# 

# Sections of a Stage 1 Registered Report

# Nature Human Behaviour

# Title page

# Abstract

# Introduction (no subheadings permitted)

# Methods

*Ethics information*

*Pilot data (if applicable)*

*Design*

*Sampling plan*

*Analysis plan*

## Data availability statement

## Code availability statement

## References

## Acknowledgements

## Author contributions

## Competing interests

## Figures & Figure captions

## Tables (A Design Table is mandatory)

**Supplementary information**

The title encapsulates the main research question in up to 100 characters (including spaces)

Aisha E. Bradshaw1\*, Jamie Horder1, Stavroula T. Kousta1, Anne-Marike Schiffer2, and Charlotte Payne1\*

1 Nature Human Behaviour, Springer Nature, London, UK

2 Nature Human Behaviour, Springer Nature, Berlin, Germany

\* Corresponding authors: Aisha E. Bradshaw ([aisha.bradshaw@springernature.com](mailto:aisha.bradshaw@springernature.com)); Charlotte Payne ([Charlotte.Payne@nature.com](mailto:Charlotte.Payne@nature.com)).

# Abstract

**The abstract of your Stage 1 Registered Report protocol should not exceed 150 words and should not contain any references. It should start with a sentence that introduces the general topic and its significance for a broad audience. It should then describe the specific question(s) your research addresses, what you will do, and broadly, what results would confirm or disconfirm your hypotheses. The abstract can be brief and will be slightly revised at Stage 2 submission to include the results.**

# Introduction

* The Introduction must include a review of the relevant literature that motivates the research question and a full description of the experimental aims (research questions) and hypotheses.

Each research question must be motivated, explained and linked to specific hypotheses (predictions). The Introduction must contain a detailed description of these hypotheses:

* Ensure that your predictions are defined precisely in terms of the specific independent and dependent variables.
* Listing them as Hypothesis 1, Hypothesis 2 etc (with corresponding H0 in each case, as appropriate) is recommended.
* The description of hypotheses must commit to interpretation of all potential data patterns (those that are predicted and those that would run counter to predictions). You cannot interpret lack of evidence (e.g. a p>0.05 in a t-test) for the existence of an effect in null hypothesis significance testing as evidence for the absence of an effect. To be able to interpret data patterns other than the predicted effect or a significant difference in the opposite direction, you must commit to using Bayesian inferential methods or frequentist equivalence testing.
* Where you describe your hypotheses, you must include a call-out to the **mandatory Design Table (Table 1)** – below.
* The only changes that will be allowed to your in-principle-accepted Stage 1 protocol when you resubmit at Stage 2 will be the tense of the sentences. In the Stage 1 protocol, describe the proposed research in the future tense to avoid confusion, e.g. with existing pilot data.
* If the evidence you will present is **correlational** or **cross-sectional**, you must ensure not to use causal language or language that implies causality
* Your manuscript should NOT **recycle text** from your own or others’ previous publications without proper acknowledgment and citation of the original work. Note that text recycling from the authors’ own work is a form of plagiarism and must be avoided (see our policy here: https://www.nature.com/authors/policies/plagiarism.html). When reusing text verbatim, you must clearly indicate this in your manuscript and identify the original source. If the portion of reused text exceeds 80 words, you must seek permission from the publisher of the original work to reproduce the text.
* The Introduction should **not** include any subheadings.
* The Introduction (and the manuscript in its entirety) must **not include priority or novelty claims** (except in the case of genetic or archaeological discovery).
* References in the main text should appear as superscript numerals, in order of mention. Only articles that have been published or accepted by a named publication or recognized preprint server should be in the **reference list**. If a manuscript is under consideration or not yet submitted, it should be mentioned in the main text only in parentheses, as follows: (Up to five author names, et al., unpublished manuscript). Published conference abstracts, numbered patents and research datasets that have been assigned a digital object identifier may be included in the reference list.

Decades of research have shown that heuristics — mental shortcuts — and the systematic biases that can result from them are a central component of human judgment decision making (Tversky & Kahneman, 1974; Gilovich & Kahneman, 2002). An important recurring question is whether these heuristics/biases operate consciously or unconsciously. For instance, tests of the availability heuristic have shown that people’s judgments of the probability of an event are influenced by how readily examples of that event come to mind (Tversky & Kahneman, 1973). To what extent do people have conscious awareness of this influence? (In the typology of Melnikoff & Bargh (2018), to what extent are people aware of the “cause/effect” relation in heuristics?)

Different theories of heuristics & biases have given different answers. The traditional model of heuristics and biases holds that they occur unconsciously. (Kahneman, 2011; Stanovich et al., 2008). In “Cognitive Illusions” (Pohl, 2004), thirty-three experts on specific heuristics and biases have written chapters on the heuristics or bias they research as a form of cognitive illusion, which, definitionally, occur involuntarily and are hard if not impossibly to avoid. However, there has been some more recent theoretical debate about whether heuristics and biases are entirely unconscious. In their model of decision-making, Gigerenzer & Gaissmaier (2011) argue that heuristics can be used both consciously and unconsciously. Melnikoff & Bargh (2018) use the “Type 1/Type 2 limbo” that they claim heuristics and biases fall into as evidence for their larger argument that the Type 1/Type 2 distinction is misguided.

