drawn at random and paid according to the participant's payment rule. Participants are told this in advance, so that every decision may be the one that counts.

[15] The estimates and the signals participants observe in each round are reported in Table 15 in Appendix E.

[16] In *No motive* × *Correlation*, this departs from Exley and Kessler (2024), where the charity receives \$1.50 regardless of participants' answers.

5.2 Procedures

The experiment was pre-registered on AsPredicted (#284113, `https://aspredicted.org/yf22j6.pdf`) and run on Prolific between the 9th and the 10th of April 2026 with 1,212 participants. Participants received a fixed payment of \$2.00 to complete the study which had a median completion time around 14 minutes. Participants also accumulated an average bonus payment of \$0.44 and an average donation to the charity of \$0.77.

5.3 Results

The pre-registered outcome is participants' naiveté parameter χ described in Section 3.2: the error in the direction of the heuristic, divided by the error a fully naive respondent would make in the same round. In the correlated conditions, both the correct and the naive answer are computed from the signals participants observe. In *Control* and *Motive only*, where no naive answer exists, I divide the error by the naive error in the correlated version of the same round, and sign it in the direction of the motive. The main analysis summarises each participant's behaviour by the median χ across rounds; all results are robust to using the mean instead.

5.3.1 Do the motive and the heuristic add up?

To test whether the motive and the heuristic add up, I estimate an OLS regression of the median χ on a dummy for each condition, with *Control* as the reference category and heteroskedasticity-robust (HC3) standard errors. Under additivity, the coefficient on *Aligned* equals the sum of the coefficients on *Heuristic only* and *Motive only*, and the coefficient on *Opposed* equals their difference. I test these two equalities with two-sided F-tests. A larger χ in *Aligned* than the sum, or a smaller one in *Opposed* than the difference, would indicate that the availability of the heuristic makes individuals more susceptible to the motive (Predictions 3a and 3b).

Table 3 reports the regression results.[17] Both the motive and the heuristic distort answers on their own. In *Control*, the median χ averages 0.06. Correlated information raises it by 0.31 (SE 0.06, $p < 0.01$). The motive alone also distorts answers: in *Motive only*, the median χ is 0.12 higher than in *Control* (SE 0.05, $p < 0.05$).

The motive also reproduces part of the pattern documented by Exley and Kessler (2024). In *Aligned*, the median χ is 0.56 above *Control*, 0.25 more than in *Heuristic only* (SE 0.08, $p < 0.01$). In *Opposed*, it is 0.28 above *Control*, only 0.03 less than in *Heuristic only* (SE 0.08, $p = 0.72$). A motive aligned with the heuristic thus exacerbates errors, but, unlike in Exley and Kessler (2024), a motive opposed to it does not mitigate them.

The comparison with *Heuristic only*, however, does not account for the distortion the motive produces on its own. Under additivity, the median χ in *Aligned* would exceed that in *Control* by

[17] Table 16 in Appendix F summarises the rusults of the tests.

Table 3: Naiveté by condition

	(1) Median χ	(2) Mean χ	(3) Median $\lvert\chi\rvert$
Constant (*Control*)	0.058^{*}	0.044^{*}	0.110^{**}
	(0.033)	(0.024)	(0.043)
Heuristic only	0.312^{***}	0.333^{***}	0.491^{***}
	(0.059)	(0.054)	(0.060)
Motive only	0.122^{**}	0.170^{***}	0.226^{***}
	(0.054)	(0.054)	(0.060)
Aligned	0.562^{***}	0.590^{***}	0.657^{***}
	(0.067)	(0.068)	(0.068)
Opposed	0.284^{***}	0.312^{***}	0.621^{***}
	(0.069)	(0.069)	(0.066)
Observations	1,212	1,212	1,212
R^2	0.039	0.033	0.071

Notes: OLS of the naiveté parameter χ on indicators for *Heuristic only*, *Motive only*, *Aligned* and *Opposed*, with *Control* as the reference category. One observation per participant: the median (column 1) and the mean (column 2) of χ over the six rounds, and the median of its absolute value (column 3). The constant is the mean in *Control*. Heteroskedasticity-robust (HC3) standard errors in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

$0.31 + 0.12 = 0.43$. The observed difference, 0.56, is 0.13 larger (SE 0.09), but the excess is not statistically significant ($p = 0.17$). Most of the exacerbation relative to *Heuristic only* is therefore what one would expect if the motive simply added the distortion it produces when the heuristic is not available.

In *Opposed*, additivity implies a median χ of $0.31 - 0.12 = 0.19$ above *Control*. The observed difference is 0.28, which is 0.09 *larger* (SE 0.10, $p = 0.32$). If anything, the motive reduces the error in the direction of the heuristic by less than it distorts answers on its own, which is the opposite of what Prediction 3b implies. Neither test rejects additivity. The same holds when each participant's answers are summarised by their mean rather than their median χ (column 2 of Table 3): the excess is 0.09 in *Aligned* (SE 0.10, $p = 0.37$) and 0.15 in *Opposed* (SE 0.10, $p = 0.13$).

Absolute errors tell a similar story. As in Section 3.3, if the motive in *Opposed* moved respondents towards the rational answer, the median of their round-specific absolute χ should fall relative to *Heuristic only*. It rises instead, by 0.13 (SE 0.06, $p = 0.046$; column 3 of Table 3). This result is fragile, however: it rests on the tails of the distribution,[18] and it is no longer significant at the 5% level when the pre-registered exclusions are applied (p between 0.07 and 0.18).

These conclusions are robust to the pre-registered exclusions of respondents who copied the

[18]Across respondents, the median of this measure is similar in the two conditions (0.64 in *Opposed* and 0.60 in *Heuristic only*), and a rank-sum test does not reject equality of the two distributions ($p = 0.11$).

signals or completed the study unusually fast: the additivity test for *Aligned* has p between 0.11 and 0.21, and the test for *Opposed* never approaches significance. Excluding the 163 respondents whose charity equivalent is zero, for whom the calibration cannot make the two payments equally valuable, makes the excess in *Aligned* significant ($p < 0.01$), but only because the effect of the motive in *Motive only* disappears in that sample; I discuss this in Appendix F.

The pre-registered exploratory analysis of motive strength points in the same direction. Its proxy is the charity equivalent X: participants with a lower X care less about the donation, and so have a stronger reason to err in the direction that secures the bonus. As a validity check, X predicts answers in *Motive only*: each additional dollar of X lowers the median χ by 0.31 (SE 0.10, $p < 0.01$). If the heuristic made individuals more susceptible to the motive, a stronger motive should raise errors in *Aligned* even more. It does not: in *Aligned*, the median χ is unrelated to X (slope 0.06, SE 0.14, $p = 0.70$). In *Opposed*, by contrast, participants with a stronger motive have a lower median χ (slope 0.52, SE 0.15, $p < 0.01$), so the motive does pull them away from the naive answer (Table 18 in Appendix F).

5.3.2 Does the motive change the representation?

To separate the two channels of the framework, I classify respondents with the bands of Section 3.2: a respondent is rational if their median χ lies in $[-1/3, 1/3]$, naive if it lies in $(1/3, 5/3)$, and makes other errors otherwise. As in Section 3.2, I estimate multinomial logits of type membership on condition. If the motive operates through the choice of representation, *Aligned* should raise the share of naive respondents relative to *Heuristic only*, and *Opposed* should raise the share of rational ones. Figure 5 in Appendix I shows the distribution of respondents' median χ in each condition. Table 4 reports the resulting changes in the share of each type in *Aligned* and *Opposed* relative to *Heuristic only*, together with the shares in *Heuristic only*.

In *Control*, 90% of respondents are classified as rational. In *Heuristic only*, the distribution of the median χ has the two peaks documented by Enke and Zimmermann (2019), at the rational and at the naive benchmark. In *Motive only*, most respondents still answer correctly (83% are classified as rational), so the effect of the motive in the regression comes from a minority of respondents who err in the favoured direction.

When the motive operates on top of the heuristic, it leaves the shape of the distribution of the median χ largely unchanged: in both *Aligned* and *Opposed*, the two peaks at the rational and at the naive benchmark remain. In *Heuristic only*, 48.5% of respondents are classified as rational, 44.6% as naive, and 6.9% as making other errors. In *Aligned*, the corresponding shares are 41.5%, 50.0% and 8.5%; in *Opposed*, 44.8%, 44.8% and 10.4%. The multinomial logit with *Heuristic only* as the reference confirms that none of these differences is statistically significant.[19] In *Aligned*, the share of

[19]The pre-registration specifies one binary logit per type. Because condition is the only regressor, both the binary logits and the multinomial logit reproduce the raw share of each type in each condition, so the changes in shares are

Table 4: Representation adopted by condition

	Rational	Naive	Other errors	Joint p
Aligned	-0.070	0.054	0.016	0.283
	(0.044)	(0.044)	(0.024)	
Opposed	-0.037	0.002	0.035	0.337
	(0.044)	(0.043)	(0.024)	
Share in *Heuristic only*	0.485	0.446	0.069	
Observations		804		

Notes: Changes in the share of each type relative to *Heuristic only*, from a multinomial logit of type membership on indicators for *Aligned* and *Opposed*, estimated on these three conditions. Because the model is saturated, the changes equal the differences in raw shares, as in the pre-registered binary logits. Types are based on the median naïveté parameter χ: rational if $\chi \in [-1/3, 1/3]$, naive if $\chi \in (1/3, 5/3)$, other errors otherwise. Delta-method standard errors in parentheses; joint p is the Wald test that the type distribution equals that in *Heuristic only*. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

naive respondents rises by 5.4 percentage points (SE 4.4) and the share of rational respondents falls by 7.0 percentage points (SE 4.4); the type distribution does not differ from that in *Heuristic only* ($p = 0.28$). In *Opposed*, the share of rational respondents does not rise, but falls by 3.7 percentage points (SE 4.4), and the share of naive respondents is unchanged (+0.2 percentage points, SE 4.3; joint $p = 0.34$). As in the data of Exley and Kessler (2024), respondents in *Aligned* are not more likely to adopt the naive representation, and respondents in *Opposed* are not more likely to adopt the rational one.

The same holds under alternative classifications (Table 17 in Appendix F): tighter windows around the two benchmarks, the requirement that a respondent's answers fall in the same type in at least four of the six rounds, and a grid of cut-offs. In none of them does the share of naive respondents rise in *Aligned*, or the share of rational respondents rise in *Opposed*. The one change that some classifications detect goes in neither predicted direction: with the narrowest windows around the benchmarks, the share of rational respondents in *Aligned* falls by 9 percentage points ($p < 0.05$), without a corresponding rise in the share of naive respondents.

Unlike in the data of Exley and Kessler (2024), the motive does not push many respondents into other errors. Their share rises only slightly, from 6.9% in *Heuristic only* to 8.5% in *Aligned* and 10.4% in *Opposed*. However, when the motive does change answers, it does so mostly at the margins of the distribution in line with the results from Exley and Kessler (2024). In *Aligned*, 5.9% of respondents overshoot the naive answer ($\chi > 5/3$), against 0.4% in *Heuristic only*; a similar share does so in *Motive only* (5.3%), where no naive representation is available. In *Opposed*, 7.7% err in the direction opposite to correlation neglect ($\chi < -1/3$), against 6.4% in *Heuristic only*.

identical.

The smaller shift into other errors than in Exley and Kessler (2024) may reflect a difference in the respondent pool: this experiment recruits on Prolific rather than MTurk.

6 Experiment 2: SN×IQ

CN×ME finds no evidence that the availability of a naive representation makes individuals more susceptible to the motive. SN×IQ tests whether this result extends to a different heuristic, selection neglect, and to a motive that does not involve money but concerns participants' own cognitive skills, a domain in which motivated beliefs are well documented. It also narrows a difference that CN×ME shares with the design of Exley and Kessler (2024): there, the problem that admits the heuristic is also the more complex one. Without the correlational structure, participants observe the four estimates and only need to average them; with it, they must first recover the estimates from the news reports. An interaction between the motive and the heuristic, such as the exacerbation documented by Exley and Kessler (2024), could therefore reflect complexity rather than the availability of a naive representation. In SN×IQ, a respondent who accounts for the selection only needs to add the missing report, whose value they know, before averaging. The simpler correction also speaks to a second reading of the absence of switches in CN×ME. A motive can lead individuals to the rational representation only if they are able to construct it. If correcting for correlation is too demanding for those who neglect it, respondents in *Opposed* may not switch because the rational representation is out of their reach, not because the motive does not act on the choice of representation. Because accounting for the selection only requires adding a known value, the rational representation is easier to construct in SN×IQ, which leaves the motive more room to move individuals towards it.

6.1 Design

Participants first take a 20-question logical reasoning test which is identical across conditions. The questions are figural pattern-completion items with a maximum time of 20 seconds per question and no feedback.[20] The questions are presented in 4 blocks of 5 questions each and are ordered by increasing difficulty.[21] Each correct answer earns a bonus of \$0.05. Participants are told that tests of this kind are widely used to measure intelligence.

Participants then estimate the percentage of the 20 questions a test-taker, either themselves or another participant, answered correctly, and earns an accuracy bonus of \$0.20 if their estimate is correct.[22] We tell participants that their estimates should be based on the reports of automated

[20] During the test, participants see a timer on the screen and are told that if they don't provide an answer within the allowed time, the question is counted as wrong. Additionally, participants who let the timer run out on more than 10 of the 20 questions, are screened out after the test.

[21] The order is based on the success rate observed in a dedicated pilot study.

[22] The other participant's score is taken from the results of a pilot study. I restrict the range of the other participants' score to 40, 50 or 60%. Again, this is to ensure that, in the *Selection* conditions, exactly one report is missing.

reviewers. Participants are told that each question was checked by exactly one reviewer, and that each reviewer checked a different set of 5 questions and computed the percentage of correct answers in that set. To keep the task simple, they are also explicitly told that the average of the percentages calculated by the reviewers equals the score to estimate.[23]

In the conditions without the possibility to engage in a heuristic (*No selection*), all four reviewers report the percentages they calculated. Therefore, averaging the reports gives the correct answer. In the conditions with the heuristic (*Selection*), participants are told that a reviewer does not provide a report if the percentage of correct answers in the set they checked equals a given value. In *Selection+*, reports of 0% are not shown, so the reports that are shown "overstate" the score if participants ignore the missing piece of information; in *Selection−*, reports of 100% are not shown, so they "understate" it. Since participants can infer whether a reviewer is not providing a report (e.g. if they observe only three reports instead of four) and they know the value of a missing report, a respondent who accounts for the selection can recover the correct answer. A respondent who neglects it and averages the reports shown commits the error of selection neglect (Enke, 2020). The two answers require the same arithmetic, an average of four or of three numbers: the rational answer only requires realising that the missing 0% or 100% is there even though it is not shown.

The motive comes from the target of the estimate. In *Own* (motive), participants estimate their own score; in *Other* (no motive), they estimate the score of another participant. Since a high score on an IQ test is a congenial belief (Eil and Rao, 2011; Möbius et al., 2022) only if the score is about oneself, participants estimating their own score have a motive to overestimate it.

Crossing the target of the elicitation with the information structure yields the conditions of Figure 1. *Other* × *No selection* is the *Control* condition, and *Other* × *Selection* form the *Heuristic only* conditions: *Heuristic only* (+) with *Selection+* and *Heuristic only* (−) with *Selection−*. *Own* × *No selection* is the *Motive only* condition; *Own* × *Selection+* is *Aligned*, since the motive and selection neglect both push the estimate up, and *Own* × *Selection−* is *Opposed*, since selection neglect pushes it down.

After the first estimate, participants make a second one about the target they have not yet estimated: participants in *Own* estimate the other participant's score, and participants in *Other* their own. The information structure stays the same, and the reports are built in the same way from the new target's score. The second estimate is not announced in advance and is not paid. Since it differs from the first in several respects at once (the target, the experience of having already completed the task, and the absence of an accuracy bonus), it cannot isolate the effect of the motive. The main analyses therefore concern only the first estimate; the analyses of the second estimate are exploratory and reported in Appendix G.

[23] Participants are not told how the sets are formed: the sets are built from the participant's answers so that the four percentages are as close as possible to 100%, 60%, 40%, and 0%. This ensures that, in the *Selection* conditions, exactly one report is missing.

