

Supplementary Information for

Hypothetical nudges provide directional but noisy estimates of real behavior change

Linnea Gandhi, Anoushka Kiyawat, Colin Camerer, & Duncan J. Watts

I. Pre-Registration Deviation Table	2
II. Supplementary Methods: Hypotheticals in the Nudge Literature	4
A. Prevalence of Hypothetical Nudge Experiments in 2022 Publications	4
B. Distribution of Hypothetical Experiments Across Domains	5
C. Comparison of Hypothetical and Real Effects in Past Literature	5
D. Why Researchers Use Hypotheticals	6
III. Supplementary Methods: Hypothetical Experiments	9
A. Identifying Target Experiments	9
B. Constructing Hypothetical Designs	10
C. Sampling Plan	17
D. Recruitment	17
E. Use of Attention Checks	19
F. Representativeness of Nudges	19
IV. Supplementary Methods: Hypothetical Experiments	21
A. Study Demographics and Balance Checks	21
B. Summary Statistics	22
V. Supplementary Methods: Regressions	24
A. Effect of Hypotheticals on Behavior	24
B. Effect of Hypotheticals x Conditions on Behavior	25
C. Effect of Setup x Descriptors x Conditions on Behavior—Hypothetical Data Only	26
D. Effect of Setup x Descriptors x Conditions on Behavior—Hypothetical and Real Data	28
VI. Supplementary Methods: Evaluating Treatment Effect Magnitudes	30
A. Rates of Under- or Overestimation	30
B. Sample Size Calculations	32
C. Replication Tests	33
D. Hypothetical effects as “less extreme” than real-world counterparts	36
VII. Supplementary Methods: Imaginability	38
VIII. Supplementary Methods: Survey Materials	39
A. Common Consent and Attention Check for All Five Surveys	39
B. Common Ending Questions (Demographics) for All Five Surveys	40
C. Survey for Consumer Choice Study	42
D. Survey for Finance Study	46
E. Survey for Health Study	51
F. Survey for Sustainability Study (two versions)	55
G. Survey for Transportation Study	61
IX. Supplementary References	65

I. Pre-Registration Deviation Table

(adapted from Willroth & Atherton, 2023)

Table S1. Pre-registration deviations with explanations.

Pre-registration deviations						
#	Details		Original wording	Deviation description	To what extent is this a deviation from the pre-registered plan?	Judgment of impact
1	Type	Analyses	Excerpt from the original sustainability study pre-reg: “... To enable comparability with our one-shot hypothetical scenarios, we focus on the impact of the first intervention on the first day only (Dec 15), at the cluster (student room) level, using the average temperature in a room (on Dec 15) as our dependent variable. ...In the real dataset, this is the average temperature in a room on the first treatment day (Dec 15). Each room is either in the treatment or control group. ...In the hypothetical dataset, this is the temperature that participants type in after the prompt.”	To enable all analyses to be on binary outcomes as well as better honor the original study design, we re-ran our sustainability study by evaluating whether participants turned down their thermostat, or not, before winter break. Excerpt from new pre-reg: “...we focus on the impact of the intervention on the average temperature in a room right after exams end, at the point at which setpoints stabilize–Dec 23..... To stay consistent with the binary analyses of our other datasets, we use whether or not that average temperature is at or below 68 (the target temperature in the intervention email) as our dependent variable. ...In the real dataset, this is driven by the average temperature in a room on Dec 23. Each room is either in the treatment or control group. ...In the hypothetical dataset, this is driven by the temperature a participant decides to set the room to before (hypothetically) leaving for winter break.	Minor	We believe this is of low risk as (i) we pre-registered our new design and (ii) results using the continuous variable are included here for transparency. Note, this change in the way the study was run <i>does</i> change the results—but we expected it to. Any Day 1 differences should be small, in the real data, versus big by the time students have left for winter break. Similarly, our control participants prior had no reason to turn their thermostat down, without the winter break timing.
	Reason	Other				
	Timing	After results known				

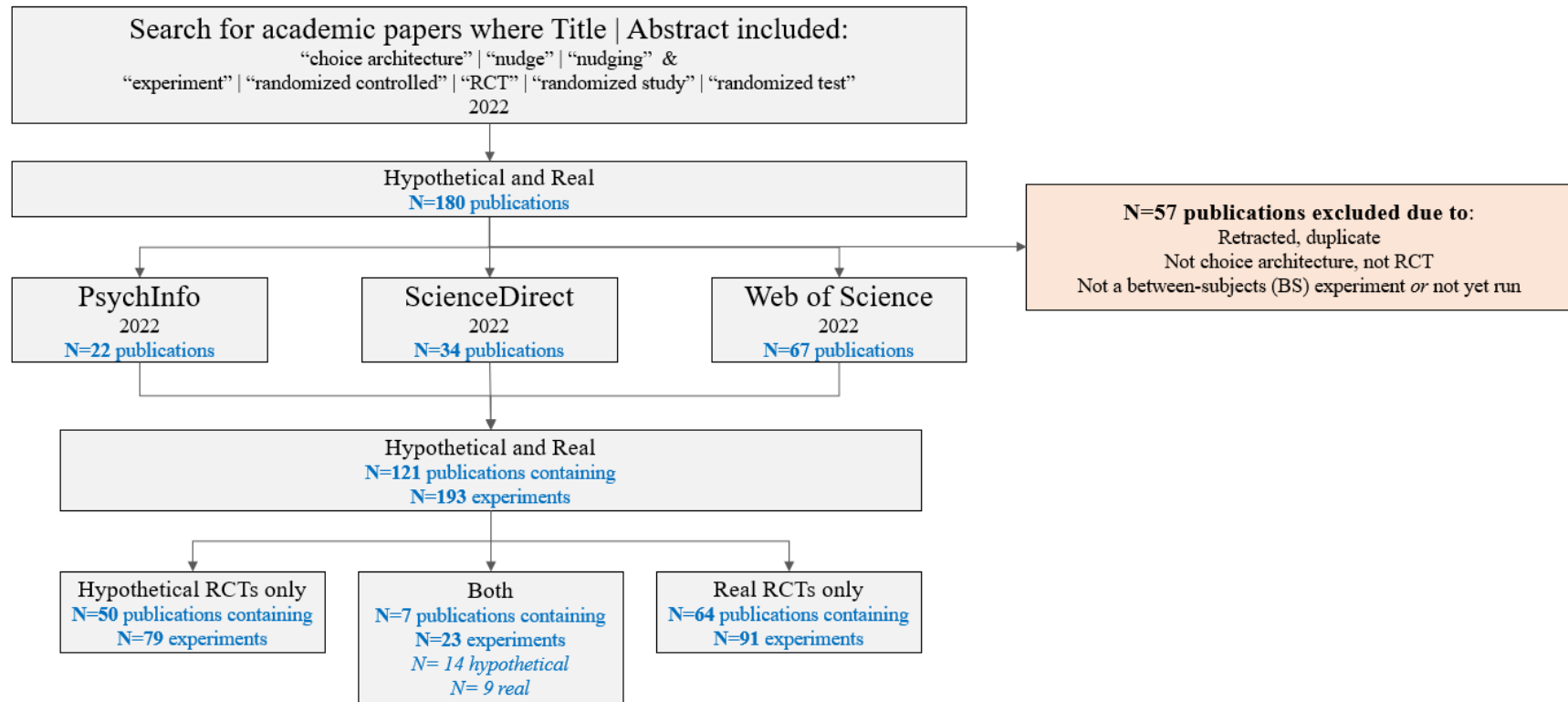
2	Type	Analyses	Excerpt from the Consumer Choice study - nearly identical across our 5 studies. <i>“1) Regression on individual participant data: First, we will run a linear probability model (and, for stopping, a logistic model) using the data from our hypothetical conditions. Our p-value cutoff will be $p=0.05$. Stop (or WTP) ~ Condition + Setting + Descriptors + Condition x Setting + Condition x Descriptors + Setting x Descriptors”</i>	In writing up the manuscript and in feedback on early drafts, we realized we needed additional analyses to highlight key contrasts as well as statistically examine patterns across study domains: (i) The same regression but dropping the real-world data, to highlight hypothetical contrasts. (ii) The same regression, simplified by pooling across hypothetical designs. <i>DV ~ Condition x Hypothetical</i> (iii) The same regression, pooled across hypotheticals and conditions. <i>DV ~ Hypothetical</i> (iv) Binomial tests evaluating whether the rate at which (all 20) hypothetical experiments (a) accurately estimated the direction of effects and (b) overestimated (vs. under-estimated) treatment effects differed from chance. Pre-registrations were domain-specific, so we had not pre-registered cross-domain analyses.	Major	These are major additions to the main text, but we do not consider these high risk deviations, as they simply provide higher-level or pooled views of the same data. They do not change the overall conclusions of the investigation. Further, we appropriately indicate these specifications were not pre-registered, for transparency. We did pre-register (i)-(iii) when re-running the sustainability study.
	Reason	Other				
	Timing	After results known				
3	Type	Methods	Excerpt from the Finance study: <i>“We will recruit 3000 individuals to participate.”</i>	We collected an additional 600 participants because of a math error in our original calculation. We estimated 350 per control and 350 per the treatment (fresh start) condition, and to calculate how that played out across the two cohorts following the original paper’s randomization, we estimated $N=380$ for cohort 1 and $N=530$ for cohort 2, for total $N = 910$. We scaled that to 3000 to be safe, as it seemed low; only to realize (after submitting our pre-registration but before data collection was complete) that we had forgotten to multiply by four for the number of hypotheticals needed. We needed $900*4 = 3600$, not 3000.	Minor	Low risk - we collected 20% more participants and followed our original logic, merely correcting a math error. The main statistic this impacts is the standard error around the hypothetical effect size estimate, widening it and making it harder to find significant differences.
	Reason	Typo / Error				
	Timing	During data collection				