Understanding the relative conscious accessibility of heuristics and biases matters both because it advances our understanding of conscious and unconscious thought and the limits of introspection more broadly and because this understanding is vital to research into debiasing. As Fischhoff (1982) points out in his review of debiasing methods, the underlying assumptions we have about the mechanisms of heuristics and biases will point us to strategies for debiasing.

Despite the theoretical discussion and the importance of understanding this question, there’s relatively little direct evidence bearing on it. Previous work has often focused on testing for conscious awareness in areas of decision making other than heuristics and biases (Newell & Shanks, 2014) such as implicit attitudes (Hahn & Gawronski, 2019; Hahn et al., 2014; Morris & Kurdi, 2023), multiple cue learning (Lagnado et al, 2006), morally-motivated beliefs (Cusimano & Lombrozo, 2023), and social self-perceptions (Bollich et al. 2015). In studies examining heuristics and biases specifically, some have tested whether warning participants about the bias leads to it influencing them less (Pohl & Hell, 1996; Harley et al., 2004; (Compen et al., 2002; Wetzel et al., 1981; Sagana et al., 2018; Starns et al., 2007; Westerberg et al., 2006; Grady et al., 2021). However, this provides at best indirect evidence of conscious access to those biases, and only a subset of the classic biases have been tested using this method, each by a different team using slightly different methods. Finally, some studies have induced a bias in participants and then asked them whether they were affected by that bias (Latane et al. 1968; Wilson et al. 1996; Pronin & Kugler, 2006; Wetzel et al. 1981; Dovido & Gaertner, 1991; Buehler et al., 1994; Gilbert et al., 2000; Cusimano & Lombrozo, 2023; Bollich et al. 2015a). Research into the bias blind spot extends these findings (Pronin et al., 2002; Pronin et al., 2004; Pronin & Kugler, 2007; Pronin, 2007; West et al., 2012). Pronin & Kugler (2007) provide evidence that the bias blind spot arises from our reliance on introspective rather than behavioral evidence when assessing our own biases, implying limited conscious access to our biases. None of these previous studies has investigated a representative array of biases which would allow us to find patterns across biases or across individuals. A second common concern for these studies is separating conscious awareness of a bias from simply inferring the presence of bias — even when participants claim to be aware of a bias, this doesn’t necessarily imply that they came to this knowledge through introspective awareness.

In this study, we introduce a novel paradigm that rigorously interrogates our conscious access to heuristics and biases across a range of canonical heuristics and biases. To select a set of representative heuristics and biases, we subjected every heuristic and bias mentioned described in In Gilovich et al.’s (2002) canonical Heuristics and Biases textbook to the following inclusion criteria:

1. One factor that is manipulated (so not just a bias relative to normative standards, e.g., the better than average effect or optimism bias)
2. Creates the presence of an effect (rather than the \*absence \*of an effect, e.g. in the base rate fallacy)
3. Can be run online in a short amount of time
4. Produces a consistent, large, replicable effect
5. The factor is easily explainable to a lay audience.
6. The manipulation is such that a factor is present in one condition and absent in the other, rather than the two conditions being very different (e.g., this rules out the "extra cost" effect, the Asian disease framing effect)
7. The factor has to be posited to play a direct cognitive role in the effect it creates, rather than indirectly influencing people in a way that could only be identified via inference (e.g., the Nisbett & Wilson dress-shopping problem)

This process produced the following seventeen heuristics and biases: 1) The Anchoring Effect, 2) The Associative Memory Effect, 3) the Availability Heuristic, 4) the Belief Bias, 5) Causal Inference, 6) the Decoy Effect, 7) the Halo Effect, 8) Hindsight Bias, 9) The Illusory Truth Effect, 10) The Imaginability Bias, 11) The Mere Exposure Effect, 12) Omission Bias, 13) The Recognition Heuristic, 14) Reference Price, 15) The Representativeness Heuristic, 16) the Status Quo Bias, and 17) The Sunk Cost Effect.

In our paradigm, participants first complete a task that induces each heuristic or bias. They then report the extent to which they believe they were influenced by that heuristic or bias. However, as argued by Bem (1972) and Nisbett & Wilson (1977) and as we saw for the prior research discussed above, the ability to report the extent to which one is influenced by a heuristic or bias could be supported either by conscious awareness of the bias or by simply inferring its presence through folk models of biases or through deducing our own biases through observing our own behavior. To address this, we introduced a critical control condition where some participants were not induced to use the heuristic or bias; we explained the experimental set-up and asked them to imagine whether they *would have used* the heuristic or bias in that situation. This control group can only use folk psychology or other kinds of inference to answer the question. Any difference between the two groups would imply the existence of conscious access to biases beyond those forms of inference.

To determine whether people are generally able to consciously assess whether or not they are affected by heuristics and biases (Aim 1), we will compare average reports of bias influence between the experimental and control groups. If heuristics and biases are conscious, that predicts participants in the factor-included group will report more influence of the bias than participants in the factor-excluded group (Hypothesis 1a). If heuristics and biases are not, that predicts no significant difference in reported influence between the two groups (Hypothesis 1b). Although this analysis will allow us to see in broad strokes whether participants show evidence of conscious access to bias, our next analysis will provide us with more fine-grained detail about participant’s conscious access to the extent to which they are affected by heuristics and biases (Aim 2). If the extent of influence is conscious, that predicts participants’ reports of how affected they were by the bias (and in which direction) will correlate with the extent and direction of their individual effect (Hypothesis 2a). If the extent of the influence is not conscious, that predicts a lack of such correlation (Hypothesis 2b).