6.2 Procedures

The experiment was pre-registered on AsPredicted (#320081, `https://aspredicted.org/9cp5yn.pdf`) and run on Prolific on 26 October 2026 with 895 participants, 762 of whom passed the screening after the test and form the sample for the main analyses.[24] Participants received a fixed payment of \$1.90 and an average bonus of \$0.56, and the median completion time was 12 minutes.

6.3 Results

The pre-registered outcome is the error of the score estimate, in percentage points: the estimate minus the true score, so that a positive error is an overestimate.

6.3.1 Do the motive and the heuristic add up?

To test whether the motive and the heuristic add up, I estimate an OLS regression of the error on a dummy for each condition, with *Control* as the reference category and HC3 standard errors.[25] Under additivity, the coefficient on *Aligned* equals the sum of the coefficients on *Heuristic only* (+) and *Motive only*, and the coefficient on *Opposed* equals the sum of the coefficients on *Heuristic only* (−) and *Motive only*. Under Prediction 3a, the departure from additivity in *Aligned* is positive: the overestimate in *Aligned* exceeds the sum of the two separate effects. Under Prediction 3b, the departure from additivity in *Opposed* is also positive: the underestimate that selection neglect produces is smaller in *Opposed* than the two separate effects imply.

Table 5 reports the estimates. The heuristic distorts estimates in the direction of selection neglect. In *Control*, estimates of the other participant's score exceed it by 4.1 percentage points on average (SE 1.3, $p < 0.01$), although the median error is zero. Relative to *Control*, the error is 7.0 percentage points higher in *Heuristic only* (+) (SE 1.7, $p < 0.01$) and 7.9 percentage points lower in *Heuristic only* (−) (SE 1.9, $p < 0.01$).

The motive, by contrast, does not shift the average estimate in the full sample. In *Motive only*, participants overestimate their own score by 1.9 percentage points more than they overestimate the other participant's score in *Control*, but the difference is not statistically significant (SE 2.0, $p = 0.34$), and the median error is zero in both conditions. The motive changes the spread of estimates instead (Figure 6 in Appendix I). Estimates of one's own score are more dispersed (standard deviation 18.7 against 13.0 percentage points; Levene test $p = 0.01$) and less often correct (35%

[24] In addition to exlcuding participants who were timed out in more than 10 questions, participants who answer fewer than 6 or more than 14 questions correctly were excluded and did not take part in the estimation task. This exclusion ensures that, in the *Selection* conditions, the reports can be built so that exactly one of them is missing; outside this range, more than one report would be hidden.

[25] This regression is a reparametrisation of the pre-registered specification, which regresses the error on a dummy for *Own*, dummies for *Selection+* and *Selection−*, and the interactions of *Own* with each selection dummy. The two models have the same fit and standard errors, and the departures from additivity below equal the pre-registered interaction coefficients.

Table 5: SN×IQ: error of the first estimate by cell

	(1) Error	(2) Absolute error
Constant (*Control*)	4.14^{***}	7.93^{***}
	(1.28)	(1.09)
Heuristic only (+)	7.02^{***}	5.87^{***}
	(1.71)	(1.40)
Heuristic only (−)	-7.94^{***}	3.54^{**}
	(1.87)	(1.46)
Motive only	1.93	4.39^{***}
	(2.04)	(1.69)
Aligned	8.09^{***}	8.04^{***}
	(1.91)	(1.56)
Opposed	-5.44^{***}	5.39^{***}
	(2.02)	(1.48)
Observations	762	762
R^2	0.125	0.037

Notes: OLS of the error of the first estimate (estimate minus target score, in percentage points; column 1) and of its absolute value (column 2) on indicators for *Heuristic only* with *Selection+* (+) and with *Selection−* (−), *Motive only*, *Aligned* and *Opposed*, with *Control* as the reference category. The constant is the mean in *Control*. Column 1 is equivalent to the pre-registered main specification. Column 2 is exploratory. Heteroskedasticity-robust (HC3) standard errors in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

against 47% within 2.5 percentage points of the target), so that the absolute error is 4.4 percentage points larger than in *Control* (SE 1.7, $p = 0.01$). In an exploratory analysis reported in Appendix H, I show that the motive does shift the average estimate among participants who scored below 50% (the sample median, and an intuitive "low score" benchmark) for whom the reports convey unfavourable news: in this subsample, estimates of one's own score exceed estimates of the other participant's score by 8.2 percentage points (SE 3.6, $p = 0.02$), against -2.4 percentage points among the other participants (SE 2.4, $p = 0.32$). The larger dispersion therefore comes mainly from participants who scored below the median and overestimate their own score.

Two features of the design can explain why the motive leaves the average estimate unchanged in the full sample. First, as implied from the subsample resutls, the motive is weaker for participants who already have a relatively high score: since the reports already convey favourable news, they have less to gain from distorting their estimate. Second, without selection the evidence provided to participants is quite conclusive as the four reports together reveal the score exactly and leaves little room to distort it within the representation.

Since the motive has no detectable effect on its own in the full sample, the additivity tests are close to tests of whether the motive has any effect when the heuristic is available, and neither detects one.[26] The departure from additivity is -0.9 percentage points in *Aligned* (SE 2.7, $p = 0.75$) and

[26]The same holds when the sample is split at the median own score (Appendix H): with selection, the motive

0.6 percentage points in *Opposed* (SE 2.9, $p = 0.85$). Neither Prediction 3a nor Prediction 3b is supported. The comparison with *Heuristic only* of Exley and Kessler (2024) gives the same answer: the error in *Aligned* exceeds that in *Heuristic only* (+) by 1.1 percentage points (SE 1.8, $p = 0.56$), and the error in *Opposed* exceeds that in *Heuristic only* (−) by 2.5 percentage points (SE 2.1, $p = 0.23$). Absolute errors do not separate the two conditions either: the absolute error in *Opposed* is 1.9 percentage points larger than in *Heuristic only* (−) (SE 1.4, $p = 0.18$; column 2 of Table 5).

These conclusions hold under each of the pre-registered robustness exclusions, when the sample is restricted to participants whose own score lies between 40% and 60%, and when the target score is interacted with each condition dummy. Across these specifications, the departure from additivity ranges from −4.0 to −0.7 percentage points in *Aligned* and from −1.3 to 3.7 percentage points in *Opposed*, and none is statistically significant.

6.3.2 Does the motive change the representation?

To separate the two channels of the framework, I classify estimates following the pre-registration: an estimate is rational if it lies within 2.5 percentage points of the true score, fully naive if it lies within 2.5 percentage points of the naive benchmark, the average of the reports shown, partially naive if it lies strictly between the two, and other otherwise.[27] Within each selection condition, I estimate linear probability models of the indicator for each type on *Own*, with HC3 standard errors. If the motive operates through the choice of representation, *Aligned* should raise the share of fully naive estimates relative to *Heuristic only* (+), and *Opposed* should raise the share of rational ones relative to *Heuristic only* (−). Table 6 reports the estimates.

Neither prediction holds. In *Aligned*, 26.8% of estimates are fully naive, against 29.5% in *Heuristic only* (+) (difference −2.6 percentage points, SE 5.5, $p = 0.63$). In *Opposed*, the share of rational estimates does not rise but falls, from 28.7% to 17.1% (−11.6 percentage points, SE 5.3, $p = 0.03$). The shares of the other types change little: in *Aligned*, the share of rational estimates is 15.9% against 18.6% (−2.7 percentage points, SE 4.7, $p = 0.57$), and in *Opposed* the share of fully naive estimates is 15.4% against 17.1% (−1.6 percentage points, SE 4.7, $p = 0.73$). The estimates that the motive moves away from the rational answer in *Opposed* end up partially naive (23.6% against 17.8%) or other (43.9% against 36.4%). The fall in the share of rational estimates in *Opposed* is between 11.0 and 14.4 percentage points ($p < 0.05$) under each single pre-registered exclusion except the one based on the ability to take averages, where it is 8.7 percentage points (SE 6.4); it is 7.2 percentage points (SE 7.3) when all exclusions are applied together, which removes about 40% of the sample.

has no significant effect either among participants who scored below 50% or among the others (all $p > 0.35$), and neither departure from additivity is significant in either group. Among low scorers, where the motive raises estimates without selection, the departures are negative (−7.3 and −6.1 percentage points, $p = 0.14$ and $p = 0.24$).

[27] Rational takes precedence if an estimate is within 2.5 percentage points of both benchmarks. Without selection, the naive benchmark equals the true score, so estimates can only be rational or other.

Table 6: SN×IQ: classification of first estimates

	(1) Rational	(2) Fully naive	(3) Partially naive	(4) Other
Panel A: *Selection+*				
Constant (*Heuristic only* (+))	18.6	29.5	23.3	28.7
	(3.5)	(4.0)	(3.7)	(4.0)
Own-first (*Aligned*)	−2.7	−2.6	2.1	3.2
	(4.7)	(5.5)	(5.3)	(5.7)
Observations	267	267	267	267
Panel B: *Selection−*				
Constant (*Heuristic only* (−))	28.7	17.1	17.8	36.4
	(4.0)	(3.3)	(3.4)	(4.3)
Own-first (*Opposed*)	−11.6**	−1.6	5.7	7.5
	(5.3)	(4.7)	(5.1)	(6.2)
Observations	252	252	252	252

Notes: Linear probability models of an indicator for each class of first estimate on an indicator for *Own-first*, estimated separately within *Selection+* (Panel A) and *Selection−* (Panel B). Rational: within 2.5 percentage points of the target score. Fully naive: within 2.5 percentage points of the average of the reports shown. Partially naive: strictly between the two. Other: anything else. The constant is the share in *Heuristic only*; the *Own-first* coefficient is the difference between the share in the cell with the motive and that in *Heuristic only*. Coefficients in percentage points. Within each panel the four columns sum to 100 (constant) and 0 (*Own-first*). The pre-registered hypotheses concern column 2 in Panel A and column 1 in Panel B; columns 3 and 4 are exploratory. Heteroskedasticity-robust (HC3) standard errors in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

7 Conclusion

This paper asked whether the motive to hold favourable beliefs shapes which mental representation individuals adopt, in addition to how they distort beliefs within a representation. I formalised this possibility in a theoretical framework in which individuals first choose among defensible representations and then distort the belief their chosen representation licenses, and I tested its predictions in three datasets: a re-analysis of the correlation-neglect experiment of Exley and Kessler (2024) and two new pre-registered experiments that cross the availability of a heuristic with a directional motive.

The three datasets give a consistent answer. In none of them does the motive move individuals towards the representation that licenses the conclusion it favours. In all three datasets, the share of naive respondents does not rise when the motive and the heuristic are aligned, compared with when the heuristic is available absent the motive, and the share of rational respondents does not rise when the motive favours the conclusion of the rational representation. Nor do the two experiments find robust departures from additivity. In CN×ME, the error in *Aligned* exceeds the sum of the effects of the motive and of the heuristic by an amount that is not statistically significant; in SN×IQ, both departures are close to zero.

For the framework of Section 2, these results locate motivated reasoning in the second stage

rather than in the first. The results suggest that motives distort beliefs within a representation, as standard models assume, but do not tip the choice between the rational and the naive representation. Within the framework, this is what one expects when individuals rarely find the two representations comparably defensible, the condition under which Prediction 3 applies: if most respondents either see the correlation or do not, the motive has no close call to tip, and acts only on the belief their representation delivers. The bimodal distributions in every condition, with peaks that the motive barely moves, are consistent with this reading.

This explanation suggests that the choice of mental representation is better seen as intuitive than as deliberate. In the framework, defensibility is measured by the salience of a representation's attention profile, following Bordalo et al. (2026): which features of the problem stand out determines which representation comes to mind. If this happens quickly and automatically, as individuals first read the problem, the representation is less chosen than *seen*. By the time the motive comes into play, the individual has already settled on a reading of the evidence, and the motive can only act on the belief that reading delivers, where some room for deliberate adjustment remains. A motivated choice between representations is also demanding in a further sense: to pick the representation that licenses the favoured conclusion, an individual must be aware, at least implicitly, that the representations lead to different conclusions, and which conclusion each leads to. That awareness may be self-defeating. An individual who has worked out both the naive and the rational answer has, in doing so, noticed the feature of the problem that the naive representation ignores, and can no longer adopt it in good faith; the heuristic is defensible precisely to those who do not see the alternative. On this view, the two stages of the framework differ not only in what they choose, a representation or a belief, but in how they choose it: the first stage is governed by attention, which the motive does not reach, and the second by reasoning, which it does. Motives would then shape which representation individuals adopt only when the representation is itself the object of deliberation, for instance when competing readings of the evidence are explicitly put to them, as in the competing narratives of Eliaz and Spiegler (2020). There, every narrative on offer is consistent with the long-run correlations individuals actually observe, so knowing the alternatives does not undermine any of them, and individuals can adopt the most hopeful narrative without contradicting the data.

Several caveats qualify these conclusions. First, the null results are bounded rather than exact zeros. The confidence intervals on the departures from additivity in SN×IQ run from about -6 to $+6$ percentage points, and the excess in *Aligned* in CN×ME (0.13 in units of χ) is imprecisely estimated; interactions of moderate size cannot be ruled out. Second, the classification of respondents relies on bands around the two benchmarks, so a motive that shifts estimates by a fixed amount can move respondents out of a type without any change in the representation they hold. In SN×IQ in particular, an additive self-enhancement in estimates of one's own score would lower the share of rational estimates in *Opposed* even if no participant changed representation; the fall in that share

should therefore not be read as evidence that the motive pulls individuals away from the rational representation. Third, the motive in CN×ME rests on a calibration of the charity donation that cannot make every participant indifferent between the two payments, and the excess in *Aligned* is sensitive to how participants for whom it fails are treated.

The experiments also share features that limit how far the conclusions travel. Both use stylised inference tasks in which the rational answer can be computed from the information shown, both run online with small stakes, and in both the heuristic is invited by the environment rather than learned. Outside the laboratory, representations are rarely laid out side by side, the evidence is often less conclusive, and motives may operate earlier, through which information individuals seek or remember rather than through how they represent the information they already have. The results speak to the narrower question of whether, holding the evidence fixed, a motive steers individuals towards whichever available representation favours it.

Within that scope, the answer is no, and the implication is a simplifying one. If motivated reasoning and heuristic reasoning largely add up, the distortion a motive produces is much the same whether or not the environment invites a heuristic, so the two can be studied, and corrected, separately. The flip side is that motives should not be expected to pull individuals away from a heuristic: a motive aligned with the truth does not lead individuals to the rational representation.

References

Sandro Ambuehl, Rahul Bhui, and Heidi C Thysen. Mental models of causal structure in economics and psychology. *arXiv preprint arXiv:2603.29070*, 2026.

Andrea Amelio and Florian Zimmermann. Motivated memory in economics—a review. *Games*, 14 (1):15, 2023.

Roland Bénabou and Jean Tirole. Mindful economics: The production, consumption, and value of beliefs. *Journal of Economic Perspectives*, 30(3):141–164, 2016.

J Aislinn Bohren and Daniel N Hauser. Misspecified models in learning and games. *Annual Review of Economics*, 17(1):427–451, 2025.

Lukas Bolte and Tony Q Fan. Motivated mislearning: The case of correlation neglect. *Journal of Economic Behavior & Organization*, 217:647–663, 2024.

Pedro Bordalo, John J Conlon, Nicola Gennaioli, Spencer Y Kwon, and Andrei Shleifer. Memory and probability. *The Quarterly Journal of Economics*, 138(1):265–311, 2023.

Pedro Bordalo, John Conlon, Nicola Gennaioli, Spencer Kwon, and Andrei Shleifer. How people use statistics. *Review of Economic Studies*, 93(1):250–285, 2026.

Anat Bracha and Donald J Brown. Affective decision making: A theory of optimism bias. *Games and Economic Behavior*, 75(1):67–80, 2012.

Andrew Caplin and John Leahy. Psychological expected utility theory and anticipatory feelings. *The Quarterly Journal of Economics*, 116(1):55–79, 2001.

Alexander Coutts. Testing models of belief bias: An experiment. *Games and Economic Behavior*, 113:549–565, 2019.