II. Supplementary Methods: Hypotheticals in the Nudge Literature

A. Prevalence of Hypothetical Nudge Experiments in 2022 Publications

We draw on a systematic review by Szaszi et al. (2018) to look at the prevalence of hypothetical nudge experiments in the historical literature.¹ For an updated view, we conducted our own search for the year 2022, drawing on three search engines. Our process is visualized below in Figure S1.

Figure S1. A flow chart in the style of PRISMA visualizing the prevalence of hypothetical experiments in the nudge literature, in 2022.



¹ Nudges are a subset of choice architecture interventions, excluding changes to strict rules (e.g., forbidding options), available options (e.g., removing options), and incentives (e.g., adding a substantial bonus to an option) (1). The interventions of our 20 experiments are technically nudges, though the behavior change literature from which we selected them, as well as the Szaszi et al., 2018 review, covers “choice architecture” more broadly. (See SI section 3a for our selection process.) We have no reason to believe our findings apply only to nudges and not to choice architecture more generally.

B. Distribution of Hypothetical Experiments Across Domains

Before selecting our target experiments, we evaluated the prevalence of hypothetical experiments in the past literature as collected by Szaszi et al. (2018), reviewing original materials where available and prose descriptions of the materials when not. While the review lists 52 hypothetical experiments, two appear to be miscoded and are instead a stated intent and a field experiment, respectively, which do not fit our definition of hypothetical. We use the corrected $n=50$ in our analysis.

Table S2 shows the percentage of experiments within a given domain category where the dependent measures were hypothetical behaviors.

Table S2. A breakdown of experiments from the Szaszi et al., 2018 database.

Domain	Across RCTS, % that were hypothetical	Across hypothetical RCTs...			
		1+ pages or customization (‘complex’)	specific, non- generic details (‘specific’)	cheap talk or certainty statements	original materials available
Finance	11 of 14 (79%)	0	0	0	1
Transportation	5 of 9 (56%)	0	0	0	3
Sustainability	16 of 29 (55%)	1	13	0	9
Consumer choice	4 of 16 (25%)	1	2	0	0
Health	12 of 67 (18%)	5	7	0	5
Prosocial	1 of 12 (8%)	0	0	0	1
Other	1 of 3 (33%)	0	0	0	0
<i>Total</i>	50 of 156 (32%)	7 of 50 (14%)	22 of 50 (44%)	0 of 50 (0%)	19 of 50 (38%)

As shown above, most hypothetical studies in the 2018 database by Szaszi et al. employed simple, generic, non-customized setups. Further, none of the studies employed strategies recommended in the economics literature (Champ et al., 2009), such as a “cheap talk” script (i.e., describing hypothetical bias and asking participants to avoid it) or certainty question (“How certain are you that you would actually do [the behavior]?”). Accordingly, when selecting what features of hypothetical designs to test in the present project, we focused on contrasting levels of complexity and specificity—features recommended by the psychological vignette literature but not apparently utilized. Future work exploring the additional use of cheap talk or certainty statements would be fruitful.

C. Comparison of Hypothetical and Real Effects in Past Literature

We could not directly compare the treatment effects from real and hypothetical studies in Szaszi et al., (2018) as that author team did not record the numeric results from each experiment. However, we could by examining another recent meta-analytic database by Mertens et al., (2022), which included summary

statistics for each condition from each study. This database did not include an indicator for whether a study was hypothetical or not, so we trained an RA to review each underlying study for whether it used a hypothetical setting or a hypothetical behavior. (All studies with hypothetical behaviors also had hypothetical settings; three studies had hypothetical behaviors within real settings.)

Across treatment effects based on binary and continuous data ($N=447$), 27.5% were hypothetical in some way; 26.9% of studies relied exclusively on hypothetical effects. Among only those based on binary outcomes ($N=254$), 29.1% were hypothetical in some way; 27.7% of studies including binary outcomes relied exclusively on hypothetical effects.

We then ran a simple statistical test comparing the average lift (difference between treatment and control condition uptake) across experiments that were hypothetical versus real, using both linear (t -test of two independent means) and logistic (z -test of two independent proportions) specifications. No statistically significant difference emerged across any specification. We report the breakdown in Table S3. The first specification (hypothetical setting, z -test) is referenced in the main text.

Table S3. A breakdown of the experiments from the Mertens et al. (2022) database with binary outcomes.

Coding Approach	Study Type	Average lift (treatment minus control)	Number of experiments	Statistical comparison	
				z -stat (p -value)	t -test (p -value)
Hypothetical Setting	Real	0.129	177	$z = 0.155$	$t = 0.154$
	Hypothetical	0.137	74	($p = .877$)	($p = .878$)
Hypothetical Behavior	Real	0.129	178	$z = 0.202$	$t = 0.199$
	Hypothetical	0.138	73	($p = .840$)	($p = .842$)

Note: All studies with hypothetical behaviors also had hypothetical settings; three studies had hypothetical behaviors within real settings. Of the $N=254$ binary hypothetical effects, only $N=251$ had data available to calculate lift.

For completeness, we also coded the techniques used within these hypothetical studies, as seen in Table S2 with the Szaszi et al., 2018 database. 34.0% of hypothetical experiments with binary outcomes used a complex setup, 13.2% used specific descriptors, only 1.9% used a cheap talk script or certainty statement, and 67.9% had readily available materials. Similar prevalence appears when looking at experiments with continuous or binary outcomes: 37.8%, 17.8%, 2.2%, and 67.8% respectively.