The next sets of analyses will focus on variation among individuals and among heuristics and biases. First, we will analyze the variation among the heuristics and biases (Aim 3). **[insert hypotheses here]** Further, we will investigate whether this variation is cognitively meaningful. If so, this would predict that the variation would cluster into sets of heuristics and biases that the same individuals tend to be more or less accurate in reporting (Hypothesis 3c). If the variation is not cognitively meaningful, this would predict no such clustering (Hypothesis 3d).

In terms of individual differences, we will first broadly investigate whether our data includes individual differences that generalize across heuristics and biases in the ability to consciously access the extent to which people are affected by heuristics and biases (Aim 4). If there are individual differences, this would predict that an individual’s accuracy in reporting the effect of one heuristic or bias will correlate with their accuracy in the other heuristics and biases (Hypothesis 4a). If there are no individual differences, this would predict no such correlation (Hypothesis 4b). The next two analysis aims will help us better understand the causes of any individual differences we find. We will examine whether people are accurate in their perception of their self-knowledge (Aim 5). If people are accurate in their perception of their self-knowledge, this would predict that overall accuracy in reporting the effect of heuristics and biases will correlate with perceived self-insight (as measured by the Self-reflection and Insight Scale) (H5a). If people are not accurate in their perception of their self-knowledge, that would predict no such correlation (Hypothesis 5b). If people are accurate in their perception of their self-knowledge, this also would predict that participants’ reported confidence levels for their reports of bias influence will correlate with their accuracy for that report (Hypothesis 5c). If people are not accurate, this would predict no such correlation (Hypothesis 5d).

Finally, we will analyze whether people who are interested in introspection are more accurate at assessing the influence of heuristics and biases on them (Aim 6). If people who are interested in introspection are more accurate than those who are not, this would predict that overall accuracy in reporting the effect of heuristics and biases will correlate with reported interest in and frequency of self-reflection (as measured by the Self-reflection and Insight Scale) (Hypothesis 6a). If people who are interested in introspection are no more accurate than those who are not, this would predict no such correlation (Hypothesis 6b).

* + - Cite the few papers we know…
    - Summarize issues with them
    - One method for establishing how much conscious control we have over biases is to provide participants with warnings about the bias before they complete a bias-induction task. If biases operate entirely unconsciously, participants should have no capacity to consciously override their bias even when they are warned, as previous studies have observed for hindsight bias (Pohl & Hell, 1996 and Harley et al., 2004), herd thinking (Compen et al., 2002), the halo effect (Wetzel et al., 1981), and choice blindness (Sagana et al., 2018). However, multiple studies on the associative memory effect (Starns et al., 2007; Westerberg et al., 2006) have demonstrated the effectiveness of warnings. And Grady et al. (2021) complicate the question further by showing that warnings reduce the partisan congruency effect, but that the effect returns when participants are tested again two weeks later. And <https://psycnet.apa.org/buy/1996-06578-004>
    - **Papers where they ask people: did you do this? People say no. anchoring, bystander**
    - **N&W: not systematic, not main point, methodological questions**
    - One sentence: Work in other areas of decision making showing that things which had been posited to be unconscious people can actually report accurately (such as implicit attitudes, multiple cue learning, and more).
      * Multiple studies have found evidence of conscious awareness of implicit biases (Hahn & Gawronski, 2019; Hahn et al., 2014; Morris & Kurdi, 2023).
      * Also cite the [Newell & Shanks paper](https://www.cambridge.org/core/journals/behavioral-and-brain-sciences/article/abs/unconscious-influences-on-decision-making-a-critical-review/86885344F7E8A44457C3FC63CFA3F3AF)
      * [Lagnado paper](https://psycnet.apa.org/record/2006-06642-003) – multiple cue learning
      * Bollich et al. (2015) found that, on average, participants accurately reported the extent and direction (positive or negative) of their biased self-perceptions.
      * Cusimano & Lombrozo (2023) found that participants are aware of their morally-biased motivations for a factual belief but believe their reasoning is ideal.
    - Research into the bias blind spot and the introspection illusion (Pronin et al., 2002; Pronin, 2007; Pronin & Kugler, 2007) further supports this view of heuristics and biases. Across a number of different heuristics and biases tasks, participants systematically deny being affected by biases that they recognize in others. If we had conscious access to biases, we would instead expect participants to more accurately report the extent to which they are affected by bias than the extent to which someone else is.
      * Why this is not relevant
    - A common concern for all these studies is separating conscious awareness of a bias from simply inferring the presence of bias. In fact, Bollich et al.’s analysis implied that participants were simply using the heuristic of guessing they were positively biased for the traits they self-rated the highest and negatively biased for the traits they self-rated the lowest.
    - Also, nobody has tested awareness of them simultaneously, to see if there’s patterns across tasks or across people
  + Here, we introduce a novel paradigm that rigorously interrogates this question across a range of canonical heuristics and biases.
    - To select a set of representative heuristics and biases, we subjected every heuristic and bias mentioned described in In Gilovich et al.’s (2002) canonical Heuristics and Biases textbook to the following inclusion criteria:
      * One factor that is manipulated (so not just a bias relative to normative standards, e.g., the better than average effect or optimism bias)
      * Creates the presence of an effect (rather than the \*absence \*of an effect, e.g. in the base rate fallacy)
      * Can be run online in a short amount of time
      * Produces a consistent, large, replicable effect
      * The factor is easily explainable to a lay audience.
      * The manipulation is such that a factor is present in one condition and absent in the other, rather than the two conditions being very different (e.g., this rules out the "extra cost" effect, the Asian disease framing effect)
      * The factor has to be posited to play a direct cognitive role in the effect it creates, rather than indirectly influencing people in a way that could only be identified via inference (e.g., the Nisbett & Wilson dress-shopping problem)
    - This process produced the following seventeen heuristics and biases: 1) The Anchoring Effect, 2) The Associative Memory Effect, 3) the Availability Heuristic, 4) the Belief Bias, 5) Causal Inference, 6) the Decoy Effect, 7) the Halo Effect, 8) Hindsight Bias, 9) The Illusory Truth Effect, 10) The Imaginability Bias, 11) The Mere Exposure Effect, 12) Omission Bias, 13) The Recognition Heuristic, 14) Reference Price, 15) The Representativeness Heuristic, 16) the Status Quo Bias, and 17) The Sunk Cost Effect.
    - In our paradigm, participants first complete a task that induces each heuristic/bias. Then they report the extent to which they believe they were influenced by that heuristic or bias.
    - However, as argued by Bem (1972) and Nisbett & Wilson (1977) and as we saw for the prior research discussed above, the ability to report the extent to which one is influenced by a heuristic or bias could be supported either by conscious awareness of the bias or by simply inferring its presence through folk models of biases or through deducing our own biases through observing our own behavior.
    - To address this, we introduced a critical control condition where some participants were not induced to use the heuristic or bias; we explained the experimental set-up and asked them to imagine whether they *would have used* the heuristic or bias in that situation. This control group can only use folk psychology or other kinds of inference to answer the question.
  + Hypotheses
    - We’ll do an “overall” assessment that aggregates across heuristics
      * Our paradigm allows us to test this in two ways…
    - We’ll do a heuristic-specific assessment, and a heterogeneity test.
      * If we find significant heterogeneity, clusters?
    - Individual differences
    - Secondary
      * Confidence
      * Correlation with SRIS