Jason Dana, Roberto A Weber, and Jason Xi Kuang. Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness. *Economic Theory*, 33(1):67–80, 2007.

Peter H Ditto and David F Lopez. Motivated skepticism: Use of differential decision criteria for preferred and nonpreferred conclusions. *Journal of personality and social psychology*, 63(4):568, 1992.

David Eil and Justin M Rao. The good news-bad news effect: asymmetric processing of objective information about yourself. *American Economic Journal: Microeconomics*, 3(2):114–138, 2011.

Kfir Eliaz and Ran Spiegler. A model of competing narratives. *American Economic Review*, 110 (12):3786–3816, 2020.

Benjamin Enke. What you see is all there is. *The Quarterly Journal of Economics*, 135(3):1363–1398, 2020.

Benjamin Enke and Florian Zimmermann. Correlation neglect in belief formation. *The Review of Economic Studies*, 86(1):313–332, 2019.

Ignacio Esponda and Demian Pouzo. Berk–nash equilibrium: A framework for modeling agents with misspecified models. *Econometrica*, 84(3):1093–1130, 2016.

Christine L Exley and Judd B Kessler. Motivated errors. *American Economic Review*, 114(4): 961–987, 2024.

Tristan Gagnon-Bartsch, Matthew Rabin, and Joshua Schwartzstein. *Channeled attention and stable errors*. Harvard Business School Boston, 2018.

Ananda Ganguly and Joshua Tasoff. Fantasy and dread: The demand for information and the consumption utility of the future. *Management Science*, 63(12):4037–4060, 2017.

Nicola Gennaioli and Andrei Shleifer. What comes to mind. Technical report, National Bureau of Economic Research, 2009.

Zvi Ginossar and Yaacov Trope. Problem solving in judgment under uncertainty. *Journal of Personality and social Psychology*, 52(3):464, 1987.

Jeanne Hagenbach and Charlotte Saucet. Motivated skepticism. Working paper, Sciences Po, 2023.

Emily C Haisley and Roberto A Weber. Self-serving interpretations of ambiguity in other-regarding behavior. *Games and economic behavior*, 68(2):614–625, 2010.

Dan M Kahan, Ellen Peters, Erica Cantrell Dawson, and Paul Slovic. Motivated numeracy and enlightened self-government. *Behavioural public policy*, 1(1):54–86, 2017.

Botond Köszegi. Ego utility, overconfidence, and task choice. *Journal of the European Economic Association*, 4(4):673–707, 2006.

Ziva Kunda. The case for motivated reasoning. *Psychological bulletin*, 108(3):480, 1990.

Gilat Levy and Ronny Razin. Correlation neglect, voting behavior, and information aggregation. *American Economic Review*, 105(4):1634–1645, 2015.

Guy Mayraz. Wishful thinking. *Unpublished Manuscript. Available at SSRN 1955644*, 2011.

Markus M Möbius, Muriel Niederle, Paul Niehaus, and Tanya S Rosenblat. Managing self-confidence: Theory and experimental evidence. *Management Science*, 68(11):7793–7817, 2022.

Robin Musolff and Florian Zimmermann. Model uncertainty. 2025.

Emily Oster, Ira Shoulson, and E Ray Dorsey. Optimal expectations and limited medical testing: Evidence from huntington disease. *American Economic Review*, 103(2):804–830, 2013.

Eyal Peer, David Rothschild, Andrew Gordon, Zak Evernden, and Ekaterina Damer. Data quality of platforms and panels for online behavioral research. *Behavior research methods*, 54(4):1643–1662, 2022.

Rasyid Sanitioso and Ziva Kunda. Ducking the collection of costly evidence: Motivated use of statistical heuristics. *Journal of Behavioral Decision Making*, 4(3):161–176, 1991.

Joshua Schwartzstein and Adi Sunderam. Using models to persuade. *American Economic Review*, 111(1):276–323, 2021.

Ran Spiegler. Bayesian networks and boundedly rational expectations. *The Quarterly Journal of Economics*, 131(3):1243–1290, 2016.

Ran Spiegler. Behavioral implications of causal misperceptions. *Annual Review of Economics*, 12 (1):81–106, 2020.

Charles S Taber and Milton Lodge. Motivated skepticism in the evaluation of political beliefs. *American journal of political science*, 50(3):755–769, 2006.

Michael Thaler. Gender differences in motivated reasoning. *Journal of Economic Behavior & Organization*, 191:501–518, 2021.

Amos Tversky and Daniel Kahneman. Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *science*, 185(4157):1124–1131, 1974.

Amos Tversky, Daniel Kahneman, et al. Evidential impact of base rates. *Judgment under uncertainty: Heuristics and biases*, 153, 1982.

Florian Zimmermann. The dynamics of motivated beliefs. *American Economic Review*, 110(2): 337–363, 2020.

A Solution of the Belief-Choice Problem

A.1 Belief choice within a mental representation

This appendix solves the second stage of the problem in Section 2.2. In this subsection, I describe how the DM chooses γ to maximize the utility function described in (2), over $\gamma \in \Delta(\Theta)$.

For a tractable, closed-form solution that allows examining comparative statics, I impose two restrictions. First, let v be locally linear around g_m: $\partial v/\partial\gamma(\theta) \approx \psi(\theta)$. $\psi(\theta)$ is independent of γ and offers the state-by-state marginal reward of the motive. This allows me to write $v(\gamma) \approx v(g_m) + \sum_\theta \psi(\theta)\big(\gamma(\theta) - g_m(\theta)\big)$.[28] Second, approximate the divergence $\mathrm{KL}(\gamma|g_m)$ by its second-order Taylor expansion around $\gamma = g_m$. Define the spin $x(\theta) := \gamma(\theta) - g_m(\theta)$, so that $\sum_\theta x(\theta) = 0$ since both γ and g_m are probability distributions. Substituting $\gamma(\theta) = g_m(\theta) + x(\theta)$ and Taylor-expanding the log to second order in x gives

$$\begin{aligned}
\mathrm{KL}(\gamma|g_m) &= \sum_\theta \gamma(\theta) \ln \frac{\gamma(\theta)}{g_m(\theta)} \\
&= \sum_\theta \big(g_m(\theta) + x(\theta)\big) \ln\left(1 + \frac{x(\theta)}{g_m(\theta)}\right) \\
&\approx \sum_\theta \big(g_m(\theta) + x(\theta)\big) \left(\frac{x(\theta)}{g_m(\theta)} - \frac{1}{2}\left(\frac{x(\theta)}{g_m(\theta)}\right)^2\right) \\
&= \sum_\theta \left(x(\theta) + \frac{1}{2}\frac{x(\theta)^2}{g_m(\theta)}\right) + O(x^3) \\
&= \frac{1}{2}\sum_\theta \frac{x(\theta)^2}{g_m(\theta)} + O(x^3),
\end{aligned} \tag{9}$$

where the last line uses $\sum_\theta x(\theta) = 0$ to cancel the first-order term. Dropping the $O(x^3)$ remainder leaves the χ^2-approximation $\mathrm{KL}(\gamma|g_m) \approx \frac{1}{2}\sum_\theta x(\theta)^2/g_m(\theta)$, valid for directions satisfying $\sum_\theta x(\theta) = 0$.

This leaves me a much easier problem: substituting both approximations into (2), the DM's problem becomes the quadratic program

$$\max_x \; v(g_m) + \sum_\theta \psi(\theta)\, x(\theta) \;-\; \frac{1}{2\lambda}\sum_\theta \frac{x(\theta)^2}{g_m(\theta)} \qquad \text{s.t.} \;\; \sum_\theta x(\theta) = 0. \tag{10}$$

[28] This is exact, not an approximation, whenever, as in Caplin and Leahy (2001), v is itself linear in γ (e.g. the anticipatory-expected-utility form $v(\gamma) = \sum_\theta \gamma(\theta)u(\theta)$ for a fixed per-state payoff u, in which case $\psi = u$). It is a genuine approximation, valid only for γ^* close to g_m, whenever v is curved in γ: any curvature of v (e.g. diminishing psychological returns to an increasingly favorable belief) is then discarded, and whatever resistance to further distortion that curvature would have provided is instead attributed entirely to the divergence cost $1/\lambda$.

The Lagrangian first-order condition on (10), with multiplier ν on the adding-up constraint, is

$$\begin{aligned} \psi(\theta) - \frac{x^*(\theta)}{\lambda g_m(\theta)} &= \nu \qquad \text{for every } \theta \in \Theta \\ \iff \quad x^*(\theta) &= \lambda g_m(\theta)\big(\psi(\theta) - \nu\big) \end{aligned} \tag{11}$$

Summing (11) across θ and imposing $\sum_\theta x^*(\theta) = 0$ together with $\sum_\theta g_m(\theta) = 1$, I get:

$$\nu = \sum_\theta g_m(\theta)\psi(\theta) =: \mathbb{E}_{g_m}[\psi] \tag{12}$$

Finally substituting ν back into (11) gives the optimal distortion $x^*(\theta)$, and therefore the optimal belief $\gamma^*(\theta) = g_m(\theta) + x^*(\theta)$:

$$\begin{aligned} \gamma^*(\theta) &= g_m(\theta) + x^*(\theta) \\ &= g_m(\theta) + \lambda\, g_m(\theta)\Big(\psi(\theta) - \mathbb{E}_{g_m}[\psi]\Big) \\ &= g_m(\theta) + \lambda\, \mathrm{Cov}_{g_m}\big(\mathbb{1}_{\{\theta\}},\, \psi\big) \end{aligned} \tag{4}$$

In the optimal belief, probability moves toward state θ in proportion to how much better the motive rewards θ than it rewards the average state under the reference belief g_m, $\psi(\theta) - \mathbb{E}_{g_m}[\psi]$, scaled by $g_m(\theta)$ itself: a reference belief that already puts little weight on θ has little room to inflate it further, and one certain of some state has no room to move at all. Equivalently, the belief moves in the direction of, and in proportion to, the covariance between believing θ and the motive's reward ψ.

A.2 The indirect value of a mental representation

Having solved the belief-choice subproblem, I can fold its solution back into $U\big(\gamma, m\big)$ to obtain $V(m)$. Retaining the two approximations from Section A.1 I obtain:

$$V(m) = v(g_m) + \sum_\theta \psi(\theta)\, x^*(\theta) \;-\; \frac{1}{2\lambda} \sum_\theta \frac{x^*(\theta)^2}{g_m(\theta)}. \tag{13}$$

The first term is the motive-assessed value of the Bayesian beliefs implied by the model: a mental representation whose implied Bayesian belief already places weight where the motive wants it is more attractive to the DM before any distortion enters. The second and third are the motivated gain and the divergence cost of the optimal motivated spin: in addition to the primary motivated value, the DM gains an additional bonus for holding a mental representation that also leaves room for profitable spins.

Plugging in the optimal distortion $x^*(\theta) = \lambda\, g_m(\theta)\big(\psi(\theta) - \mathbb{E}_{g_m}[\psi]\big)$ from (4), the gain term collapses to a variance:

$$\sum_\theta \psi(\theta)\, x^*(\theta) = \lambda \sum_\theta g_m(\theta)\, \psi(\theta)\big(\psi(\theta) - \mathbb{E}_{g_m}[\psi]\big) = \lambda\big(\mathbb{E}_{g_m}[\psi^2] - \mathbb{E}_{g_m}[\psi]^2\big) = \lambda\, \mathrm{Var}_{g_m}(\psi), \quad (14)$$

and the divergence cost is exactly half of it:

$$\frac{1}{2\lambda} \sum_\theta \frac{x^*(\theta)^2}{g_m(\theta)} = \frac{\lambda}{2} \sum_\theta g_m(\theta)\big(\psi(\theta) - \mathbb{E}_{g_m}[\psi]\big)^2 = \frac{\lambda}{2}\, \mathrm{Var}_{g_m}(\psi). \quad (15)$$

The DM thus keeps half the gross motivated gain, and $V(m)$ collapses to:

$$\begin{aligned} V(m) &= v(g_m) + \frac{\lambda}{2} \sum_\theta g_m(\theta)\Big(\psi(\theta) - \mathbb{E}_{g_m}[\psi]\Big)^2 \\ &= v(g_m) + \frac{\lambda}{2}\, \mathrm{Var}_{g_m}(\psi). \end{aligned} \quad (5)$$

The bonus gained in addition to the primary motivated value of having a mental representation that implies a Bayesian belief g_m is increasing in $\mathrm{Var}_{g_m}(\psi)$: the more g_m spreads probability mass across states with different values of ψ, the more room the DM has to profitably spin their belief within the model.[29] At the other extreme, a mental representation that places all its weight on a single state has $\mathrm{Var}_{g_m}(\psi) = 0$ and earns no bonus, since there is no room left to spin.

Formally, let g' be a mean-preserving spread of g in ψ: $\mathbb{E}_g[\psi] = \mathbb{E}_{g'}[\psi]$, but g' shifts weight toward more extreme values of ψ. Then $\mathrm{Var}_{g'}(\psi) \geq \mathrm{Var}_g(\psi)$, so $V(g') \geq V(g)$ whenever $v(g') = v(g)$: two mental representations that the motive rates equally on their honest merits can still differ in how attractive they are to the DM, purely because one leaves more room for motivated spin than the other.

B Questions in the Correlation-Neglect Study of Exley and Kessler (2024)

Table 7 reproduces Table 6 of Exley and Kessler (2024). For each of the ten questions, it reports the four estimates, the three news-channel reports participants observe, the correct answer, and the naive answer. Participants observe Estimate 1 and the three news-channel reports, and see the ten questions in random order.

[29] Since $\mathrm{Var}_{g_m}(\psi) = \mathrm{Cov}_{g_m}(\psi, \psi)$, this bonus is the same covariance object that governs the optimal spin in (4), now instantiated with ψ against itself instead of against the indicator $\mathbb{1}_{\{\theta\}}$.

Table 7: Questions in the correlation-neglect study of Exley and Kessler (2024)

			Estimate				News channel		
Q	Correct	Naive	1	2	3	4	1	2	3
1	50	29	8	40	59	93	24	33.5	50.5
2	50	71.5	93	8	40	59	50.5	66.5	76
3	47	24.5	2	43	50	93	22.5	26	47.5
4	47	70	93	2	43	50	47.5	68	71.5
5	54	29	4	59	55	98	31.5	29.5	51
6	54	76	98	4	59	55	51	78.5	76.5
7	49	27	5	51	48	92	28	26.5	48.5
8	49	70.5	92	5	51	48	48.5	71.5	70
9	52	27.5	3	52	54	99	27.5	28.5	51
10	52	75.5	99	3	52	54	51	75.5	76.5

Notes: Reproduced from Table 6 of Exley and Kessler (2024). Each news channel reports the average of Estimate 1 and one of the other estimates (channel j averages Estimates 1 and $j+1$). The correct answer is the average of the four estimates. The naive answer ("correlated answer" in the original) is the average of Estimate 1 and the three news-channel reports, i.e. the answer of a respondent who treats them as independent signals. Questions come in pairs with the same four estimates: in odd-numbered questions Estimate 1 is the lowest estimate, so correlation neglect pulls the answer down; in even-numbered questions it is the highest, so correlation neglect pulls the answer up.

C Re-analysis of the Correlation-Neglect Study of Exley and Kessler (2024): Robustness

This appendix probes the classification of Section 3.2 along four dimensions: how *No motive* answers are aggregated (Appendix C.1), the dependent variable (Appendix C.2), the cutoffs that define the types (Appendix C.3), and the sample (Appendix C.4). Appendix C.5 summarises.

C.1 Aggregation of *No motive* answers

In the main analysis of Section 3.2, each *No motive* participant contributes a single median, computed over all ten questions, whereas motive-treatment participants contribute one median per condition, computed over five questions. Table 8 repeats the analysis with each *No motive* participant split into two units, one per half of their questions, so that every unit rests on five answers.