D. Why Researchers Use Hypotheticals

We provide here several examples of the various ways in which researchers use hypothetical designs (e.g., for existence proofs, unpacking mechanisms, and making recommendations for real-world policies), as least as mentioned directly in their published manuscripts. These come from papers in Mertens et al., (2022) which we identified (above) as relying on hypothetical designs. A more thorough survey is beyond the scope of the present work.

- **Barnes et al., 2021 rely exclusively on hypothetical scenarios to simulate choices in health insurance markets. They manipulate choice designs to (a) demonstrate the existence of a**

potential treatment effect and (b) extrapolate to policy: *“In conclusion, millions of Americans face the difficult task of choosing health insurance in the ACA marketplaces. How these markets are designed has a crucial impact on consumers’ ability to compare and choose health plans that provide adequate risk protection and are affordable. Our results indicate that a simple, cheap, and policy feasible mechanism—namely, sorting by and highlighting the total estimated cost in a prominent place—could help improve the quality of health insurance choices, saving money for consumers and the government alike. Our data also reveal that including additional educational material does little to improve consumers’ choices, suggesting the need for other strategies to make further improvements in consumer decision-making. Such findings can be extended to improving insurance choices for consumers shopping in Medicare managed care markets, stand-alone Medicare drug plan markets, and employer-sponsored markets as well.”* **They further note:** *“The study was hypothetical by nature and may not capture the complexity of real life health insurance decision-making. Nonetheless, empirical evidence, such as that provided by our three studies, is necessary to inform policy makers on how to optimize designs of health insurance markets for consumers.”*

- **Bacon et al., 2019 run only online hypothetical scenarios, varying menu design. They use this as an existence proof and draw direct implications to policy:** *“Overall, the findings from this research suggest that, even if certain restaurant menus can encourage pro-environmental food choice for infrequent vegetarian eaters, they can also backfire for frequent vegetarian eaters and have an undesirable impact on food selection. Our experiment therefore points out that any contextual interventions aimed at nudging pro-environmental behavior need to be carefully examined in relation to people’s past eating choices to avoid undesirable behavioral effects, and suggests that achieving sustainable eating may require more personalized interventions.”* **They further note:** *“One of the limitations is that the experiment was conducted online rather than in a real restaurant. We implemented a “restaurant scenario” task in the experimental procedure to minimize the disadvantages of conducting our research online and to make the food choice more convincing. More precisely, we asked participants to imagine a scenario that was supposed to influence them to adopt a mental state like the one they would experience in a real restaurant.... To make this scenario easier to imagine, we also presented participants with an image of a cozy table in a restaurant. Given that some other impactful menu studies (e.g. Liu, Roberto, Liu, & Brownell, 2012) were also conducted online and reported to obtain similar results to experiments conducted in naturalistic locations, previous research indicates that the online mode of administration should not be considered a serious disadvantage.”*
- **Ansher et al., 2014 run a series of hypothetical case studies with medical students, varying opt in or opt out defaults. From these scenarios, they conclude:** *“The defaults used in electronic order sets influence medical treatment decisions when the consequences to a patient’s health are low. This pattern suggests that physicians cognitively override incorrect default choices but only to a point, and it implies tradeoffs that maximize accuracy and minimize cognitive effort. Results indicate that defaults for low-impact items on electronic templates warrant careful attention because physicians are unlikely to override them.”* **They further note:** *“Although our use of case scenarios could also arguably affect the external validity of our findings, we doubt this would be the case when considering the expansive literature on default effects on real-world behavior, including physicians’ decision making.”*

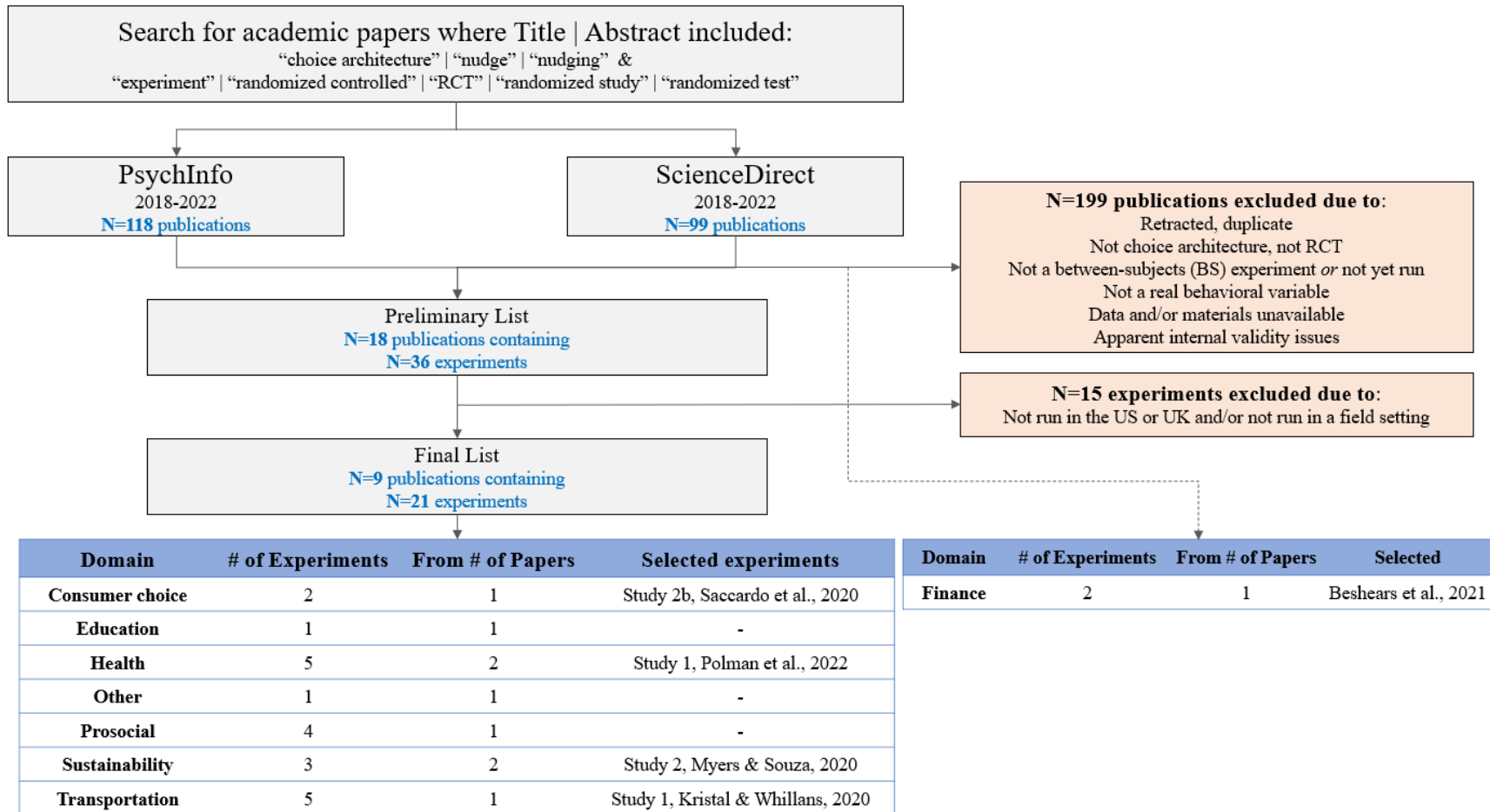
- **Abhyankar et al., 2014 run a computer based task with patients making hypothetical choices about a real-world medical trial. They evaluated the hypothetical impact of choice design, and conclude:** *“...we expect that these results would be replicated in the real world and in other contexts, because the study explores how an individual’s construction of a decision problem is influenced by the presentation of options, rather than the evaluation of the information contained within the decision problem. It is likely that the same metacognitive processes would be used by individuals whether or not they were patients...”*. **The team further acknowledges potential limits:** *“There are ethical concerns about carrying out this type of research in a sample of patients making actual trial participation choices...In this applied context, we need to have confidence that any manipulation, at the least, causes no additional harm and may even benefit the patient making the choice. This study is therefore carried out in a sample of healthy women making a hypothetical choice about trial participation but using information from a real-world cancer treatment trial. This study is expected to provide some proof-of-concept data, as in a phase II trial addressing whether these framing effects influence health care choices...Third, the framing bias may be greater or smaller depending on the values and experiences of the individuals, so studies of people making real-world versus hypothetical decisions may find different effect levels...”*
- **Goldstein et al., 2011 run a field test with reciprocity as an intervention. Then they use hypothetical scenarios to unpack potential psychological mechanisms of the treatment effects, which was impossible to do in the real-world setting:** *“Because it was not possible to ask the participants additional questions, we collected additional data from a different group of participants to examine the psychological mechanism that might underlie the effects demonstrated in the field experiment.”*
- **Soon et al., 2018 run a series of vignette studies with clinicians, testing interventions to nudge better decisions around treatment of lower back pain. They conclude:** *“Our study suggests that embedding a default option nudge into clinical decision support tools can reduce the choice of low-value care options for low back pain. This has the potential to significantly improve patient care while preserving autonomous clinical decision-making.”* **Further, they note:** *“It is reasonable to ask whether findings from hypothetical vignettes can be fully indicative of clinicians’ behaviour in a ‘real world’ setting. Nonetheless, vignette-based studies are widely used and accepted in research into medical choices.”*