1. Five research questions with hypotheses. The first is a general question about our capacity to introspect. The other four aim to help us understand the mechanisms of PIME by understanding patterns in the kinds of processes that are and are not consciously accessible (see [design table](#bookmark=id.iu6wnujqs5rt)):
   * 1. Are people generally able to consciously access the extent to which they are affected by heuristics and biases?
     + We predict that on average, participants who were affected by each heuristic and bias will report they were more strongly affected than those who were not affected by each heuristic or bias. (H1)
   * 2. Are there individual differences that generalize across heuristics and biases in the ability to consciously access the extent to which people are affected by heuristics and biases?
     + We predict that an individual’s accuracy in reporting the effect of one heuristic or bias will predict their accuracy in the other heuristics and biases. (H2)
   * 3. Do individual differences in people’s second-order abilities to introspect on heuristics and biases mirror people’s first-order tenancy to fall into heuristics and biases?
     + We predict that individual differences in accuracy in reporting the effect of one heuristic or bias will cluster similarly to Ceschi et al.’s (2019) clustering of first order tenancy to fall into heuristics and biases. (H3)
   * 4. Are people accurate in their perception of their self-knowledge?
     + We predict that overall accuracy in reporting the effect of heuristics and biases will correlate with perceived self-insight (as measured by the Self-reflection and Insight Scale). (H4)
   * 5. Does self-reported interest in and practice with introspection correlate with accuracy in reporting the extent to which people are affected by heuristics and biases?
     + We predict that overall accuracy in reporting the effect of heuristics and biases will correlate with reported interest in and frequency of self-reflection (as measured by the Self-reflection and Insight Scale). (H5)

We will code all responses such that up direction means direction of effect

* Find which ones are reverse coded. Make this 50%

1. Manipulation section
   1. Are people biased (not under hypothesis – under manipulation check)
   2. Data collada post
      1. What percentage affected
   3. Aggregating across all heuristics, do people report being more influenced by effect in included than excluded?
   4. Are people significantly above midpoint for both
2. Between condition comparison for each task separately
3. Formal test of variability across biases
4. Comparing people who were more or less affected
   1. Dichotomous split
   2. Gradient
5. People who claim to have seen the task before: are they more accurate
   1. And run all tests again excluding people? Or exclude them in the first place
6. Testing for stable individual differences
7. Self-reflection & insight
8. Clustering thing

Exclusion criteria

Power analyses

Make these into hypotheses

People reporting influence in the opposite direction

Question 1: Aggregating across all tasks, do people report being influenced by the factor (in the expected direction) a priori (in the control condition)? Do people report being influenced by it after experiencing it?

Question 2: Do people more report being influenced after experiencing it, compared to the control condition?

Question 3: Is there variability in these results across tasks?

* Running the analyses separately for each task
* Formally testing whether there’s meaningful variation
* If there is meaningful variation, clustering analysis

Question 4: Do people who were *more* influenced by the factor report it as having influenced them more?

Question 5: Do these objective results correlate with self-reported insight?

* Test against SRIS and confidence ratings

Question 6: Are there stable individual differences?

Start filling out table (minus statistics) and draft intro paragraphs

# Methods

## Ethics information

# If your protocol describes research with human participants, the Methods section must start with a statement confirming that the research complies with all relevant ethical regulations; naming the board and institution that approved the study protocol; and confirming that informed consent will be obtained from all human participants. Information on participant compensation must also be included.