A median over five answers is noisier than one over ten, so the share of other errors in the bias-only condition doubles, from 5.4% to 11.9%, and the marginal effects shrink to about half their size. The direction of every effect is unchanged except one: in *Opposed*, the share of rational respondents no longer falls but rises by an insignificant 2.0 percentage points. The main conclusion holds: in both motive conditions, other errors rise significantly (by 12.9 and 15.3 percentage points) at the expense of the representation the motive disfavours, and the representation the motive favours does

Table 8: Representation adopted by condition: *No motive* participants split into halves

	Rational	Naive	Other errors
Aligned	-0.039*	-0.090***	0.129***
	(0.022)	(0.025)	(0.021)
Opposed	0.020	-0.173***	0.153***
	(0.021)	(0.024)	(0.021)
Share in *Heuristic only*	0.267	0.614	0.119
Units		2,400	

Notes: As Table 1, except that each *No motive* participant contributes two units, one per half of their questions (median over five questions each), so that every unit rests on five answers. Average marginal effects from a multinomial logit of type membership on indicators for *Aligned* and *Opposed*; the reference condition is *Heuristic only*. Types are based on the median naïveté parameter χ: rational if $\chi \in [-0.33, 0.33]$, naive if $\chi \in (0.33, 1.66)$, other errors otherwise. Standard errors in parentheses. Data from Exley and Kessler (2024). $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

not gain.

C.2 Alternative dependent variables

I repeat the analyses of Section 3.2 with three alternatives to the respondent-level median of χ. The first is the respondent-level *mean* of χ, which uses every answer rather than the central one. The second is χ for each individual answer, pooling all rounds. The third is the *directional error*, the error $a - c$ signed so that positive values point in the direction in which correlation neglect pulls the answer, used at the answer level and as the respondent-level mean and median. Unlike χ, the directional error is not normalised by the gap between the naive and the correct answer. In this design, however, that gap varies little across questions, between 21 and 25 points, so the directional error is close to a rescaling of χ by a factor of about 23. Because it has no benchmarks of its own, and classifying it relative to each question's naive answer would reproduce the classification by χ, I do not classify respondents by their directional error; I report its mean by condition.

Types. Tables 9 and 10 repeat the classification and the multinomial logit of Table 1 with the mean of χ and with χ of each answer. The answer-level model pools the ten answers of every participant, adds question-pair fixed effects, and clusters standard errors by participant.

With the mean of χ, the results are close to those of the main specification. In *Aligned*, the shares of rational and naive respondents fall by 14.8 and 11.6 percentage points, and the share of other errors rises by 26.5. In *Opposed*, they fall by 9.6 and 20.5 percentage points, and the share of other errors rises by 30.1 (all $p < 0.01$). The answer-level classification resembles the halves specification of Table 8: single answers are noisier than medians, so 15.8% of bias-only answers already count

Table 9: Representation adopted by condition: mean naïveté parameter

	Rational	Naive	Other errors
Aligned	-0.148***	-0.116***	0.265***
	(0.028)	(0.034)	(0.038)
Opposed	-0.096***	-0.205***	0.301***
	(0.027)	(0.033)	(0.037)
Share in *Heuristic only*	0.370	0.583	0.047
Units		1,992	

Notes: As Table 1, except that types are based on the respondent-level *mean* of χ instead of the median: rational if $\chi \in [-0.33, 0.33]$, naive if $\chi \in (0.33, 1.66)$, other errors otherwise. Standard errors in parentheses. Data from Exley and Kessler (2024). $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 10: Representation adopted by condition: individual answers

	Rational	Naive	Other errors
Aligned	-0.043**	-0.082***	0.125***
	(0.021)	(0.022)	(0.019)
Opposed	0.002	-0.145***	0.143***
	(0.020)	(0.022)	(0.019)
Share in *Heuristic only*	0.272	0.570	0.158
Answers		12,000	

Notes: As Table 1, except that each answer is classified separately by its own χ: rational if $\chi \in [-0.33, 0.33]$, naive if $\chi \in (0.33, 1.66)$, other errors otherwise. The multinomial logit pools all ten answers of every participant and includes question-pair fixed effects. Standard errors clustered by participant in parentheses. Data from Exley and Kessler (2024). $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

as other errors, and the effects are about half as large. Other errors rise by 12.5 percentage points in *Aligned* and 14.3 in *Opposed*. Naive answers fall in both conditions (by 8.2 and 14.5 percentage points), and rational answers fall in *Aligned* (by 4.3 percentage points, $p < 0.05$) and do not change in *Opposed* (+0.2 percentage points, $p = 0.92$).

Averages. Table 11 regresses each variable on the condition indicators. Every specification reproduces the pattern of average errors documented by Exley and Kessler (2024): relative to the bias-only condition, the average error in the direction of correlation neglect is larger in *Aligned* and smaller in *Opposed*. For instance, the median χ rises by 0.09 in *Aligned* ($p < 0.05$) and falls by 0.29 in *Opposed* ($p < 0.01$), from 0.47 in the bias-only condition. The coefficients of the mean and of the answer-level specifications coincide because every participant answers the same number of questions in each condition. These shifts in averages are what Section 3.2 decomposes: they arise even though the motive does not move respondents towards the representation it favours.

Table 11: Naiveté and directional error by condition

	Naïveté parameter χ			**Directional error**		
	Median	**Mean**	**Answer**	**Median**	**Mean**	**Answer**
Constant	0.472***	0.431***	0.416***	10.663***	9.752***	8.946***
	(0.024)	(0.023)	(0.024)	(0.555)	(0.513)	(0.533)
Aligned	0.094**	0.118***	0.118***	2.045**	2.631***	2.631***
	(0.041)	(0.038)	(0.038)	(0.915)	(0.867)	(0.867)
Opposed	-0.287***	-0.262***	-0.262***	-6.491***	-5.889***	-5.889***
	(0.042)	(0.039)	(0.039)	(0.943)	(0.884)	(0.884)
Question-pair FE	No	No	Yes	No	No	Yes
N	1992	1992	12000	1992	1992	12000
R^2	0.039	0.042	0.029	0.039	0.042	0.029

Notes: OLS of each dependent variable on indicators for *Aligned* and *Opposed*; the constant is the predicted value in the *Heuristic only* condition. *Median* and *Mean*: respondent-level median and mean over the questions a participant answers in a condition (one unit per *No motive* participant, two per motive-treatment participant). *Answer*: one observation per answer, with question-pair fixed effects. The directional error is $a - c$, signed so that positive values point in the direction in which correlation neglect pulls the answer, in points of the 0–100 scale. Standard errors clustered by participant in parentheses. Data from Exley and Kessler (2024). $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

C.3 Classification cutoffs

The bands that define the types, $[-0.33, 0.33]$ for rational and $(0.33, 1.66)$ for naive, are a choice. Table 12 repeats the main specification with a narrower ($[-0.25, 0.25]$) and a wider ($[-0.50, 0.50]$) rational band, and with a lower (1.50) and a higher (2.00) upper bound of the naive band. Wider

bands mechanically shift the baseline shares, but the marginal effects barely move. In every specification and both motive conditions, the shares of rational and naive respondents both fall, and the share of other errors rises by between 22 and 33 percentage points (all $p < 0.01$). The smallest rise in other errors comes with the widest naive band, which reclassifies some of the answers that overshoot the naive benchmark as naive; even then, the share of naive respondents falls in both conditions.

Table 12: Representation adopted by condition: alternative classification cutoffs

	Rational	Naive	Other errors
Rational band $[-0.25, 0.25]$, naive band $(0.25, 1.66)$			
Aligned	-0.119***	-0.141***	0.260***
	(0.026)	(0.034)	(0.036)
Opposed	-0.081***	-0.208***	0.289***
	(0.025)	(0.033)	(0.035)
Share in *Heuristic only*	0.311	0.627	0.061
Rational band $[-0.50, 0.50]$, naive band $(0.50, 1.66)$			
Aligned	-0.185***	-0.093***	0.278***
	(0.030)	(0.033)	(0.038)
Opposed	-0.085***	-0.198***	0.283***
	(0.030)	(0.033)	(0.038)
Share in *Heuristic only*	0.431	0.527	0.042
Rational band $[-0.33, 0.33]$, naive band $(0.33, 1.50)$			
Aligned	-0.133***	-0.184***	0.317***
	(0.028)	(0.034)	(0.039)
Opposed	-0.074***	-0.252***	0.325***
	(0.027)	(0.033)	(0.038)
Share in *Heuristic only*	0.328	0.618	0.054
Rational band $[-0.33, 0.33]$, naive band $(0.33, 2.00)$			
Aligned	-0.120***	-0.101***	0.221***
	(0.027)	(0.034)	(0.036)
Opposed	-0.059**	-0.207***	0.266***
	(0.026)	(0.033)	(0.035)
Share in *Heuristic only*	0.328	0.625	0.047

Notes: As Table 1, with the bands of the respondent-level median χ that define the rational and the naive type given in each panel title; other values are classified as other errors. The main specification uses $[-0.33, 0.33]$ and $(0.33, 1.66)$. Standard errors in parentheses. Data from Exley and Kessler (2024). $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

C.4 Participants with a failed calibration

The motive treatments rely on the calibration: X is meant to make a bonus of X cents worth as much to the participant as 150 cents for the charity. For participants whose calibration has a lower bound of 0, X is set to 5 cents. They prefer even a minimal bonus to the donation, so for them the calibration cannot equalise the two payments and the motive treatments do not hold stakes constant. In total, 200 participants have $X = 5$, most of them with a lower bound of 0 and a few with a lower bound of 5 cents; I exclude all of them. Tables 13 and 14 repeat the classification and the absolute-error regression without them.

Table 13: Representation adopted by condition: excluding failed calibrations

	Rational	Naive	Other errors
Aligned	-0.113***	-0.130***	0.242***
	(0.030)	(0.036)	(0.039)
Opposed	-0.074**	-0.189***	0.263***
	(0.030)	(0.036)	(0.038)
Share in *Heuristic only*	0.342	0.606	0.052
Units		1,655	

Notes: As Table 1, on a restricted sample. Excludes the 200 participants whose charity equivalent X was set to 5 cents (calibration lower bound of 0 or 5 cents), leaving 1,000 participants. Standard errors in parentheses. Data from Exley and Kessler (2024). $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 14: Absolute errors by condition: excluding failed calibrations

	Absolute error
Heuristic only (constant)	15.856***
	(0.495)
Aligned	3.184***
	(0.679)
Opposed	1.512**
	(0.663)
Question-pair fixed effects	Yes
Observations	10,000
R^2	0.010

Notes: As Table 2, on a restricted sample. Excludes the 200 participants whose charity equivalent X was set to 5 cents (calibration lower bound of 0 or 5 cents), leaving 1,000 participants. Standard errors clustered by participant in parentheses. Data from Exley and Kessler (2024). $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Excluding these participants shrinks the effects slightly but leaves the conclusions intact. Other

errors rise by 24.2 percentage points in *Aligned* and 26.3 in *Opposed*, and the shares of rational and naive respondents fall in both conditions (all $p < 0.05$). Absolute errors still rise in both motive conditions, by 3.2 points in *Aligned* ($p < 0.01$) and 1.5 points in *Opposed* ($p < 0.05$).

C.5 Summary

The main conclusions of Section 3 hold in every specification. In both motive conditions, the share of answers that neither representation implies rises significantly, the representation the motive disfavours loses ground (in the halves specification, the fall in the rational share in *Aligned* is significant only at the 10% level), and the representation the motive favours does not gain: the naive share never rises in *Aligned*, and the rational share never rises significantly in *Opposed*. Absolute errors rise in both conditions. One result of the main specification is not robust: the fall in the rational share in *Opposed* disappears when each unit rests on at most five answers (the halves and the answer-level specifications), where this share is unchanged.

D Charity Equivalent

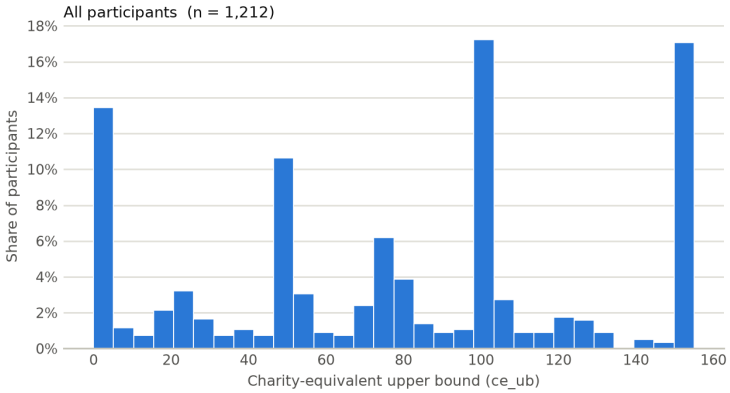


Figure 4: Distribution of elicited charity-equivalent upper bounds (*ce_ub*) across all subjects.

E Estimation Rounds in CN×ME

Table 15 reports, for each of the six rounds of CN×ME, the four estimates and the information participants observe. In *Control* and *Motive only*, participants observe the four estimates directly. In *Heuristic only* and *Motive×Heuristic*, they observe one estimate directly and three news-channel

reports; the observed estimate is the lowest of the four for participants for whom correlation neglect pulls answers down, and the highest for those for whom it pulls answers up. The estimates underlying the six rounds are the same for every participant and are shown in random order.

Table 15: Estimation rounds in CN×ME

				Estimates				Correlated information	
Round	CN pulls	Correct	Naive	1	2	3	4	Observed	News channels 1–3
1	down	31	25	19	25	37	43	19	22, 28, 31
1	up	31	37	19	25	37	43	43	31, 34, 40
2	down	45	35.75	27	39	48	66	27	33, 37, 46
2	up	45	55.25	27	39	48	66	66	46, 52, 57
3	down	55	42.75	31	52	60	77	31	41, 45, 54
3	up	55	65.75	31	52	60	77	77	54, 64, 68
4	down	65	53.75	43	58	71	88	43	50, 57, 65
4	up	65	76.25	43	58	71	88	88	65, 73, 79
5	down	73	63.25	54	63	79	96	54	58, 66, 75
5	up	73	84.25	54	63	79	96	96	75, 79, 87
6	down	80	70.75	62	74	85	99	62	68, 73, 80
6	up	80	89.25	62	74	85	99	99	80, 86, 92

Notes: The correct answer is the average of the four estimates. Each news channel reports the average of the observed estimate and one of the other three estimates, rounded down to the nearest integer. The naive answer is the average of the observed estimate and the three news-channel reports, i.e. the answer of a respondent who treats them as independent signals. Without rounding, the naive answer equals the average of the observed estimate and the correct answer.

F Robustness of the CN×ME Results

F.1 Ordering and additivity tests

Table 16 reports the linear combinations of the coefficients in Table 3 that Section 5.3 discusses: the ordering contrasts of Exley and Kessler (2024), which compare *Aligned* and *Opposed* with *Heuristic only*, and the pre-registered additivity contrasts.

F.2 Participants with a zero charity equivalent

The 163 participants whose charity equivalent is zero prefer even the smallest bonus to the donation, so the calibration cannot make the two payments equally valuable to them: they face not only a motive, but also a small material incentive to err in the direction that secures the bonus. Excluding them ($N = 1{,}049$), the excess of *Aligned* over the additive benchmark rises from 0.13 to 0.24 and becomes significant ($p < 0.01$). This, however, is mostly because the benchmark itself shrinks: in this sample, the effect of the motive in *Motive only* falls from 0.12 to 0.05 (SE 0.05) and is no longer

Table 16: Ordering and additivity tests

	Contrast
Ordering	
Aligned − *Heuristic only*	0.250***
	(0.077) [0.001]
Heuristic only − *Opposed*	0.028
	(0.078) [0.720]
Additivity	
Aligned − (*Heuristic only* + *Motive only*)	0.128
	(0.093) [0.171]
Opposed − (*Heuristic only* − *Motive only*)	0.094
	(0.095) [0.321]

Notes: Linear combinations of the coefficients in column (1) of Table 3 (each row states the contrast). Under additivity, both additivity contrasts are zero. Heteroskedasticity-robust (HC3) standard errors in parentheses, two-sided p-values in brackets. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

significant, while the excess in *Opposed* grows to 0.15 ($p = 0.11$), again in the direction opposite to Prediction 3b. Because the significance of the *Aligned* test hinges on a sample in which the motive has no detectable effect on its own, I do not read it as evidence of an interaction.