Notably, there is less evidence around the use of hypotheticals for power calculations. It may be that when researchers run hypotheticals as pilots in advance of field studies, in order to estimate effect sizes in ex ante power calculations, those do not always or often get reported in the final manuscript. Measuring the prevalence of this usage is also subject to publication bias. If a hypothetical pilot is run and is unsuccessful, there is no reporting and no subsequent manuscript. If a hypothetical pilot study is run and is successful, the followup field experiment may only be reported if it, too, is deemed successful. And this is often likely to not be the case, as warned by Albers & Lakens (2018) about using pilot studies to inform power calculations in general: “If researchers only follow up on pilot studies when they observe an effect size in the pilot study that, when entered into a power analysis, yields a sample size that is feasible to collect for the follow-up study, these effect size estimates will be upwardly biased, and power in the follow-up study will be systematically lower than desired.”

III. Supplementary Methods: Hypothetical Experiments

A. Identifying Target Experiments

Figure S2. A flow chart in the style of PRISMA visualizing how we arrived at our target studies.



First, we identified candidate experiments evaluating choice architecture techniques on real behaviors that (i) were published in 2018-2022, (ii) shared their intervention materials and data, (iii) analyzed at least 100 individuals per condition, and (iv) had no apparent internal validity issues (Simonsohn et al., 2022). 36 experiments from 18 papers met these criteria. We filtered to those run in the United States or the United Kingdom—countries with large online samples—and conducted in field settings, leaving us with 21 experiments from eight papers. We selected four experiments, one each from four papers representing distinct domains—sustainability (Myers & Souza, 2020: study 2), health (Polman et al., 2022: study 1), transportation (Kristal & Whillans., 2020: study 1), and consumer choice (Saccardo et al., 2020: study 2b)—where a one-shot online setting could reasonably approximate the intervention deployment and behavioral measure. We ensured our focal domains included those where our field has historically relied more on hypotheticals (see SI Table S2).

Given our field’s relatively heavy reliance on hypothetical scenarios in the domain of finance (as noted in the introduction), it was unsurprising that none of the 21 experiments we could choose from came from that domain. In order to include a finance study, we relaxed our inclusion criteria to include papers where individual level data were not publicly available. Five experiments from three papers met this relaxed version of our criteria and related to the domain of finance. Four experiments from two papers (Roll et al., 2019; Roll et al., 2020) evaluated how well choice architecture interventions embedded in TurboTax, a tax filing software, could encourage low-to-moderate income tax filers to save more of their tax refunds. We chose not to use these, given we could only simulate the focal interventions and not the step-by-step filing process in which they were embedded. This left us with one finance experiment (from Beshears et al., 2021; study 1) evaluating employer-sent letters encouraging engagement in retirement savings—an intervention we judged easier to fully simulate online.

Table S4 describes the five experiments we chose as our target, real-world experiments from which to derive hypothetical scenarios. All interventions in these experiments are technically “nudges”, a subset of choice architecture techniques that neither forbid options nor substantially change economic incentives.

B. Constructing Hypothetical Designs

We then constructed four hypothetical scenarios to represent each target experiment in an online setting, following a 2x2 factorial design that varied setup (simple or complex) and descriptors (generic or specific). As described in the main text, we focused on varying features that were recommended as best practices in decision-making research, but not always employed.

The “simple” setup had no personalization or images, fit on a single screen, and covered only the immediate behavioral context. In contrast, the “complex” setup had two personalization questions and two to four contextual images, flowed across four to five screens, and provided contextual build-up. The contrast between “generic” versus “specific” descriptors was more subtle; we either obscured or included proper nouns that identified the exact location of the study or the companies providing study-specific materials.

Figure S3 shows the template. Figure 2 in the main manuscript shows the design for the consumer choice studies. Figures S4-S7 show the designs for the four other domains. We re-ran the sustainability study (Figure S6) to turn it to a binary variable and better reflect the winter break timing of the behavior.

Table S4. Below are the five real-world experiments from 2020-2021 for which we constructed hypothetical scenarios.

Paper	Domain	Experiment Description	Target Behavior(s)	Intervention(s)
<i>Nudging generosity in consumer elective pricing</i> (Saccardo et al., 2020)	Consumer Choice	Study 2b: Researchers set up a donut stand on the University of California San Diego’s campus. They varied the language on the sign, either saying “Pay What You Want” or “Pay What You Can” for donuts from Krispy Kreme. They measured who stopped at the stand for a donut and how much they paid. We focus on the former binary variable. (In study 4, the authors also run a simple hypothetical similar to study 2b, only measuring the latter outcome variable).	Stopping for a donut	Framing <i>e.g., pay what you can vs. want</i>
<i>Using curiosity to incentivize the choice of “should” options</i> (Polman et al., 2022)	Health	Study 1: Researchers approached individuals in a large university library and asked them to choose between two fortune cookies: one plain (healthier), and one dipped in chocolate with sprinkles (less healthy). With half of the participants, they added a “curiosity lure” to encourage uptake of the plain cookie. Their main outcome variable was whether the individual chose the target plain (healthier) cookie.	Choosing the healthier cookie	Curiosity lures <i>e.g., we know something about you (written in this cookie)</i>
<i>Using fresh starts to nudge increased retirement savings</i> (Beshears et al., 2021)	Finance	Study 1: Researchers sent different mailings to university employees to encourage them to take up a target retirement plan in which they were not yet invested or were underinvested. The mailers offered an option to invest more now or to do so at a delayed time (randomized frame and time). Their main outcome variable was whether employees took up the “delayed” option provided.	Choosing the delayed investment option	Fresh start effect (framing) <i>e.g., start saving in your birthday month</i>
<i>Social comparison nudges without monetary incentives: Evidence from home energy reports</i> (Myers & Souza, 2020)	Sustainability	Study 2: In this study, researchers partnered with the University of Urbana-Champaign to run an experiment with college students to encourage them to turn their thermostats down to 68°F before Winter break. Their main outcome variable was the temperature to which the thermostat was set. We turn this into a binary variable (at or below 68°F), taking the average from the day by which students have left for winter break and settings have stabilized (December 23; see Figure 7 from the original text in Myers & Souza, 2020).	Setting the thermostat to 68°F or below	Reminders <i>e.g., remember to lower your thermostat</i>
<i>What we can learn from five naturalistic field experiments that failed to shift commuter behavior</i> (Kristal & Whillans., 2020)	Transportation	Study 1: Researchers sent letters to encourage employees of a European airport (those who typically drove themselves alone) to register to carpool to work. Some employees received a letter (three treatment groups), others did not (the do-nothing control group). Their main outcome variable was whether or not employees signed up (or didn’t) for the employer’s carpooling service over the next two months.	Signing up for the company carpool program	Salience Reduced friction Peer testimonials <i>e.g., four easy steps</i>

Note: The consumer choice paper included a hypothetical design as its final study. Its design is similar to our “simple and generic” design, although it takes place in a different hypothetical location than the real-world study we compare to. We replicated this design as part of our data collection for exploratory purposes only; we do not report any corresponding analyses here.