* If your manuscript reports the results of research with **non-human animals**, the Methods section starts with a statement confirming that the research complies with all relevant ethical regulations; naming the board and institution that approved the study protocol; and confirming that the ARRIVE guidelines were used to report the research.
* If your manuscript reports the results of a **clinical trial**, the Ethics information section also includes the trial registration number from ClinicalTrials.gov or an equivalent approved trials registry.

## Pilot data

* You may include pilot data, for example to demonstrate the feasibility of your approach. Your pilot studies and results should be described briefly in the main manuscript and reported in full in Supplementary Information.
* Pilot data and custom analyses code should be made available and referred to in the Data Availability statement and Code availability statement. You may also include simulated data, for example to support your power analysis. This should also be made available.
* If you report analyses of pilot data using NHST (either in the main text or in Supplementary Information), you must report statistics **in full**:

statistic(degrees of freedom) = value, p = value, effect size statistic = value, % Confidence Intervals = values

## Design

* Your Methods section must include a description of experimental procedures in sufficient detail to allow another researcher to repeat the methodology exactly, without requiring further information other than that included in the protocol, your Supplementary Information file (if used) and, if applicable, the linked Code and Data (please refer to the **Code Availability** and **Data availability** statements below).

* Provide full descriptions of any outcome-neutral criteria and positive controls. These quality checks might include the absence of floor or ceiling effects in data distributions, positive controls, or other quality checks that are orthogonal to the experimental hypotheses.

# You must have a statement on randomization in the Methods, if applicable.

# For experimental studies, make it clear whether the design is within-subjects, between-subjects, mixed, or other.

# You must have a statement indicating whether blinding will be used in the Methods, if applicable. If there will be no blinding, this must be clearly stated in the manuscript, as follows: "Data collection and analysis will not be performed blind to the conditions of the experiments.”

* If your manuscript reports the results of a **Phase 2 or 3 randomized controlled trial**, you should also attach the CONSORT checklist with your submission.

### Sampling plan

* Studies involving Neyman-Pearson inference must include a statistical **power analysis**. Estimated effect sizes should be justified with reference to the existing literature. Since publication bias overinflates published estimates of effect size, power analysis must be based on the **lowest** available or meaningful estimate of the effect size. For frequentist analysis plans, the a priori power must be **0.95 or higher** for all proposed hypothesis tests. In the case of highly uncertain effect sizes, a variable sample size and interim data analysis is permissible but with inspection points stated in advance, appropriate Type I error correction for ‘peeking’ employed, and a final stopping rule for data collection outlined.
* Methods involving Bayesian hypothesis testing are encouraged. For studies involving analyses with Bayes factors, the predictions of the theory must be specified so that a Bayes factor can be calculated. Authors should indicate what distribution will be used to represent the predictions of the theory and how its parameters will be specified. For inference by Bayes factors, authors must be able to guarantee data collection until theBayes factor is at least 10 times in favour of the experimental hypothesis over the null hypothesis (or vice versa). Authors with resource limitations are permitted to specify a maximum feasible sample size at which data collection must cease regardless of the Bayes factor; however to be eligible for advance acceptance this number must be sufficiently large that inconclusive results at this sample size would nevertheless be an important message for the field.
* Regardless of sampling method, you must list all criteria for **data inclusion** and/or **data exclusion** and how this affects your sampling strategy. This includes a full description of proposed sample characteristics. Procedures for objectively defining exclusion criteria due to technical errors or for any other reasons must be specified, including details of how and under what conditions data would be replaced. These details must be summarized in the mandatory **Design table** (Table 1).

## Analysis Plan

* Your proposed analysis pipeline must include all pre-processing steps, and a precise description of all planned analyses (including appropriate correction for multiple comparisons if applicable). Any covariates or regressors must be stated. Where analysis decisions are contingent on the outcome of prior analyses, these contingencies must be specified and adhered to.
* Do not include exploratory analyses in the Stage 1 protocol. These should be reported in the Stage 2 manuscript, under the heading **Exploratory Analyses**.

**Should you need to deviate in any way from the description of your work in the Methods after acceptance in principle, you must seek editorial feedback first (before implementing these changes).**