F.3 Alternative classifications

Table 17 repeats the classification of Section 5.3 under alternative definitions of the types. The first two rows are pre-registered: the bands of the main text, and bands set empirically from the *Control* and *Heuristic only* distributions. The others are exploratory: tight bands around the rational and the naive benchmark, with partial neglect as a separate class; a classification that requires a respondent's answers to fall in the same class in at least four of the six rounds; and symmetric windows of half-width h around the two benchmarks. In no classification is the share of naive respondents significantly higher in *Aligned*, or the share of rational respondents significantly higher in *Opposed*, than in *Heuristic only*.

F.4 Motive strength

Table 18 reports the slope of the median χ on the charity equivalent X in each condition, from the pre-registered regression that interacts X with the condition indicators. A lower X means that the participant cares less about the donation relative to the bonus, and hence has a stronger reason to err in the direction that secures the bonus. Relative to *Control*, the slope is significantly different in *Motive only* (-0.32, $p < 0.01$) and in *Opposed* ($+0.51$, $p < 0.01$), but not in *Heuristic only* (-0.14, $p = 0.28$) or *Aligned* ($+0.05$, $p = 0.77$).

Table 17: Representation adopted under alternative classifications

	Aligned vs *Heuristic only*			*Opposed* vs *Heuristic only*		
Classification	Rational	Naive	Joint p	Rational	Naive	Joint p
Pre-registered						
$[-\frac{1}{3}, \frac{1}{3}]$, $(\frac{1}{3}, \frac{5}{3})$	-0.070	0.054	0.283	-0.037	0.002	0.337
	(0.044)	(0.044)		(0.044)	(0.043)	
Empirical boundaries	-0.051	0.019	0.373	-0.024	-0.024	0.226
	(0.045)	(0.044)		(0.044)	(0.042)	
Tight bands, partial neglect separate						
$\lvert x\rvert \leq 0.05$, $\lvert x-1\rvert \leq 0.1$	-0.089**	0.013	0.209	-0.057	0.019	0.615
	(0.044)	(0.037)		(0.043)	(0.037)	
Round-level consistency						
≥ 4 of 6 rounds	-0.078*	0.033	0.201	-0.066	0.003	0.185
	(0.044)	(0.044)		(0.043)	(0.043)	
Symmetric windows $\lvert x\rvert \leq h$, $\lvert x-1\rvert \leq h$						
$h = 0.05$	-0.089**	-0.022	0.042	-0.057	0.018	0.410
	(0.044)	(0.033)		(0.043)	(0.034)	
$h = 0.1$	-0.081*	0.013	0.157	-0.059	0.019	0.392
	(0.044)	(0.037)		(0.043)	(0.037)	
$h = 0.2$	-0.079*	0.008	0.129	-0.058	0.005	0.308
	(0.044)	(0.040)		(0.043)	(0.039)	
$h = 1/3$	-0.070	0.046	0.293	-0.037	0.016	0.683
	(0.044)	(0.042)		(0.044)	(0.041)	
$h = 0.5$	-0.060	0.043	0.397	-0.002	-0.025	0.577
	(0.044)	(0.044)		(0.044)	(0.043)	

Notes: Each row is a multinomial logit of type membership on indicators for *Aligned* and *Opposed*, estimated on these and *Heuristic only*, under a different classification of the median naïveté parameter x (0 = rational, 1 = naive). Entries are changes in the share of rational and naive respondents relative to *Heuristic only*; joint p is the Wald test that the whole type distribution equals that in *Heuristic only*. The first two rows are pre-registered; the others are exploratory. Delta-method standard errors in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 18: Motive strength: slope of the median naïveté on X by condition

	Slope on X
Control	0.008
	(0.068)
Motive only	-0.307***
	(0.098)
Heuristic only	-0.130
	(0.107)
Aligned	0.055
	(0.143)
Opposed	0.517***
	(0.154)
Observations	1,212

Notes: Slopes of the median naïveté parameter χ on the charity equivalent X (in dollars) in each condition, from an OLS regression of χ on the condition indicators, X, and their interactions. Heteroskedasticity-robust (HC3) standard errors in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

G The Second Estimate in SN×IQ

G.1 Design of the second estimate

After the first estimate, participants make a second one about the other target: *Own-first* participants estimate the other participant's score, and *Other-first* participants their own. The selection condition is the same as in the first estimate, and the reports are built in the same way from the new target's score. The second estimate differs from the first in four respects at once: the target, the experience of having already completed the task once, the absence of an announcement, and the absence of an accuracy bonus. No comparison involving the second estimate isolates the effect of the target, so all the analyses in this appendix are exploratory. The first two (Sections G.2 and G.3) are pre-registered; the others are not.

Because the order of the two targets is randomised, the second estimate does allow one comparison that the first does not: the same participant estimates both targets. When a behaviour moves in opposite directions in the two orders, increasing when the target switches to one's own score and decreasing when it switches away from it, the change follows the target rather than the round. A difference between rounds that is common to both orders, such as learning or the absence of the bonus, would move both orders in the same direction.

G.2 Pre-registered: estimates of the other participant's score

The pre-registration conjectures that participants keep the representation they adopted for their own score when they then estimate the other participant's score, since they no longer have a motive to change it. I compare the estimates of the other participant's score made by *Own-first* participants in the second round with those made by *Other-first* participants in the first round, by regressing the error on PriorOwn (one for *Own-first* participants), the selection dummies and their interactions:

$$\text{error}_i = \gamma_0 + \gamma_1 \text{PriorOwn}_i + \gamma_2 \text{Sel}+_i + \gamma_3 \text{Sel}-_i + \gamma_4 \text{PriorOwn}_i \times \text{Sel}+_i + \gamma_5 \text{PriorOwn}_i \times \text{Sel}-_i + u_i.$$

The pre-registered predictions are $\gamma_4 > 0$ and $\gamma_5 > 0$, to be read as carry-over only if the motive changed the representation in the first estimate (Predictions 3a and 3b). Since PriorOwn $= 0$ selects first-round estimates of *Other-first* participants, γ_0, γ_2 and γ_3 coincide with the corresponding coefficients of the main specification in Section 6.3.1.

Table 19 reports the estimates. Neither interaction has the predicted sign and is significant: $\gamma_4 = -2.0$ (SE 2.5, $p = 0.42$) and $\gamma_5 = 1.5$ (SE 2.6, $p = 0.56$). Since neither Prediction 3a nor Prediction 3b holds in the first estimate, these coefficients are not interpreted as carry-over.

Table 19: Round 2 across participants: estimates of the other participant's score

	Estimate	SE	p
Constant (γ_0)	4.14^{***}	(1.28)	[0.001]
PriorOwn (γ_1)	0.26	(1.82)	[0.889]
Selection+ (γ_2)	7.02^{***}	(1.71)	[<0.001]
Selection− (γ_3)	-7.94^{***}	(1.87)	[<0.001]
PriorOwn × *Selection+* (γ_4)	−1.99	(2.48)	[0.422]
PriorOwn × *Selection−* (γ_5)	1.52	(2.62)	[0.561]
N	762		
R^2	0.129		

Notes: OLS of the error of the estimate about the other participant's score (in percentage points) on PriorOwn, the selection dummies and their interactions. For *Own-first* participants (PriorOwn = 1) the estimate is the round-2 estimate, made after estimating their own score; for *Other-first* participants it is the round-1 estimate. Pre-registered predictions: $\gamma_4 > 0$, $\gamma_5 > 0$. Heteroskedasticity-robust (HC3) standard errors in parentheses, two-sided p-values in brackets. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

G.3 Pre-registered: change in error between the two estimates

Let Δ_i be the error of participant i's second estimate minus the error of their first. I regress Δ_i on the six cell dummies. Since Δ_i also reflects the switch of target, each selection cell is compared with the *No selection* cell of the same order. The pre-registration predicts that Δ is larger in *Other-first* × *Selection−* than in *Other-first* × *No selection*: participants who used the naive representation

for the other participant's score may notice the mistake once it lowers the estimate of their own.

Table 20 reports the results. The predicted difference is positive but small and not significant (2.1 percentage points, SE 2.1, $p = 0.31$), and none of the other three differences is significant either. Mean errors change by less than three percentage points between rounds in every cell.

Table 20: Round 2 within participants: change in error

	Error R1	Error R2	Δ	SE	N
Panel A: change in error by cell					
Other-first, *No selection*	4.1	4.3	0.1	(1.5)	104
Other-first, *Selection+*	11.2	9.1	−2.1	(1.6)	129
Other-first, *Selection−*	−3.8	−1.5	2.3	(1.5)	129
Own-first, *No selection*	6.1	4.4	−1.7	(1.5)	139
Own-first, *Selection+*	12.2	9.4	−2.8	(1.4)	138
Own-first, *Selection−*	−1.3	−2.0	−0.7	(1.3)	123
Panel B: difference in Δ					
		Estimate	SE	p	
Other-first: *Selection− − No selection*[†]		2.1	(2.1)	[0.313]	
Other-first: *Selection+ − No selection*		−2.2	(2.2)	[0.312]	
Own-first: *Selection− − No selection*		1.0	(1.9)	[0.622]	
Own-first: *Selection+ − No selection*		−1.1	(2.0)	[0.576]	

Notes: Δ = round-2 error − round-1 error, in percentage points. Rows are cells by order of the targets and selection condition; the target switches between rounds and the selection condition stays the same. Panel A: mean errors and the coefficients of an OLS regression of Δ on the six cell dummies without a constant. Panel B: differences between each selection cell and the *No selection* cell of the same order. [†]Pre-registered, with the prediction that the difference is positive. Heteroskedasticity-robust (HC3) standard errors in parentheses, two-sided p-values in brackets. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

G.4 Exploratory: correct and fully naive estimates

Table 21 classifies both estimates of each participant as in Section 6.3.2, and Table 22 reports how participants move between classes in the cells with selection.

The share of fully naive estimates barely changes between rounds: the change is below 2.5 percentage points in every cell, and in every selection cell about half of the participants whose first estimate is fully naive give a fully naive second estimate (between 10 of 22 and 24 of 38). The rate is similar whether the target switches to one's own score or away from it. If participants keep their representation across targets, they do so equally often when the motive switches on and when it switches off.

The share of correct estimates, instead, follows the target. It falls in every *Other-first* cell, where the second estimate concerns one's own score (by 2.3 to 13.5 percentage points), and rises in every *Own-first* cell, where it concerns the other participant's (by 8.7 to 16.3 percentage points). In *Opposed*, the share rises from 17.1% to 33.3% ($p < 0.01$, exact McNemar test): 30 of the 102

participants whose first estimate is not correct give a correct estimate of the other participant's score. The effect of the target on the share of correct estimates is therefore not specific to *Opposed*: within participants, it is as large without selection, and the differences between each selection cell and the *No selection* cell of the same order are not significant at the 5% level (bottom rows of Table 21). This is consistent with the first-round comparison between *Motive only* and *Control* in Section 6.3.1, where estimates of one's own score are less often correct: the motive makes estimates less precise whether or not selection neglect is available.

Table 21: Round 2 within participants: correct and fully naive estimates

	Correct (%)				Fully naive (%)			
	R1	R2	Diff.	p	R1	R2	Diff.	p
Other-first, *No selection*	47.1	33.7	-13.5^{**}	0.013	–	–	–	–
Other-first, *Selection+*	18.6	16.3	-2.3	0.648	29.5	27.1	-2.3	0.690
Other-first, *Selection−*	28.7	21.7	-7.0	0.163	17.1	18.6	1.6	0.845
Own-first, *No selection*	35.3	46.8	11.5^{**}	0.020	–	–	–	–
Own-first, *Selection+*	15.9	24.6	8.7^{*}	0.050	26.8	28.3	1.4	0.868
Own-first, *Selection−*	17.1	33.3	16.3^{***}	0.002	15.4	14.6	-0.8	1.000
Difference in the change of the correct share								
Other-first: *Selection− − No selection*			6.5 (6.7)	0.331				
Other-first: *Selection+ − No selection*			11.1^{*} (6.0)	0.064				
Own-first: *Selection− − No selection*			4.7 (6.8)	0.483				
Own-first: *Selection+ − No selection*			-2.8 (6.1)	0.646				

Notes: Share of estimates classified as correct (within 2.5 percentage points of the target score) and fully naive (within 2.5 percentage points of the average of the reports shown) in round 1 (R1) and round 2 (R2), by cell. The target switches between rounds: *Other-first* participants estimate their own score in round 2, *Own-first* participants the other participant's. Diff.: R2 − R1, in percentage points; p: exact McNemar test. Without selection, estimates cannot be fully naive. Bottom rows: difference between each selection cell and the *No selection* cell of the same order in the change of the correct share (OLS of the change in the correct indicator on the cell dummies, HC3 standard errors in parentheses). $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

G.5 Summary

The second estimate confirms the first-round results. The pre-registered round-2 tests on mean errors detect neither carry-over of the representation nor the predicted correction. Switching the target to one's own score leaves the share of fully naive estimates unchanged and lowers the share of correct ones, with and without selection.

H Low Scorers in SN×IQ

Section 6.3.1 shows that, without selection, the motive does not shift the average estimate but makes estimates more dispersed. This appendix asks whether the effects of the motive and of the

Table 22: Round 2 within participants: transitions between types

	Type in round 2				
Type in round 1	Correct	Partially naive	Fully naive	Other	Total
Other-first, *Selection+*					
Correct	13	1	2	8	24
Partially naive	2	8	5	15	30
Fully naive	1	7	24	6	38
Other	5	9	4	19	37
Other-first, *Selection−*					
Correct	16	4	1	16	37
Partially naive	6	3	5	9	23
Fully naive	1	2	10	9	22
Other	5	6	8	28	47
Own-first, *Selection+*					
Correct	12	5	3	2	22
Partially naive	8	6	9	12	35
Fully naive	2	11	20	4	37
Other	12	9	7	16	44
Own-first, *Selection−*					
Correct	11	2	1	7	21
Partially naive	10	11	3	5	29
Fully naive	5	0	10	4	19
Other	15	8	4	27	54

Notes: Number of participants by the type of their round-1 estimate (rows) and of their round-2 estimate (columns), for the cells with selection. Types as in Table 21; partially naive: strictly between the correct and the fully naive benchmark; other: anything else. The target switches between rounds.

heuristic differ between participants who scored low and participants who scored high on the test. The analysis is exploratory and was not pre-registered.

Let $\text{Low}_i = 1$ if participant i's own score is below 50%, the sample median, and $\text{Low}_i = 0$ otherwise; the 15% of participants who scored exactly 50% have $\text{Low}_i = 0$. This gives 306 low scorers and 456 other participants. I interact every regressor of the main specification with Low:

$$\begin{aligned}\text{error}_i = {} & \beta_0 + \beta_1 \text{Own}_i + \beta_2 \text{Sel}+_i + \beta_3 \text{Sel}-_i + \beta_4 \text{Own}_i \times \text{Sel}+_i + \beta_5 \text{Own}_i \times \text{Sel}-_i \\ & + \text{Low}_i \times (\delta_0 + \delta_1 \text{Own}_i + \delta_2 \text{Sel}+_i + \delta_3 \text{Sel}-_i + \delta_4 \text{Own}_i \times \text{Sel}+_i + \delta_5 \text{Own}_i \times \text{Sel}-_i) + \varepsilon_i,\end{aligned}$$

so that β_0–β_5 are the coefficients of the main specification for participants with Low = 0, and δ_0–δ_5 are the differences for low scorers. I also estimate the same model with the absolute error as the dependent variable. Table 23 reports both.

Taken together, the low-scorer terms are not significant for the error ($p = 0.23$, joint F-test of $\delta_0 = \cdots = \delta_5 = 0$), and they are significant for the absolute error ($p = 0.03$). The heuristic works in the same way for both groups: neither δ_2 nor δ_3 is significant, so *Selection+* raises and *Selection−* lowers estimates by similar amounts among low scorers and the other participants.