Figure S3. Our 2x2 factorial design, varying the setup and descriptors of any given hypothetical. This figure represents the high-level design, agnostic to the field domain.

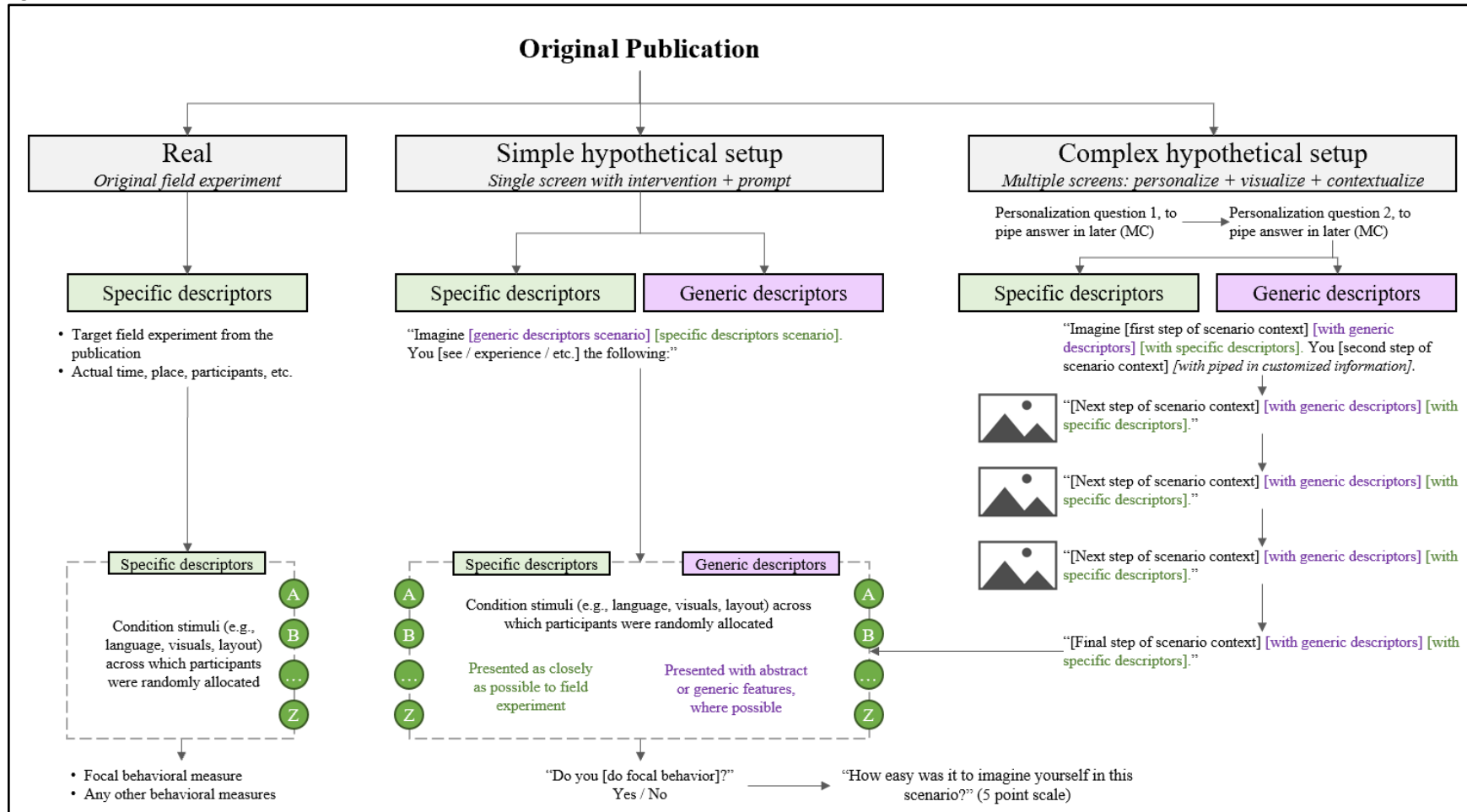


Figure S4. Our 2x2 factorial design, varying the setup and descriptors of the hypothetical version of the finance study.