* Our study protocol has been approved by the Yale University Institutional Review Board and will comply with their regulations. All participants will provide informed consent in accordance with the regulations of the review board. Participants will be paid $12 per hour.
* showing that we’ve gotten the tasks to work with pilot data (not that we’ve collected pilot data for our core hypotheses)
  + Pilot study: 65 participants. We excluded 21 for either failing to pass our three attention checks or switching out of the tab where the study was running more than 12 times while the study was running. Within each task, we also excluded participants who responded that they were previously familiar with a task similar to this one. Tasks in that study: anchoring, availability, belief bias, causal inference, decoy, halo, hindsight, mere exposure, reference price, representativeness, status quo. Tasks added in this full study: moral omission bias, imaginability, sunk cost, illusion of truth, recognition heuristic.
  + **insert more data analysis once we’ve talked over any potential changes in method of analysis**
* Study design:
  + Selected all heuristics and biases that fit the following criteria and are described in the canonical Heuristics and Biases textbook by Gilovich et al. (2002)
    - One factor that is manipulated (so not just a bias relative to normative standards, e.g., the better than average effect or optimism bias)
    - Creates the presence of an effect (rather than the \*absence \*of an effect, e.g. in the base rate fallacy)
    - Can be run online in a short amount of time
    - Produces a consistent, large, replicable effect
    - The factor is easily explainable to a lay audience.
    - The manipulation is such that a factor is present in one condition and absent in the other, rather than the two conditions being very different (e.g., this rules out the "extra cost" effect, the Asian disease framing effect)
    - The factor has to be posited to play a direct cognitive role in the effect it creates, rather than indirectly influencing people in a way that could only be identified via inference (e.g., the Nisbett & Wilson dress-shopping problem)
  + The protocol for each task was based on a previous large online study that demonstrated that effect. In the supplement, we will provide the precise protocol for each task.
  + At the beginning of the study, we will randomly assign each participant to the factor-included or factor-excluded group.
  + For instance, the factor-included condition of the anchoring-estimation task will present participants with an initial anchoring question: “is the average winter Antarctic temperature greater or less than -20 degrees Fahrenheit?” before asking the estimation question: “what is your estimate of the average winter Antarctic temperature?” The factor-excluded group will only be asked the second estimation question without the anchor. Finally, both groups will be asked how they think the heuristic or bias affected them or would have affected them. For example, in the factor-included condition of the anchoring-estimation task, participants will be asked “on a scale from -10 to 10, how do you think you were affected by being shown the initial question? -10 means being shown the initial question pulled your subsequent estimation closer to -20 degrees Fahrenheit. 0 means you were unaffected. 10 means being shown the initial question pushed your subsequent estimation further away from -20 degrees Fahrenheit.” In the factor-excluded condition of the anchoring-estimation task, participants will be asked “Imagine that before being asked to estimate the average winter Antarctic temperature, you had first been asked whether the average winter Antarctic temperature is greater or less than -20 degrees Fahrenheit. On a scale from -10 to 10, how do you think you would have been affected by being shown the initial question? -10 means being shown the initial question would have pulled your subsequent estimation closer to -20 degrees Fahrenheit. 0 means you would have been unaffected. 10 means being shown the initial question would have pushed your subsequent estimation further away from -20 degrees Fahrenheit.”
  + By comparing factor-included to factor-excluded we will be able to see if there is any such thing as perception of internal mental events for H&B or whether all understanding of how affected we are by H&B comes from inference. The factor-excluded group will be able to infer their internal mental events, but they will have nothing to perceive. The factor-included group will be able to infer their internal mental events, and if there is such a thing as perception of internal mental events for H&B, they will be able to use that as well. Any difference between the two groups will reflect this additional use of perception of internal mental events.
  + **add all details from Sophia’s thesis**
  + **list of all tasks and how we analysed each**
  + **Add more about analysis plan**

# Data availability

For Registered Reports, public sharing of data and materials upon acceptance for publication of the Stage 2 manuscript is mandatory. Please include a statement committing to sharing your raw data and materials on acceptance of your Stage 2 manuscript. Please deposit any pilot data that you may have already collected. Pilot data should be made accessible for peer-review, but can be placed under embargo until Stage 2 acceptance.

# Code availability

For Registered Reports, public sharing of all code upon acceptance for publication of the Stage 2 manuscript is mandatory. Please include a statement committing to sharing all code on acceptance of your Stage 2 manuscript. The Code availability statement must be included separately from the Data availability statement. Please provide a link (e.g. GitHub, osf) to a live version of your code. Code used to simulate data, conduct power analyses, and analyse pilot data should be made accessible in the same location. The code must be made available for peer-review, but can be placed under public embargo until Stage 2 acceptance.

# Results

Do **not** include a **Results** section.

# Discussion

Do **not** include a **Discussion** section.

# References

1. Rosenzweig, C. et al. Attributing physical and biological impacts to anthropogenic climate change. *Nature* **453,** 353–357 (2008).
2. Jones, R. A. L. *Soft Machines: Materials and Life*(Oxford Univ. Press, 2004).
3. Hao, Z., AghaKouchak, A., Nakhjiri, N. & Farahmand, A. Global Integrated Drought Monitoring and Prediction System (GIDMaPS) data sets. figshare <http://dx.doi.org/10.6084/m9.figshare.853801> (2014).
4. VanderWeele, T. J., Mathur, M. B. & Chen, Y. Outcome-wide longitudinal designs for causal inference: a new template for empirical studies. Preprint at *arXiv* <http://arxiv.org/abs/1810.10164> (2019).
5. No unpublished manuscript (i.e., a manuscript that is in preparation, submitted, under review, or under revision) should be included in the reference list. Only mention such work parenthetically in the main text. No main argument or conclusion can rely on an unpublished manuscript.

# Acknowledgements

Please ensure that you acknowledge all funding sources that supported the work reported in your manuscript and provide grant or contribution numbers in an Acknowledgments section after the references. Indicate what role the funder(s) had in the conceptualization, design, data collection, analysis, decision to publish, or preparation of the manuscript. If any of this information could be perceived as a competing interest, ensure that it is also included in your competing interests statement. If the funder(s) have/had no role, please include the following statement: “**The funders have/had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.**” If no specific funding supported the work, include the following statement: “**The authors received no specific funding for this work.**” Keep other acknowledgements brief and do not include effusive comments.