The motive, instead, differs between the two groups without selection. Among participants with Low = 0, estimates of one's own score are not higher than estimates of the other participant's score ($\beta_1 = -2.4$, SE 2.4, $p = 0.32$). Among low scorers, they are 8.2 percentage points higher ($\beta_1 + \delta_1$, SE 3.6, $p = 0.02$), and the difference between the groups is significant ($\delta_1 = 10.6$, SE 4.3, $p = 0.01$). With selection, the difference disappears. The effect of the motive among low scorers is 1.0 percentage points with *Selection+* (SE 3.4, $p = 0.78$) and 2.2 percentage points with *Selection−* (SE 3.7, $p = 0.56$), close to the effects among the other participants (1.2 and 2.4 percentage points). Among low scorers, the departures from additivity are therefore negative but not significant: $\beta_4 + \delta_4 = -7.3$ (SE 4.9, $p = 0.14$) in *Aligned* and $\beta_5 + \delta_5 = -6.1$ (SE 5.1, $p = 0.24$) in *Opposed*. Relative to the other participants, both departures are 10.8 percentage points lower, a difference that is only marginally significant (δ_4 and δ_5, $p = 0.06$ and $p = 0.08$).

The absolute error follows the same pattern without selection: the motive raises it by 7.4 percentage points among low scorers ($\beta_1 + \delta_1$) and by 2.3 percentage points among the other participants (β_1), although the difference is not significant ($\delta_1 = 5.2$, SE 3.6, $p = 0.15$). The larger dispersion of estimates of one's own score without selection thus comes mainly from low scorers, who overestimate their score. This is consistent with participants resisting unfavourable news about their own ability. Two caveats apply. First, in *Own-first* a low own score is also a low target, while in *Other-first* the target is unrelated to the participant's own score. The pattern is therefore also consistent with estimates of one's own score being pulled towards the middle of the range of scores, for instance by participants' impression of how they did on the test, which would raise the estimates of low scorers without any motive. The design cannot separate the two

explanations. Second, the split is not pre-registered, and with twelve coefficients per model, a few marginally significant estimates are expected by chance.

Table 23: SN×IQ, exploratory: low scorers

	Error	Absolute error
Constant (β_0)	5.49^{***}	8.10^{***}
	(1.60)	(1.41)
Own (β_1)	-2.36	2.28
	(2.39)	(1.97)
Selection+ (β_2)	5.43^{**}	5.84^{***}
	(2.24)	(1.84)
Selection− (β_3)	-8.04^{***}	5.03^{***}
	(2.48)	(1.80)
Own × *Selection+* (β_4)	3.55	-1.19
	(3.21)	(2.56)
Own × *Selection−* (β_5)	4.71	-4.09
	(3.50)	(2.53)
Low (δ_0)	-3.42	-0.41
	(2.65)	(2.25)
Own × Low (δ_1)	10.60^{**}	5.16
	(4.29)	(3.60)
Selection+ × Low (δ_2)	3.98	0.10
	(3.50)	(2.87)
Selection− × Low (δ_3)	0.55	-3.41
	(3.80)	(3.01)
Own × *Selection+* × Low (δ_4)	-10.84^{*}	-2.32
	(5.85)	(4.77)
Own × *Selection−* × Low (δ_5)	-10.78^{*}	4.28
	(6.21)	(4.64)
Joint test $\delta_0 = \cdots = \delta_5 = 0$, p	0.232	0.027
N	762	762
R^2	0.136	0.057

Notes: OLS of the error of the first estimate (estimate minus target score, in percentage points) and of its absolute value on the regressors of the main specification, each also interacted with Low. Low = 1 if the participant's own score is below 50% (the sample median); participants scoring exactly 50% have Low = 0. β_0–β_5 are the effects for participants with Low = 0; δ_0–δ_5 are the differences for low scorers. The joint test is an F-test. Not pre-registered. Heteroskedasticity-robust (HC3) standard errors in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

I Additional Figures

Figure 5 shows the distribution of respondents' median χ in each condition of CN×ME (Section 5.3.2). Conditions without the heuristic are on the left and conditions with it on the right, so that each row compares cells that differ only in the availability of the heuristic.

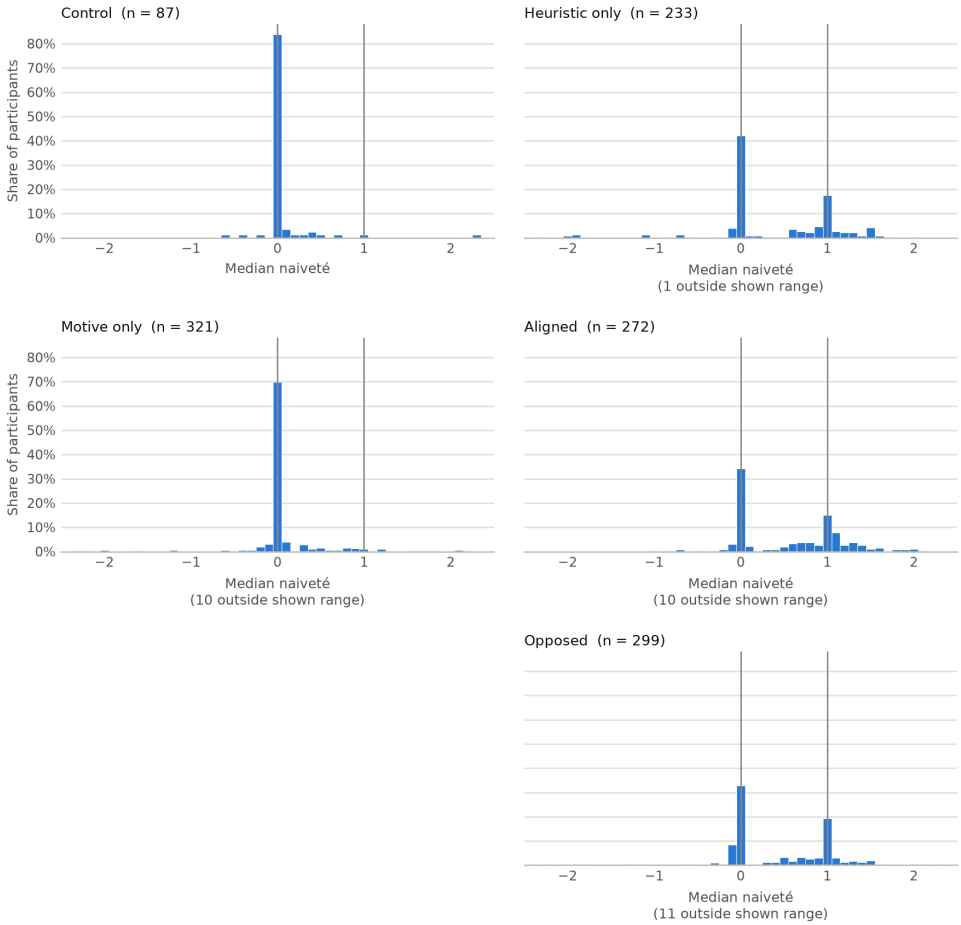


Figure 5: Distribution of respondents' median naiveté parameter by condition

Notes: $\chi = 0$ corresponds to the rational answer and $\chi = 1$ to the naive answer. Left column: conditions without the heuristic; right column: conditions with it. In *Control* and *Motive only*, χ is computed relative to the naive answer in the correlated version of the same round and signed in the direction of the motive. Bars show the share of respondents in each bin; respondents with $|\chi| > 2.5$ are not shown, and their number is reported below each panel.

Figure 6 shows the distribution of errors in the first estimate in each cell of SN×IQ (Section 6.3.1).

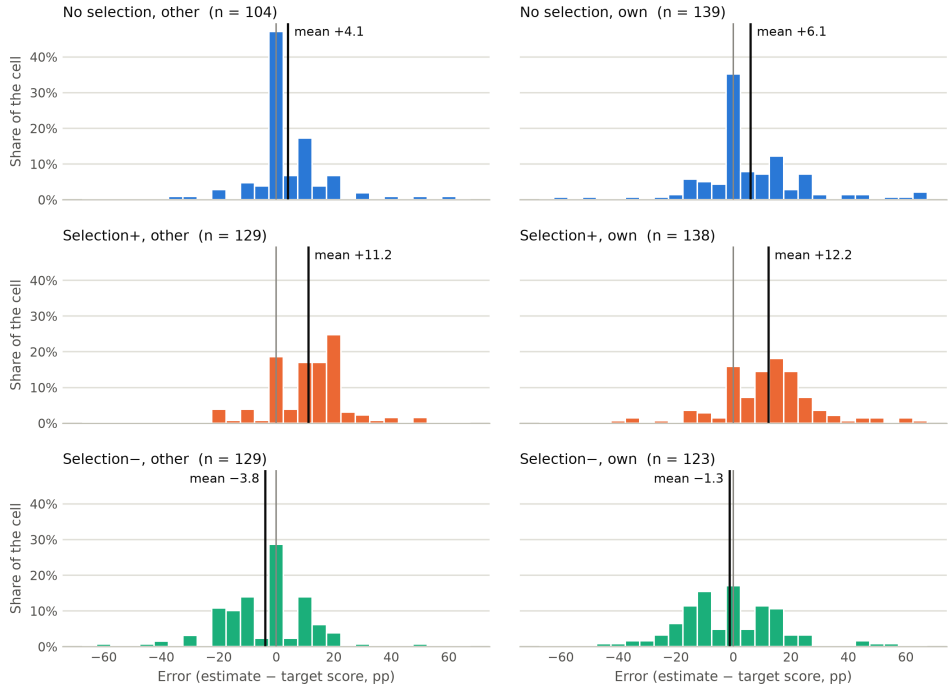


Figure 6: Distribution of errors in the first estimate by cell

Notes: Error = estimate minus target score, in percentage points; a positive error is an overestimate. The black line marks the mean of the cell. *Other* and *own* refer to the target of the first estimate (*Other-first* and *Own-first*).

J Experimental Instructions

This appendix reports screenshots of the pages shown to participants in the two experiments. For each experiment, I report the full sequence of pages for one condition. Where a page differs across conditions, I describe the differences in the text instead of repeating the screenshot.

J.1 CN×ME

The screenshots below are from the condition with an upward motive and *Correlation+* (*Aligned*). Only the two pages of instructions of Part 2 differ across conditions; all other pages are identical, except that the summary figures repeated on later pages are those of the participant's condition.

The amounts in the payment rule depend on the participant's choices in Part 1: the screenshots are from a participant whose charity equivalent is \$1.00. After the consent form, participants see a welcome page (Figure 7) and a page that introduces the bonus part and the charity (Figure 8). Each part opens with a cover page that is shown for 3 seconds and that I do not report ("Part 1 will begin shortly." and "You completed Part 1. Part 2 will begin shortly."). Participants who answer a comprehension question incorrectly twice are asked to return their submission.

Part 1: calibration. Figure 9 reports the instructions of the price list and Figure 10 the price list itself, after a participant has clicked on the row where they switch to the bonus for themselves. Before the click, no option is highlighted and the button to continue is hidden.

Part 2: estimation. Figure 11 reports the instructions on the payment rule and Figure 12 those on the information structure. A reminder page then shows again the two summary figures of the instructions, the table of payments and the diagram of the information structure, with no other content. Figure 13 shows the first of the six rounds; the other rounds differ only in the numbers. The values shown differ across rounds and participants (Table 15 in Appendix E).

The instructions differ across conditions in two ways.

Payment rule. With a downward motive, the two outcomes of a wrong guess are swapped: "If your guess is too high, Emergency USA will be allocated a donation of \$1.50", "If your guess is too low, you will be allocated a bonus payment of \$1.00", and the ticks and crosses in the rows "Too high" and "Too low" of the summary table are swapped accordingly. In *No motive*, the list of payments reads: "If your guess is right, you will be allocated a bonus payment of \$1.00. If your guess is too high or too low, you will not be allocated any bonus payment." The summary table then has a single column, "Bonus payment for you", with a tick only in the row "Correct", and is introduced by "The table below summarizes the payments allocation based on your guess." The options of the three comprehension questions on the payment become "A bonus payment of \$2.00.", "A bonus payment of \$1.00." and "You will not be allocated any bonus payment."

Information structure. Correlation+ and *Correlation−* have the same instructions: participants are not told which of the four estimates is reported directly. In *No correlation*, each news channel reports one estimate: "Channel 1 News reports Estimate 1. Channel 2 News reports Estimate 2. Channel 3 News reports Estimate 3." The example reads accordingly ("Channel 1 News reports 15", and so on), and the two diagrams show each estimate going to one channel only, without the arrows from Estimate 0 to the channels. In the first comprehension question, the option "The average of the signals you see." is replaced by "Equal to Estimate 0."; the correct answers to the questions on Channel 2 News and on Channel 1 News become "Estimate 2." and "12".

End of the study. The last page reports the part drawn for payment and the resulting payments, and asks for comments. When Part 2 is drawn (Figure 14), it reports the round drawn, whether the guess in that round was higher than, lower than or equal to the correct answer, and who receives a payment. When Part 1 is drawn, the page reports instead the row drawn from the price list and the option chosen in it. Participants are then redirected to Prolific.

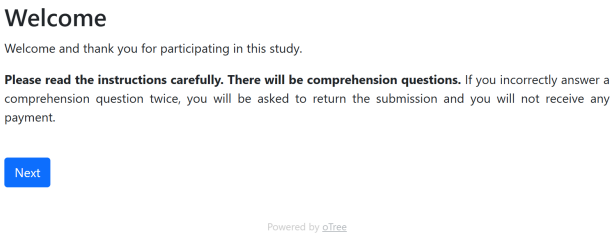


Figure 7: CN×ME: welcome page.

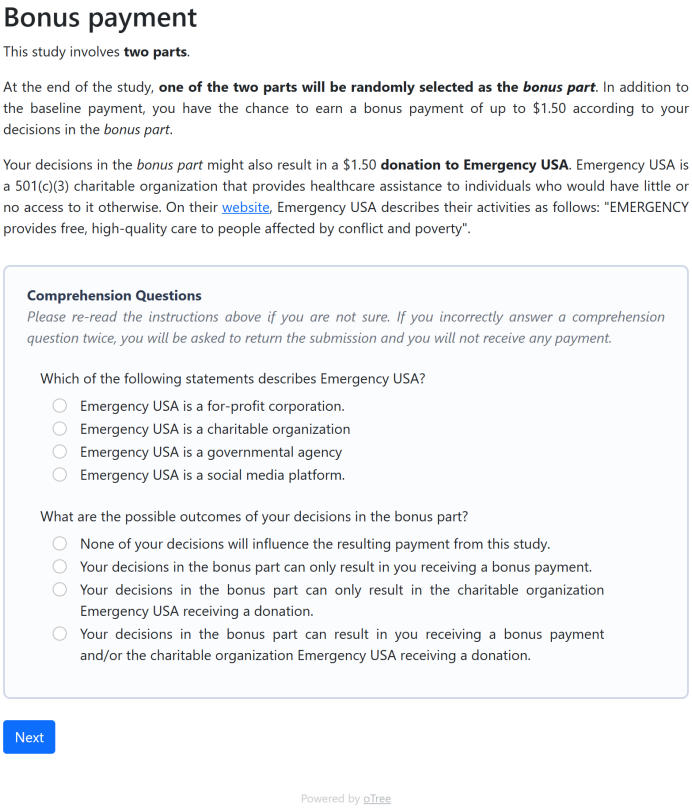
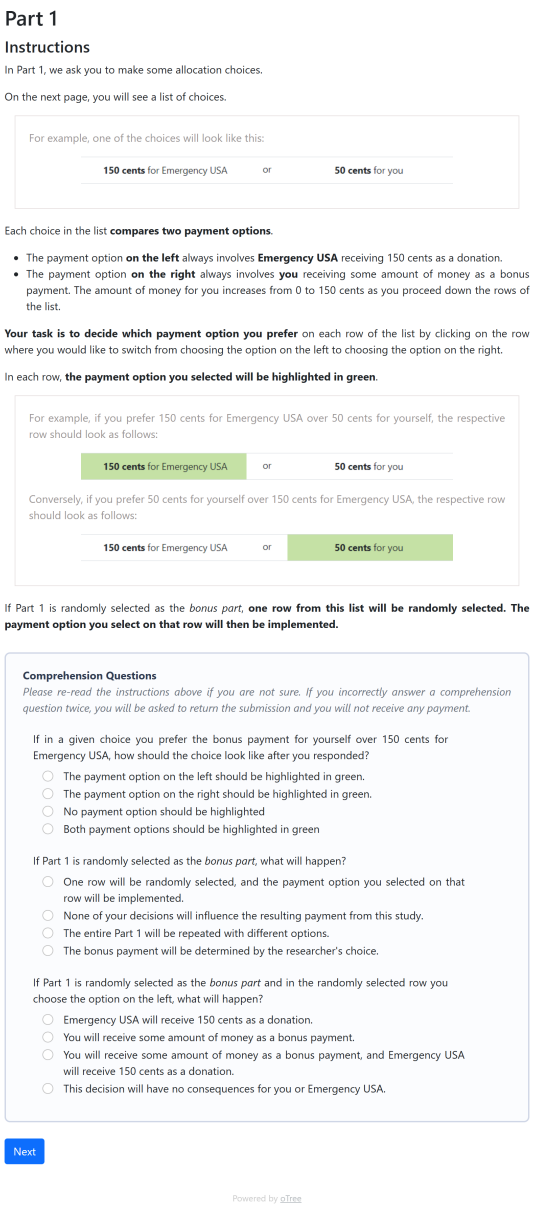


Figure 8: CN×ME: bonus payment and charity.