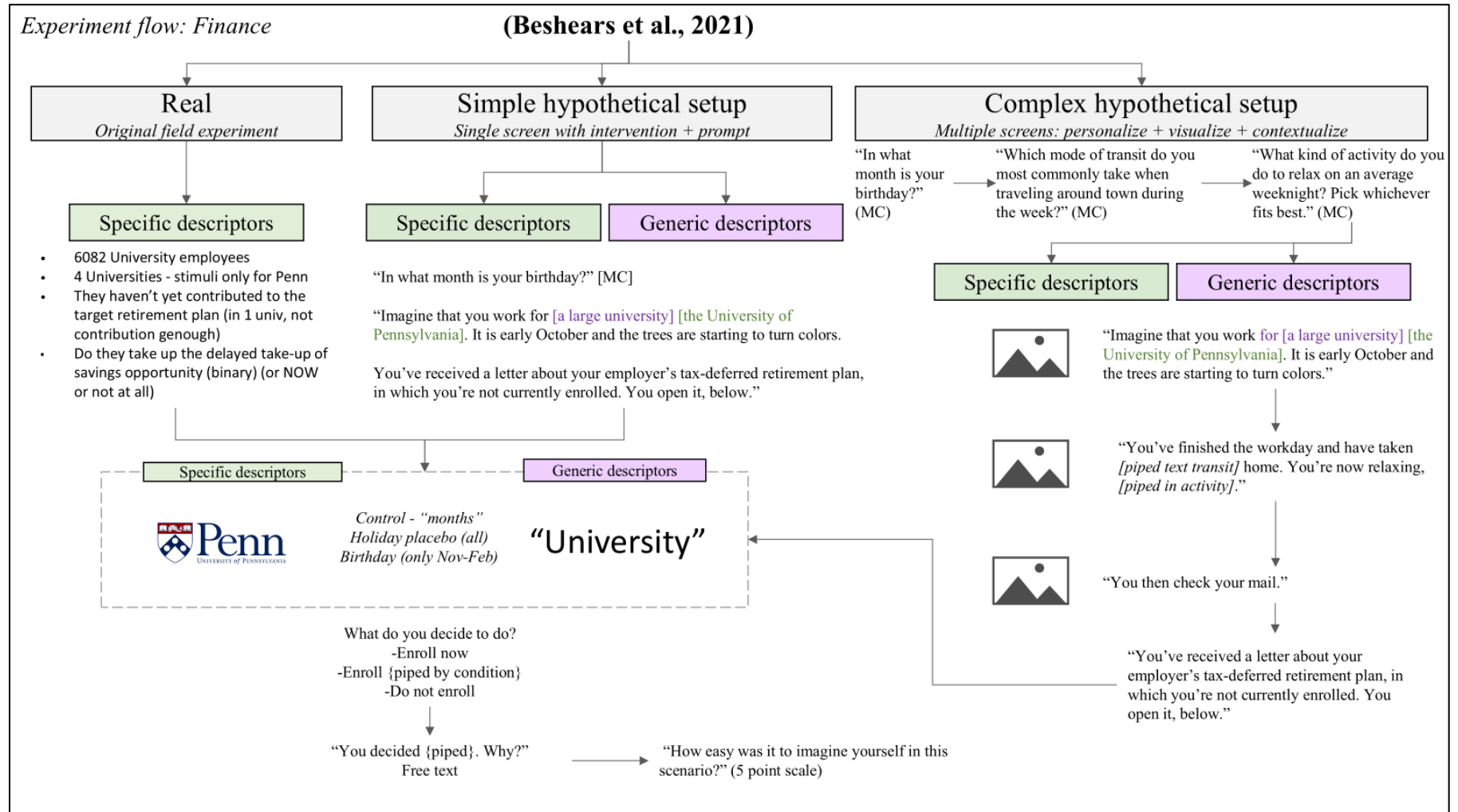


Figure S5. Our 2x2 factorial design, varying the setup and descriptors of the hypothetical version of the health study.

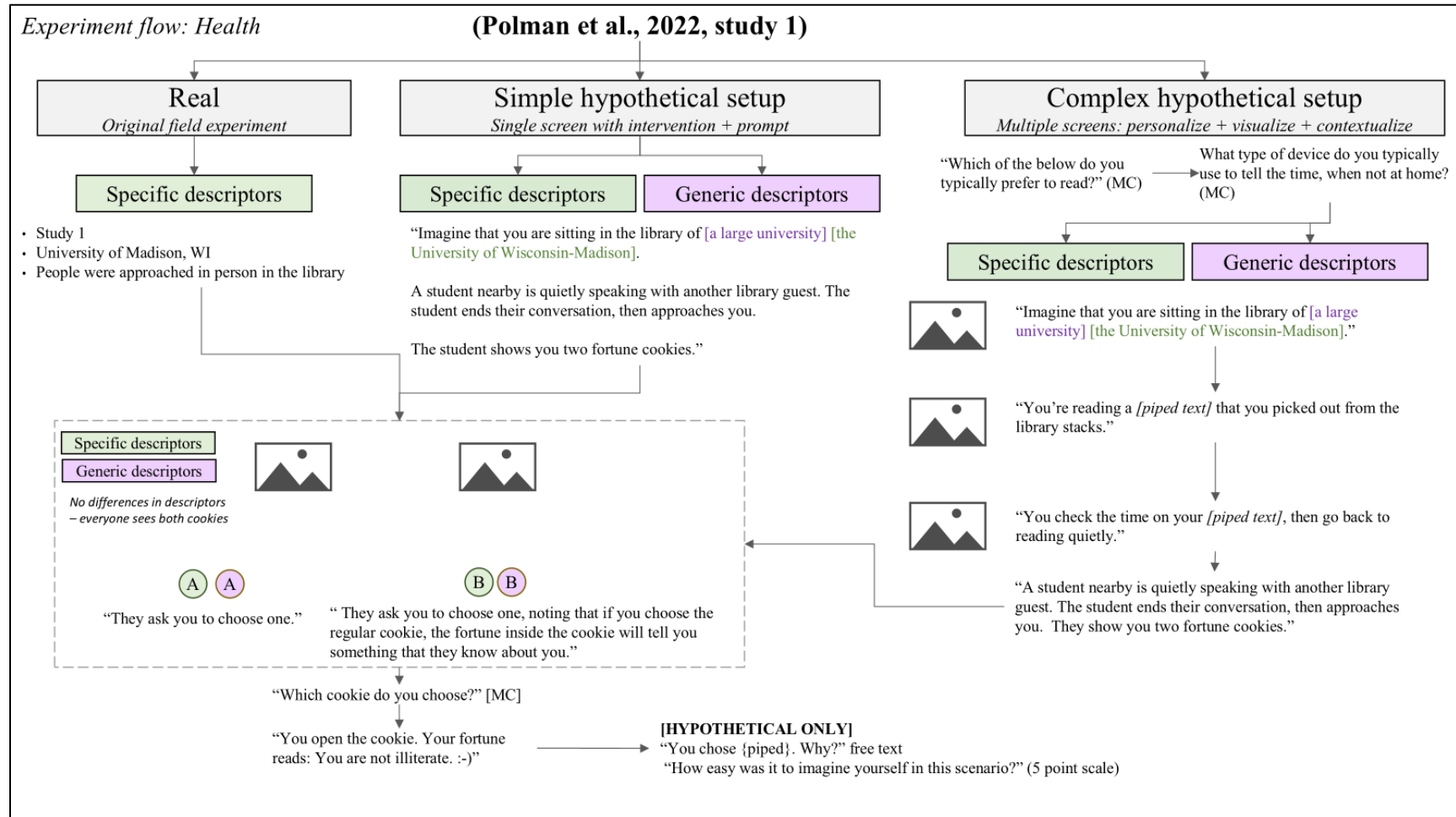


Figure S6. Our 2x2 factorial design, varying the setup and descriptors of the hypothetical version of the sustainability study. The highlighted text indicates what we added when we re-ran this study (“binary” version); the text that is struck-through is what we originally ran (“continuous” version). Both were pre-registered.