# Author contributions

We require authors to include an author contributions statement of their individual contributions to the paper -- such as experimental work, project planning, data analysis, etc (see the CRediT taxonomy for relevant contributor roles: <https://casrai.org/credit/>). The statement should be short, and refer to authors by their initials. For details please see the Authorship section of our joint Nature Research Editorial policies at http://www.nature.com/authors/editorial\_policies/authorship.html

# Competing interests

We ask authors to declare both financial and non-financial competing interests. For more details, see https://www.nature.com/authors/policies/competing.html. If you have no financial or non-financial competing interests, please state so: “The authors declare no competing interests.”

# Figures

You are encouraged to include Figures in the text or at the end of the protocol. Keep in mind that a total of 8 display elements (i.e., combination of Tables and Figures) is permitted in the final, Stage 2, submission. Figures/Tables that are not essential should be included in your Supplementary Information file.

# Figure Legends

**Figure 1. Guidelines for the preparation of figure captions.** Figure captions should be concise. Begin with a brief title and then describe what is presented in the figure and detail all relevant statistical information. If you show pilot data, list the N of each plot and report full statistics.

# Table 1. Design Table

You must include this mandatory **Design table**. The columns are prescribed; the number of rows will depend on the number of research questions you will address in your Registered Report.

* Ensure that there is an **exact** correspondence between each scientific hypothesis and each statistical test. For example, it is not appropriate to write: Condition A will affect performance differently from Condition B. Instead, you must define the performance measure (e.g. Reaction Time) and the predicted direction of the difference. This would translate to, e.g.: Reaction times will be significantly higher in Condition A than Condition B.
* If your analysis strategy will depend on the results (e.g. normal vs. non-normal distribution) then specify the contingencies for making different choices, i.e. IF-THEN statements.
* You cannot interpret lack of evidence for the existence of an effect in NHST (e.g. a p>0.05 in a t-test) as evidence for the absence of an effect. To be able to interpret null results, you must commit to using Bayes Factors or equivalence testing.

| **Question** | **Hypothesis** | **Sampling plan (e.g. power analysis)** | **Analysis Plan** | **Interpretation given to different outcomes** |
| --- | --- | --- | --- | --- |
| 1. Aggregating across tasks, are participants aware that they are being influenced by the heuristics or biases? | If participants are aware that they are being influenced by the heuristics or biases, this predicts that, within the factor-included group, participants will on average report more than 0 effect on the scale from -10 to 10. (H1a)  If participants are unaware that they are being influenced by the heuristics or biases, this predicts that, within the factor-included group, participants will on average report 0 effect on the scale from -10 to 10. (H1b) | Two-hundred participants will be recruited from the online platform, Prolific. **[add power analysis here]** | Bayesian one-sample t-test on whether average introspection report is greater than 0. We will find the posterior probability. | A posterior probability greater than 0.95 will suggest that participants are aware that they are being influenced by the heuristics or biases.  A posterior probability less than 0.05 will suggest that participants are unaware that they are being influenced by the heuristics or biases. |
| 2. Does that awareness come from actual experience of the heuristic or bias, or is it attributable to lay theories? | If heuristics are conscious, that predicts participants in the factor-included group will report more influenceof the bias than participants in the factor-excluded group (H1a) if heuristics are not, that predicts no significant difference in reported influence between the two groups (H1b) |  | Bayesian two-sample t-test on whether average introspection reports differ between the factor-excluded and factor-included groups. We will find the posterior probability. |  |
| 3. Are participants who are more impacted by a bias consciously aware of the extent to which they are affected by heuristics and biases? | If the extent of influence is conscious, that predicts participants’ reports of how affected they were by the bias (and in which direction) will correlate with the extent and direction of their individual effect (H2a). If the extent of the influence is not conscious, that predicts a lack of such correlation (H2b). |  |  |  |
| 4. Is there variation in conscious access between these heuristics and biases? | **insert** |  |  |  |
| 5. Are there individual differences that generalize across heuristics and biases in the ability to consciously access the extent to which participants are affected by heuristics and biases? | If there are individual differences, this would predict that an individual’s accuracy in reporting the effect of one heuristic or bias will correlate with their accuracy in the other heuristics and biases (H4a). If there are no individual differences, this would predict no such correlation (H4b). |  |  |  |
| 6. Are participants accurate in their perception of their self-knowledge? | If participants are accurate in their perception of their self-knowledge *within this study*, this also would predict that participants’ reported confidence levels for their reports of bias influence will correlate with their accuracy for that report (H5a). If participants are not accurate, this would predict no such correlation (H5b).  If participants are accurate in their perception of their self-knowledge *more broadly*, this would predict that overall accuracy in reporting the effect of heuristics and biases will correlate with perceived self-insight (as measured by the Self-reflection and Insight Scale) (H5c). If participants are not accurate in their perception of their self-knowledge, that would predict no such correlation (H5d).  If participants who are interested in introspection are more accurate than those who are not, this would predict that overall accuracy in reporting the effect of heuristics and biases will correlate with reported interest in and frequency of self-reflection (as measured by the Self-reflection and Insight Scale) (H5e). If participants who are interested in introspection are no more accurate than those who are not, this would predict no such correlation (H5f). |  |  |  |

# Supplementary information

Please report pilot data in detail here and include any other material that provides background information.