Part 1

Instructions

In Part 1, we ask you to make some allocation choices.

On the next page, you will see a list of choices.

For example, one of the choices will look like this:

150 cents for Emergency USA	or	**50 cents** for you

Each choice in the list **compares two payment options**.

- The payment option **on the left** always involves **Emergency USA** receiving 150 cents as a donation.
- The payment option **on the right** always involves **you** receiving some amount of money as a bonus payment. The amount of money for you increases from 0 to 150 cents as you proceed down the rows of the list.

Your task is to decide which payment option you prefer on each row of the list by clicking on the row where you would like to switch from choosing the option on the left to choosing the option on the right.

In each row, **the payment option you selected will be highlighted in green**.

For example, if you prefer 150 cents for Emergency USA over 50 cents for yourself, the respective row should look as follows:

150 cents for Emergency USA	or	**50 cents** for you

Conversely, if you prefer 50 cents for yourself over 150 cents for Emergency USA, the respective row should look as follows:

150 cents for Emergency USA	or	**50 cents** for you

If Part 1 is randomly selected as the *bonus part*, **one row from this list will be randomly selected. The payment option you select on that row will then be implemented.**

Comprehension Questions

Please re-read the instructions above if you are not sure. If you incorrectly answer a comprehension question twice, you will be asked to return the submission and you will not receive any payment.

If in a given choice you prefer the bonus payment for yourself over 150 cents for Emergency USA, how should the choice look like after you responded?

- The payment option on the left should be highlighted in green.
- The payment option on the right should be highlighted in green.
- No payment option should be highlighted
- Both payment options should be highlighted in green

If Part 1 is randomly selected as the *bonus part*, what will happen?

- One row will be randomly selected, and the payment option you selected on that row will be implemented.
- None of your decisions will influence the resulting payment from this study.
- The entire Part 1 will be repeated with different options.
- The bonus payment will be determined by the researcher's choice.

If Part 1 is randomly selected as the *bonus part* and in the randomly selected row you choose the option on the left, what will happen?

- Emergency USA will receive 150 cents as a donation.
- You will receive some amount of money as a bonus payment.
- You will receive some amount of money as a bonus payment, and Emergency USA will receive 150 cents as a donation.
- This decision will have no consequences for you or Emergency USA.

Next

Powered by oTree

Figure 9: CN×ME: instructions of the price list (Part 1).

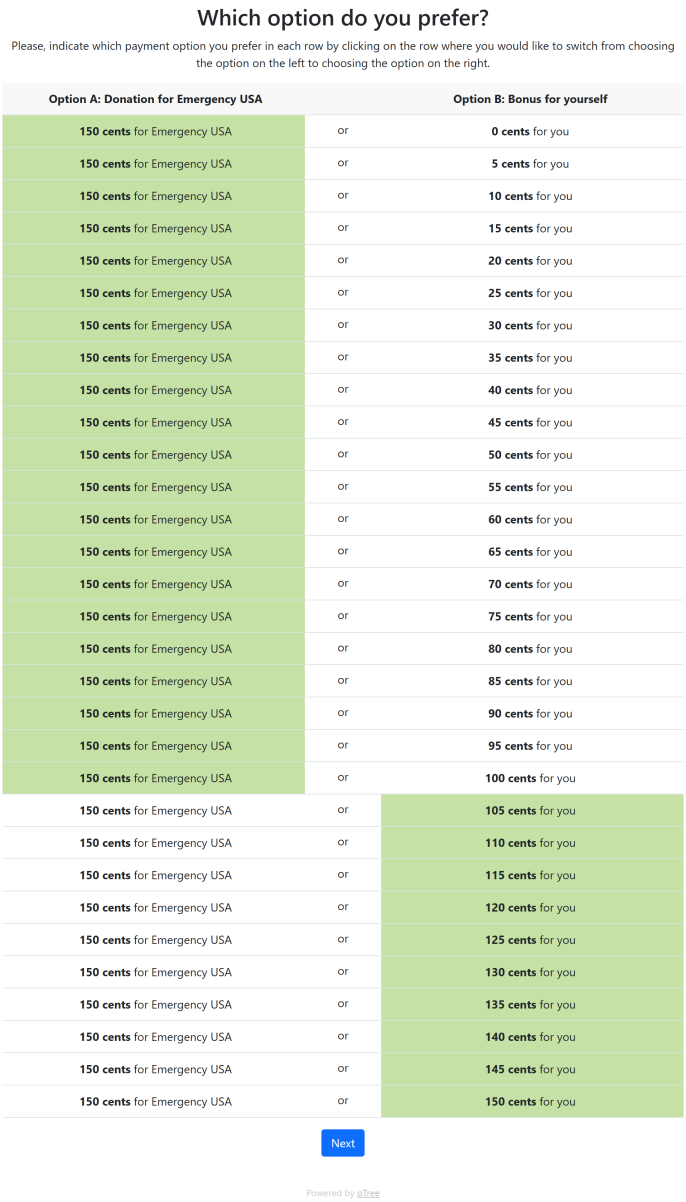
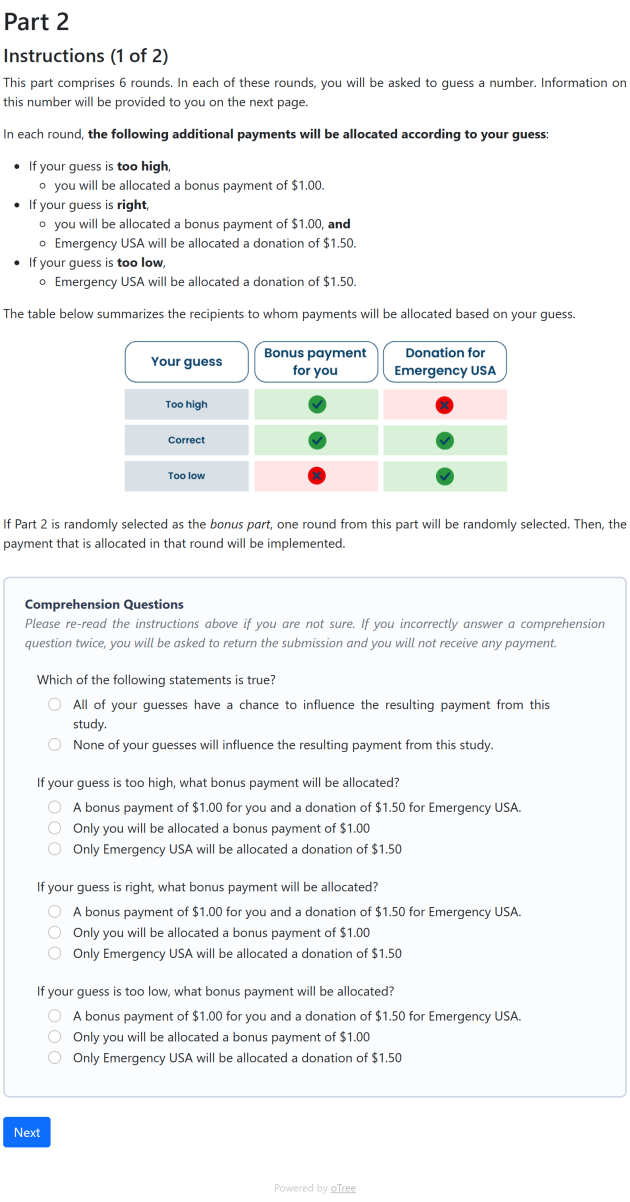


Figure 10: CN×ME: price list, after a participant has chosen to switch at 105 cents for themselves (charity equivalent of $1.00). On the screen, the list is in a scrollable box.



Part 2

Instructions (1 of 2)

This part comprises 6 rounds. In each of these rounds, you will be asked to guess a number. Information on this number will be provided to you on the next page.

In each round, **the following additional payments will be allocated according to your guess**:

- If your guess is **too high**,
  - you will be allocated a bonus payment of $1.00.
- If your guess is **right**,
  - you will be allocated a bonus payment of $1.00, **and**
  - Emergency USA will be allocated a donation of $1.50.
- If your guess is **too low**,
  - Emergency USA will be allocated a donation of $1.50.

The table below summarizes the recipients to whom payments will be allocated based on your guess.

Your guess	Bonus payment for you	Donation for Emergency USA
Too high	✓	✗
Correct	✓	✓
Too low	✗	✓

If Part 2 is randomly selected as the *bonus part*, one round from this part will be randomly selected. Then, the payment that is allocated in that round will be implemented.

Comprehension Questions

Please re-read the instructions above if you are not sure. If you incorrectly answer a comprehension question twice, you will be asked to return the submission and you will not receive any payment.

Which of the following statements is true?

- ○ All of your guesses have a chance to influence the resulting payment from this study.
- ○ None of your guesses will influence the resulting payment from this study.

If your guess is too high, what bonus payment will be allocated?

- ○ A bonus payment of $1.00 for you and a donation of $1.50 for Emergency USA.
- ○ Only you will be allocated a bonus payment of $1.00
- ○ Only Emergency USA will be allocated a donation of $1.50

If your guess is right, what bonus payment will be allocated?

- ○ A bonus payment of $1.00 for you and a donation of $1.50 for Emergency USA.
- ○ Only you will be allocated a bonus payment of $1.00
- ○ Only Emergency USA will be allocated a donation of $1.50

If your guess is too low, what bonus payment will be allocated?

- ○ A bonus payment of $1.00 for you and a donation of $1.50 for Emergency USA.
- ○ Only you will be allocated a bonus payment of $1.00
- ○ Only Emergency USA will be allocated a donation of $1.50

Next

Powered by oTree

Figure 11: CN×ME: instructions on the payment rule (Part 2), upward motive.

Part 2

Instructions (2 of 2)

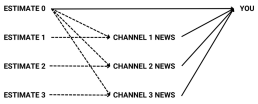
As mentioned before, you will face 6 rounds. In each of these round, **you will be asked to guess a number**. The number considered is different in each round and each number is between 0 and 100.

In each round, **the number you have to guess is equal to the average of four estimates**: Estimate 0, Estimate 1, Estimate 2 and Estimate 3.

You will be directly informed of Estimate 0.

You will also get some information about the other three estimates from three news channels.
Channel 1 News reports the average of Estimate 0 and Estimate 1.
Channel 2 News reports the average of Estimate 0 and Estimate 2.
Channel 3 News reports the average of Estimate 0 and Estimate 3.

The figure below represents the structure of the information you will get.



For example, suppose that:

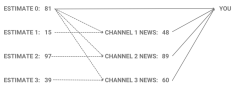
- Estimate 0 is 81.
- Estimate 1 is 15.
- Estimate 2 is 97.
- Estimate 3 is 39.

Then, the number you are asked to guess is 58 (this is the average of 81, 15, 97 and 39)

You will be provided with the following information:

- Estimate 0 is 81.
- Channel 1 News reports 48 (the average of 81 and 15).
- Channel 2 News reports 89 (the average of 81 and 97).
- Channel 3 News reports 60 (the average of 81 and 39).

The figure below represents the information structure applied to the example.



Comprehension Questions

Please re-read the instructions above if you are not sure. If you incorrectly answer a comprehension question twice, you will be asked to return the submission and you will not receive any payment.

In each round, the number you are asked to guess is
- The average of the four estimates.
- The average of the three news reports.
- The average of the signals you see.
- None of the above.

In each round, you observe:
- The four estimates.
- Estimate 0 and three news reports.
- Four news reports.
- Only estimate 0.

In each round, Channel 2 News reports:
- Estimate 2.
- The average of estimate 0 and estimate 2.
- The average of estimate 1 and estimate 2.
- The average of estimate 0, estimate 1, and estimate 2.

Suppose that Estimate 0 is 6, Estimate 1 is 12 and Estimate 2 is 16. What would Channel 1 News report?
- 6
- 12
- 9
- 11

Next

Powered by oTree

Figure 12: CN×ME: instructions on the information structure (Part 2), *Correlation*.

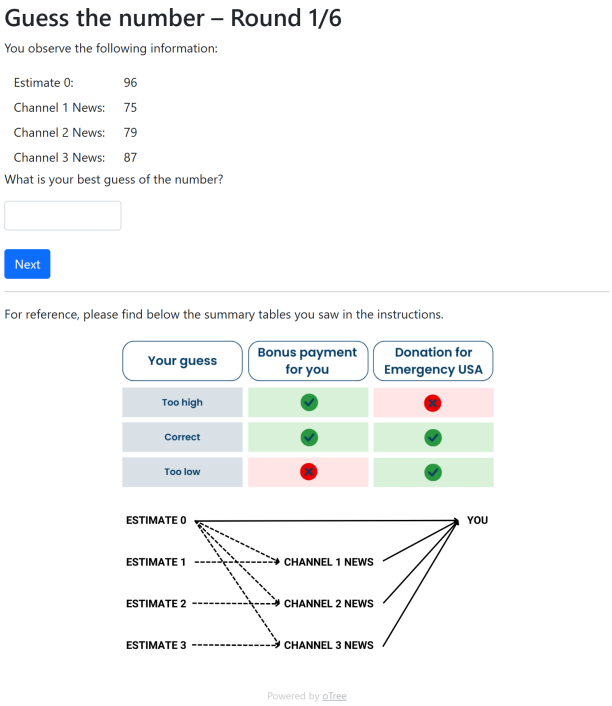


Figure 13: CN×ME: first round of Part 2, upward motive and *Correlation+*.

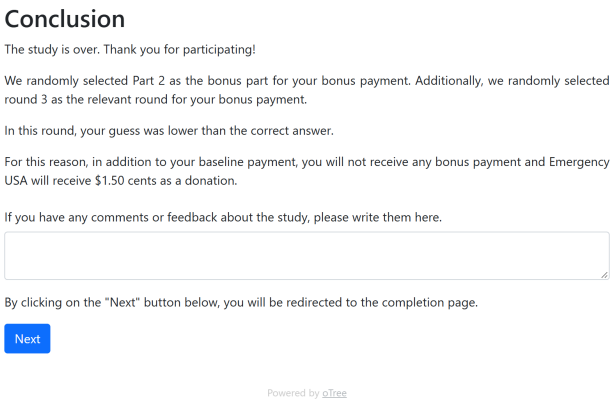


Figure 14: CN×ME: last page, when Part 2 is drawn for payment (upward motive and *Correlation−*, guess below the correct answer).

J.2 SN×IQ

The screenshots below are from the *Own-first* × *Selection+* condition (*Aligned*). Only the instructions of Part 2 and the estimation pages differ across conditions; all other pages are identical. After the consent form, participants see an overview of the study (Figure 15). Each part opens with a cover page that is shown for 2.5 seconds and that I do not report ("Part 1 – Logical Reasoning Test" and "You have completed Part 1. Part 2 is about to begin.").

Part 1: test. Figure 16 reports the instructions of the test, with the example question expanded (participants open it with a click). Participants must answer the comprehension questions at the bottom of the page correctly; those who fail them twice are asked to return their submission. Each block of questions opens with a short page (Figure 17); from Block 2 onward, it adds that the questions are harder than those in the previous block. Figure 18 shows the first question of the test. Participants see no feedback during the test.

Part 2: estimates. Figure 19 reports the instructions of the first estimate and Figure 20 the page on which it is entered. Participants must answer all comprehension questions correctly to continue, and can try again as often as needed. The page of the second estimate, which is not announced in advance, is in Figure 21; it is followed by an estimation page identical to Figure 20, except that it is headed "The other participant's performance in the test" and asks for "the percentage of the questions of the test that the other participant answered correctly". The reports on the estimation pages differ across participants; in the *Selection* conditions, one of the four reports is missing.