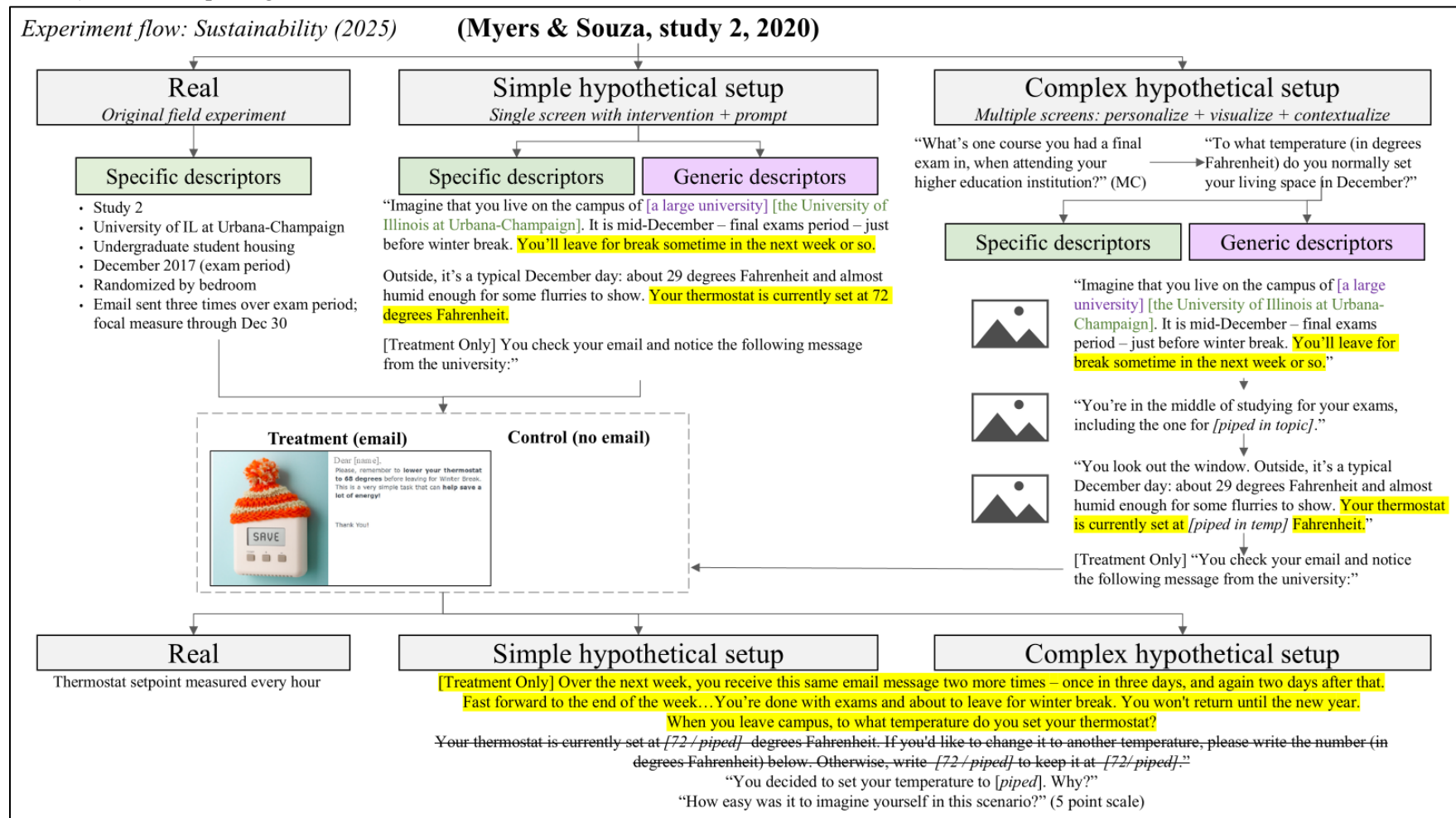
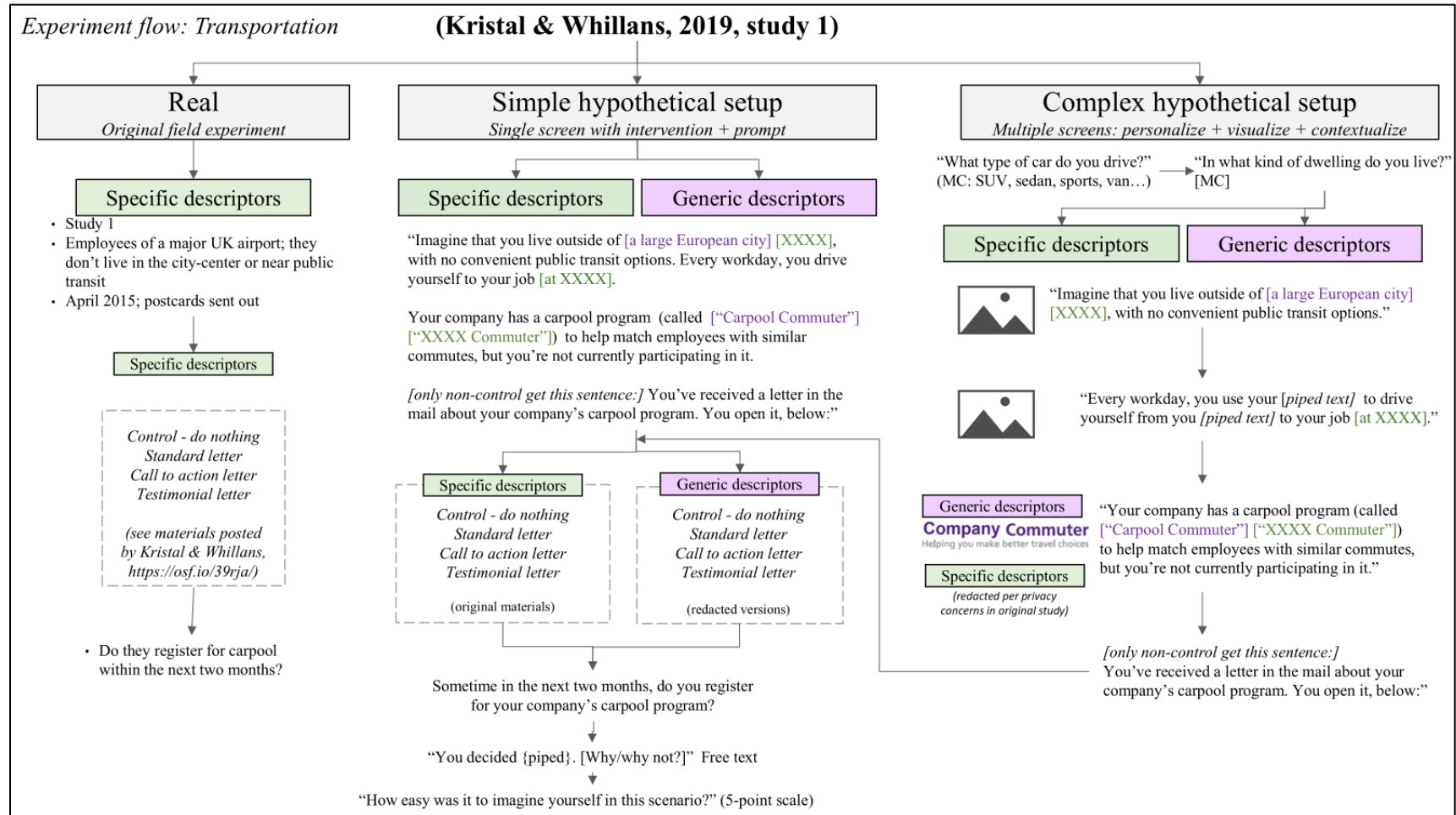


Figure S7. Our 2x2 factorial design, varying the setup and descriptors of the hypothetical version of the transportation study. Specifics are redacted per the original study's text, to keep the field partner private.



C. Sampling Plan

To determine how many participants to recruit for each collection of hypotheticals, we ran a series of power analyses. We faced two sets of considerations. First, we wanted to see how well each hypothetical variant could detect the original effect. However, a recent meta-analysis showed academic studies yield far larger effects (by a factor of 6) than practitioner field studies (Dellavigna & Linos, 2022).² Where the target sample size was prohibitive with online samples, we aimed to detect at least the latter, however imperfect a proxy. Second, four of our five sets of studies originally employed binary outcome variables (all but sustainability). In some cases, the original effects were estimated treating those binary variables as binary (using proportion tests or logistic regressions). In other cases they were estimated treating those binary variables as continuous (using mean tests or linear regressions). Accordingly, when estimating required sample sizes, we examined binary variables using both specifications. When we re-ran our sustainability study to focus on a binary interpretation of the variable and around winter break timing, we followed the original sampling plan to be conservative; the sample needed to detect our new binary effect (27 to 32 participants per condition) was even smaller.

We were able to target the original effect size for our health, sustainability, and consumer choice³ domains. The original effect size was far smaller in transportation and finance. We could detect it if the effect size was estimated by treating the binary variables as binary; we could not detect it without the 6x scale if the effect size was estimated by treating them as continuous. Altogether, across studies, we aimed for n=200 to n=450 per randomized condition. See Table S5 for a full reporting of these power analyses; related R code is available in our ResearchBox.

D. Recruitment

In recent years, researchers running hypothetical studies have recruited from online platforms such as mTurk or Prolific. We followed suit, recruiting from Prolific and excluding participants from taking more than one of our studies. For each study, we added filters to target sub-populations that mimicked the population sampled in the original study as closely as possible. For example, the original transportation experiment was run with employees who commuted by car to work; we recruited online participants who self-reported owning a car (of any brand) and commuted to work by car.⁴ When such filters would prohibitively reduce the available pool of participants, we relaxed them. For example, the original sustainability experiment was run with college students. If we limited our data collection to only current students on Prolific, we would have had a pool of about 5,000 participants (active in the last 90 days) and could expect, at best, 40-50% uptake and risk missing our target sample size.⁵ To keep our pool larger but still filter for participants likely to be familiar with university campuses and student experiences, we recruited participants with any higher education experience.