**Power**

100-200 for continuous

100 for dichotomous. 80% power on effect size 0.4

250 total given 25% exclusion

highlighted are not online

1. Anchor: Mussweiler & Strack, 1999
2. Association: Roediger & McDermott, 1995
3. Availability: Tversky & Kahneman, 1973
4. Belief: Evans et al. 1983
5. Causal Inference: (Morris et al., 2019)
6. Decoy Effect (Kaptein et al., 2016)
7. Halo (Palmer & Peterson, 2015)
8. Hindsight (Groß et al. 2023)
9. Illusion of Truth: Ly et al, 2024
10. Imaginability (Ceschi et al. 2019)
11. Mere Exposure: Stang (1974)
12. Omission: (Cushman et al. 2006)
13. Recognition: (Pachur et al. 2009)
14. Reference Price: Thaler, 2008
15. Representativeness: Gigerenzer et al. 1988
16. Status Quo: Miceli master’s thesis 2020 <https://scholarworks.calstate.edu/downloads/s4655p04g>
17. Sunk Cost: (Petrov et al. 2023)

| **Task name** | **citation** | **validated in Sofia’s thesis** | **validated summer 2024** | **Ceschi** |
| --- | --- | --- | --- | --- |
| Anchor | Mussweiler & Strack, 1999 | yes | yes | Anchoring and adjustment |
| Association | Roediger & McDermott, 1995 | yes | yes |  |
| Availability | Tversky & Kahneman, 1973 | yes | yes | Mindware gaps |
| Belief | Evans et al. 1983 | yes | yes |  |
| Causal Inference | (Morris et al., 2019) | yes | no p-value = 0.8861 |  |
| Decoy Effect | (Kaptein et al., 2016) | yes | no p-value = 0.5 |  |
| Halo | (Palmer & Peterson, 2015) |  | yes |  |
| Hindsight | (Groß et al. 2023) |  | no |  |
| Illusion of Truth | Ly et al, 2024 |  |  |  |
| Imaginability | (Ceschi et al. 2019) |  |  | Mindware gaps |
| Mere Exposure | Stang (1974) | yes | yes |  |
| Omission | (Cushman et al. 2006) |  |  |  |
| Recognition | (Pachur et al. 2009) |  | yes |  |
| Reference Price | Thaler, 2008 |  | no p-value = 0.4913 | Anchoring and adjustment |
| Representativeness | Gigerenzer et al. 1988 |  | yes | Mindware gaps |
| Status Quo | Miceli master’s thesis 2020 <https://scholarworks.calstate.edu/downloads/s4655p04g> |  | no p-value = 0.5 |  |
| Sunk Cost | (Petrov et al. 2023) |  |  | Valuation bias |

Best practice to test for significance/strength of null & positive effects in brms? Is it bayes factor?

What should we use for priors?

Absolute value distance from the anchor on y axis.

Predicting Distance from anchor

Based on factor presence, which task,

Individual number in equation: how much am I affected by H&B. as constant in model

## Graveyard

1. Review of literature on introspective accuracy
   * longstanding view that our ability to introspect is limited and our feelings of introspection are illusory: N&W, Gigerenzer, Haidt, etc.
   * recent work showing a more complicated story: Newell & Shanks 2014, Morris, Carlson
   * open questions: limits of introspection, mechanism of introspection, trainability
   * finding answers relies on first creating and validating measures and paradigms that will allow us to study introspection
2. In this study, we develop a measure of introspective accuracy and use it to better understand individual differences in introspective accuracy and the cognitive processes involved in introspection.
3. Flaws in previous measures
   * Self-reports are unreliable
     + e.g. self-reflection and insight scales
     + if only it were that easy! History of studies going back to N&W show we are unreliable at reporting whether or not we accurately know the mental events and processes in our minds
   * subtle but important point: accuracy itself does not imply introspection. (Bem, Gopnick, Carruthers, Haidt, etc.) We could be relying on theory of mind, folk psychology… This is why our measure includes a control group who are only making use of those processes to deduce mental processes
     + Vocabulary: distinction between [Perception of Internal Mental Events (PIME) or Perception Related to Internal Mental Events? Or should I use Sofia’s acronym for continuity?] and [Inferring Internal Mental Events]. Both PIME and IIME could potentially be accurate or inaccurate.
   * This project directly inspired by studies described in N&W
     + using social-psychological paradigms that have not been replicated. We selected tasks whose effects have been recently replicated in large online samples
     + Tasks did not allow for degrees of introspective access – only its absence or existence. Our tasks use a scale to capture gradients between all-or-nothing
   * flaws in metacognition research
     + **[what do we want to say here? Just that they are looking at confidence judgement accuracy and confidence metacognition is a separate object of study than introspection into how we are affected by heuristics and biases? Some expansion of your Gorillas argument that this is just an “abstracted feature” of our cognitive processes?]**
4. Why use heuristics and biases as the effect participants are introspecting on?
   * **[could we discuss this reasoning some more?]**
   * H&B as traditionally unconscious
     + Kahneman, 2011; Stanovich et al. 2008 (who, in fact, point out that ‘bias blind spot’ is precisely the problem of assuming incorrectly that our biases accessible to conscious introspection)
   * The same heuristics can be used consciously and unconsciously, implying that this might be a process that could be consciously introspected on **[is this reasoning logical?]** (Gigerenzer, G., & Gaissmaier, W. 2011)

Further, if this variation is cognitively meaningful, this would predict that the variation would cluster into sets of heuristics and biases that the same individuals tend to be more or less accurate in reporting (H3c). If the variation is not cognitively meaningful, this would predict no such clustering (H3d).