The instructions differ across conditions in two ways.

Reporting rule. Only the third bullet point under "How the reports work" changes (the same bullet point appears on the page of the second estimate). In *Selection+* it reads as in Figure 19. In *Selection−* it reads: "Each reviewer then reports the percentage they calculated, with one exception: a reviewer does not provide a report when you answered all questions they checked correctly (that is, when the percentage they calculated is 100%). In that case, you will not see a report from this reviewer." In *No selection* it reads: "Each reviewer then reports the percentage they calculated. A reviewer always provides a report." The options of the comprehension question "When does a reviewer not provide a report?" are the same in all conditions; only the correct one changes.

Target. In *Other-first*, the instructions of the first estimate are those of Figure 19 with "the other participant" in place of "you" throughout (e.g. "your task is to estimate the percentage of questions that another participant answered correctly in the test", "we will never explicitly tell you the other participant's score"). After the first paragraph, they add the same sentence on the other participant as Figure 21: "The other participant took part in an earlier session of this study and completed the same Part 1 as you with the same questions, in the same order, and under the same time limit." In the first comprehension question, the wrong option "The percentage of

questions that the reviewers checked" is replaced by "The percentage of questions of the test that you answered correctly". Symmetrically, the page of the second estimate in *Other-first* is that of Figure 21 with "you" in place of "the other participant", without the sentence on the earlier session.

Final questions. After the two estimates, participants answer three unincentivised questions, identical across conditions: what the reporting rule was (Figure 22), the average of four numbers (Figure 23; the numbers vary across participants), and five statements on how much their self-esteem depends on their intelligence (Figure 24). The study ends with a page that reports the bonus and asks for comments (Figure 25), after which participants are redirected to Prolific.

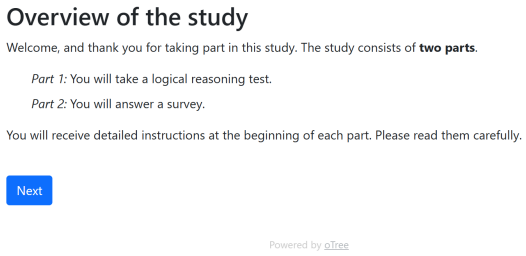


Figure 15: SN×IQ: overview of the study.

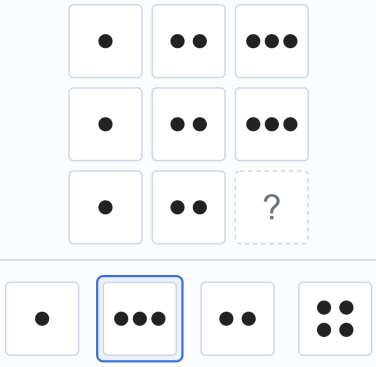
Instructions

In this part, we ask you to complete a **logical reasoning test**.

Tests of this kind are widely used to **measure intelligence**: according to several scientific studies, people who score higher in them tend, on average, to do better in education and at work. Your score therefore says something about how well you reason compared to other people.

- The test consists of **4 blocks** of **5 questions** each.
- In each question, you will see a pattern of figures with one figure missing, marked by a question mark. Your task is to **choose the figure that completes the pattern**.

▼ Click here to see an example



In each row, the number of dots increases by one from left to right. The missing figure therefore has three dots: the right answer is the highlighted option.

- The blocks are of **increasing difficulty**: the questions in Block 1 are the easiest, and those in Block 4 are the hardest. Only few people can solve the questions in Block 4.
- You have **at most 20 seconds** to answer each question. If you do not submit an answer within 20 seconds, your answer is counted as **wrong**.
- **Earn a bonus for every correct answer:** 5 cents each, paid on top of your base payment.

Comprehension Questions

If you answer a comprehension question incorrectly twice, you will be asked to return your submission and will not be paid.

How are the four blocks of questions ordered?

- ○ All blocks are equally difficult.
- ○ By increasing difficulty: Block 1 is the easiest and Block 4 is the hardest.
- ○ By decreasing difficulty: Block 1 is the hardest and Block 4 is the easiest.
- ○ In a random order: the difficulty of a block has nothing to do with its number.

What happens if you do not submit an answer to a question within 20 seconds?

- ○ You get more time to answer the question.
- ○ The question is skipped and does not count.
- ○ Your answer to the question is counted as wrong.
- ○ The test ends and you cannot answer the remaining questions.

Next

Powered by oTree

Figure 16: SN×IQ: instructions of the test (Part 1), with the example expanded.

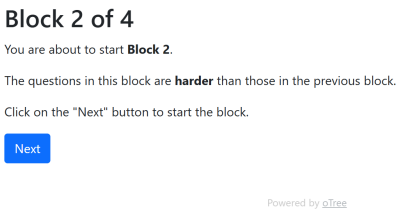


Figure 17: SN×IQ: page before Block 2 of the test. The page before Block 1 omits the sentence on difficulty.

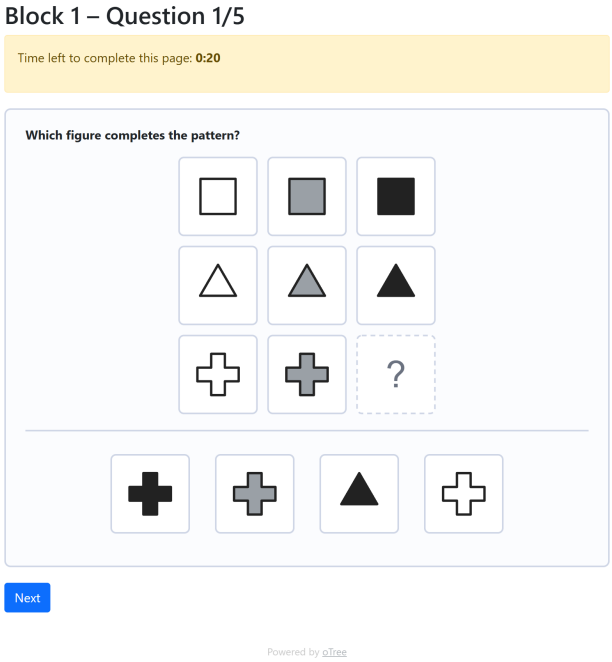


Figure 18: SN×IQ: first question of the test. The timer counts down from 20 seconds.

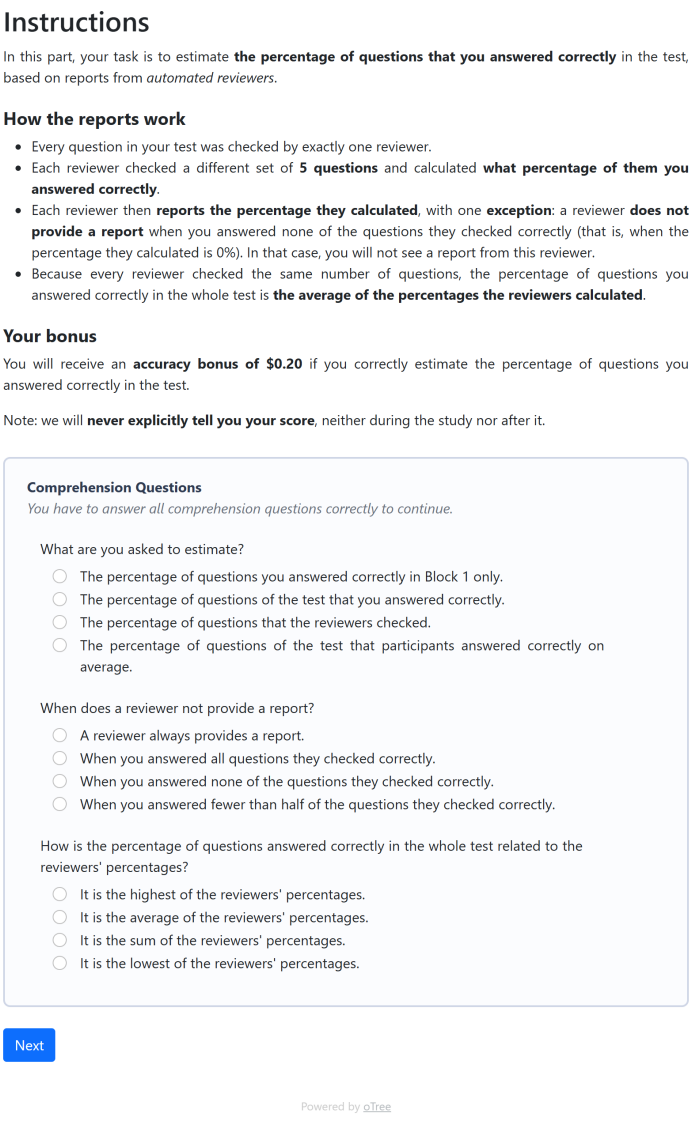


Figure 19: SN×IQ: instructions of the first estimate (Part 2), *Own-first* × *Selection+*.

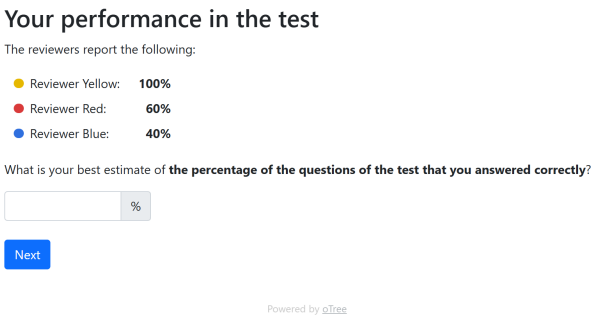


Figure 20: SN×IQ: first estimate, *Own-first* × *Selection+*. The report of 0% is not shown.

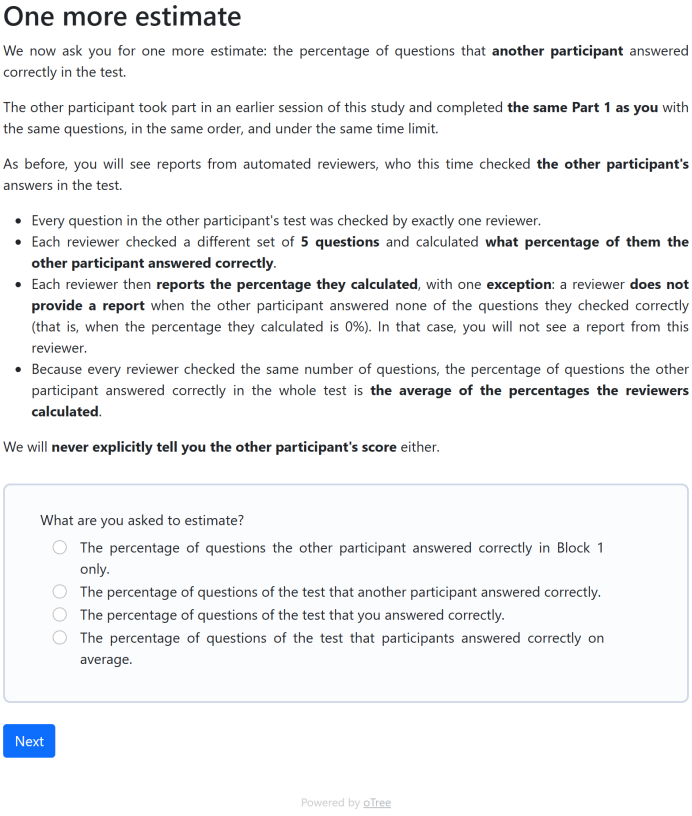


Figure 21: SN×IQ: instructions of the second estimate, *Own-first* × *Selection+*.

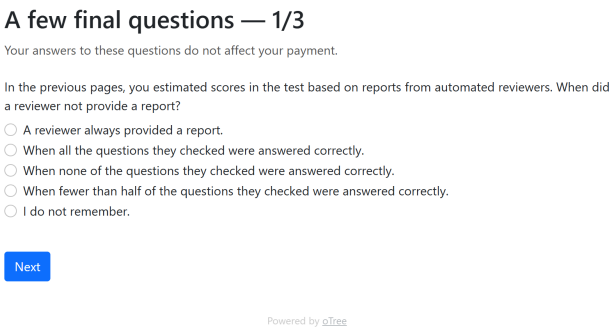


Figure 22: SN×IQ: final question on the reporting rule.

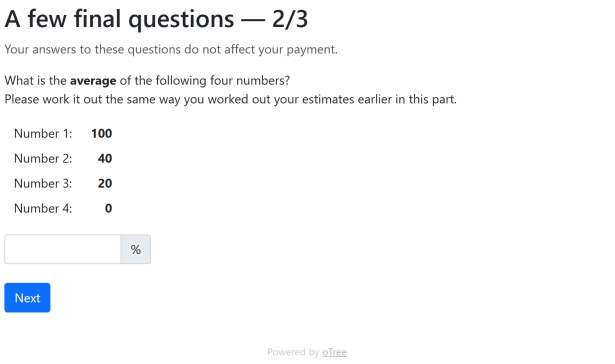


Figure 23: SN×IQ: final question on averaging.

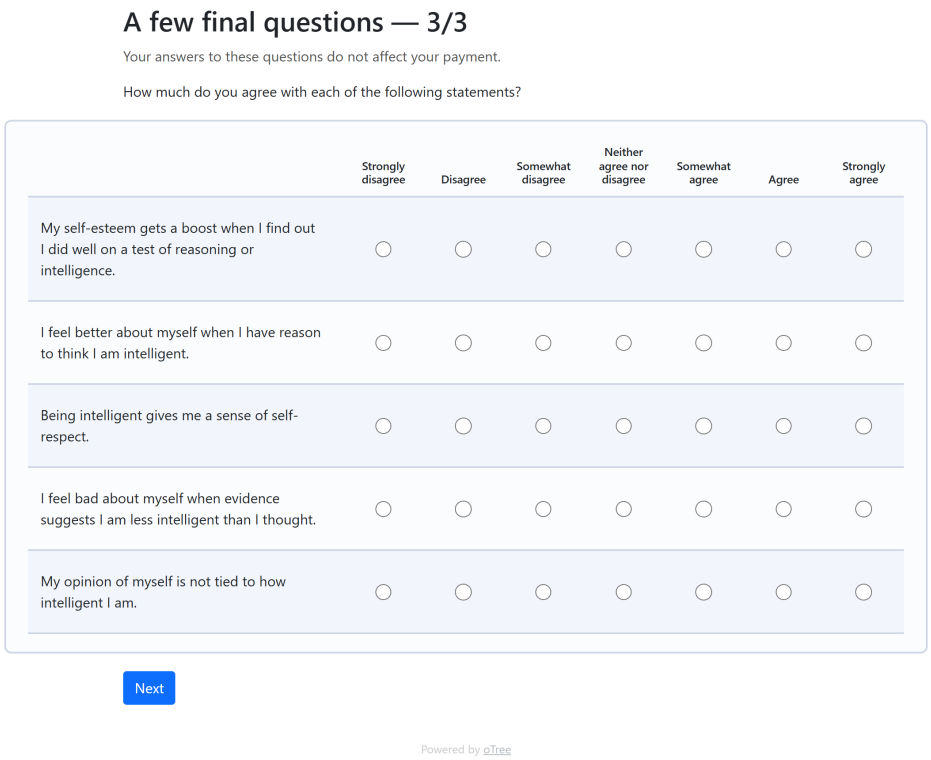


Figure 24: SN×IQ: final questions on self-esteem and intelligence.

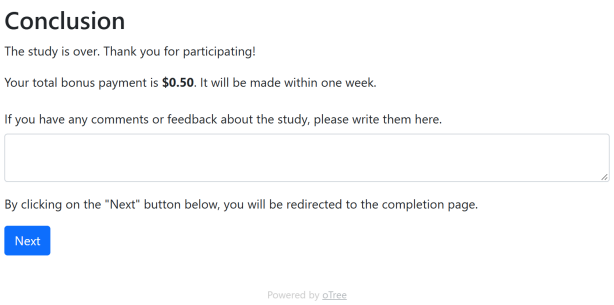


Figure 25: SN×IQ: last page.