² “In papers published in academic journals, the average impact of a nudge is very large—an 8.7 percentage point take-up effect, a 33.5% increase over the average control. In the Nudge Unit trials, the average impact is still sizable and highly statistically significant, but smaller at 1.4 percentage points, an 8.1% increase.” (Dellavigna & Linos, 2022)

³ The primary outcome in the original study was price per donut, which had a large enough effect size that we could target it. The researchers also measured the rate of stopping for a donut, which was far smaller; we could afford to target 6x that effect.

⁴ After discussion with the original authors, we were also able to target participants in the specific country of the original experiment. However, to respect the privacy concerns in the original study, we do not report it here.

⁵ “We typically recommend expecting a 40-50% response rate from eligible participants.” Source: Prolific, researcher-help.prolific.co/hc/en-gb/articles/360009221093 accessed November 26, 2023.

Table S5. Summary of power calculation considerations and recruitment methodology across domains. For binary variables, we calculated effect size treating them as continuous variables for a more conservative estimate. The updated sustainability study is in blue. (See code files for each specific study to reproduce the effect size calculations.)

Domain	Original ES, Cohen's d^1 (6x the ES)	Sample per condition to detect original ES (6x the ES)	Total recruited (N per condition * conditions + ~100 buffer)	Total responded / Total analyzed	Dates Run	Screening requirements (no participants in more than one of our studies)	Eligible Prolific users
Consumer Choice ²	<u>WTP: Mean-to-d:</u> 0.346 (2.076) <u>Stop: Mean-to-d:</u> -0.04 (-0.24) <u>Stop: Prop-to-d:</u> 1.258 (7.548)	<u>WTP: Mean-to-d:</u> 177 (6) <u>Stop: Mean-to-d:</u> 46850 (1302) <u>Stop: Prop-to-d:</u> 13175 (367)	Round up to N=450 1 control & 1 treatment 4 hypothetical styles +2 from paper's hyp. $450*(2*4 + 2) + 100 = 4600$	<u>Responded:</u> 4611 <u>Analyzed:</u> 4525	July 13 - July 15, 2023	95%+ approval; US; fluent in English	~42,000
Health	<u>Mean-to-d:</u> 1.192 (7.152) <u>Prop-to-d:</u> 1.258 (7.548)	<u>Mean-to-d:</u> 16 (na) <u>Prop-to-d:</u> 14 (na)	Round up to N=200; 1 control & 1 treatment 4 hypothetical styles $200*2*4 + 100 = 1700$	<u>Responded:</u> 1804 <u>Analyzed:</u> 1750	June 13 - June 13, 2023	95%+ approval; US; fluent in English; Highest education above high school (no NA); student status is available (whether Y or N)	~12,000
Finance	<u>Beta-to-d:</u> 0.076 (0.457) <u>Prop-to-d:</u> 0.249 (1.496)	<u>Beta-to-d:</u> 3618 (101) <u>Prop-to-d:</u> 339 (10)	Conditions partially endogenous due to participant birthdays. ⁴ Roughly aim for: $900*2*2 = 3600$	<u>Responded:</u> 3640 <u>Analyzed:</u> 3586	July 14 - July 21, 2023	95%+ approval; US; fluent in English; having a retirement plan status is available (whether Y or N)	~27,000
Sustainability ³ Second, updated version highlighted in blue text	<u>Beta-to-d:</u> 0.307 (1.845) <u>Mean-to-d:</u> 0.816 (4.894) <u>Prop-to-d:</u> 0.891 (5.346)	<u>Beta-to-d:</u> 223 (7) <u>Mean-to-d:</u> 33 (2) <u>Prop-to-d:</u> 27 (2)	Round up to 300: 1 control & 1 treatment 4 hypothetical styles $300*2*4 + 100 = 2500$ (re-use 2500 to be conservative)	<u>Responded:</u> 2563 <u>Analyzed:</u> 2438 <u>Responded:</u> 2552 <u>Analyzed:</u> 2481	June 14 - June 21, 2023 Dec 23, 2024	95%+ approval; US; fluent in English; Highest education above high school (no NA); student status is available (whether Y or N) (same)	Original run: ~12,000 Updated run: ~29,000
Transportation	<u>Mean-to-d:</u> 0.055 (0.329) <u>Prop-to-d:</u> 0.817 (4.901)	<u>Mean-to-d:</u> 7006 (196) <u>Prop-to-d:</u> 32 (2)	Round up to 300; 1 control & 3 treatments 4 hypothetical styles $300*4*4 + 100 = 4900$	<u>Responded:</u> 4966 <u>Analyzed:</u> 4678	June 13 - June 28, 2023	95%+ approval; UK; fluent in English; Commute to work by car; Car brand status (except "rather not say/na")	~11,000

1. Cohen's d calculated in two ways for binary variables of Consumer Choice (stopping), Health, Finance, and Transportation, using linear ((P1-P2)/SDpooled) or logistic (ln(OR)*sqrt(3)/ π) approaches.

2. The stopping variable (versus the WTP variable) had a smaller effect size so we began with that as our benchmark. 3. Sustainability study target behavior is to lower the temperature. Technically, Cohen's d in this case is negative. In our original results using this continuous variable, we flipped the sign to positive to indicate the effect was in the desired direction.

4. Differs from pre-registration ($N=3000$); we collected an additional 600 participants because of a math error in our original calculation. We estimated 350 per control and 350 per the treatment (fresh start) condition, and to calculate how that played out across the two cohorts following the original paper's randomization, we estimated $N=380$ for cohort 1 and $N=530$ for cohort 2, for total $N = 910$. We scaled that to 3000 to be safe, as it seemed low; only to realize we had forgotten to multiply by four for the number of hypotheticals needed. We needed $900*4 = 3600$, not 3000.

We recruited participants from the online platform Prolific, first in June and July 2023 and then in December 2024 for the second run of the sustainability study. The advertised title and description of the task was the same for all experiments: “Imagine yourself in an everyday scenario for \$0.54: This survey is research for the [university anonymized for peer review]. In this survey you will imagine yourself in an everyday scenario and share what you would do in it. You will be paid \$0.54 for the 3 min survey. If you would like to participate, please click the link below.” We screened for participants that would be close to the target real-world population, as long as the screener left us with at least ten thousand potential participants or more; Prolific cautions that about half may be active at the time of your survey. Where a participant could be eligible for more than one study, we excluded them from taking more than one.

See Table S5 for recruitment details. Survey materials and images by experiment can be found in our Research Box; the text is also available in SI section 8.

Across all studies, the sample size was $N=17,573$ ($N=17,584$ when using the original sustainability study data) before dropping any data. The sample size was $N=16,114$ ($N=16,071$ when using the original sustainability study data) after dropping failed attention checks, incompletes, duplicate IPs, and, where relevant, outlier data (per pre-registered specifications). Neither aforementioned final sample sizes include a fifth exploratory hypothetical for the consumer choice study ($N=906$ after cleaning), which was designed to replicate the hypothetical study run by the original authors and is not included in this paper.

E. Use of Attention Checks

Across all online surveys, we used a single attention check at the start (beyond a captcha). We followed guidelines on our survey platform (Prolific) that indicate a single attention check for surveys under 5 minutes of length.⁶ Work by Roth and Yakobi (2021) also indicates that a single attention check at the start is sufficient for distinguishing attentive from inattentive survey takers. Notably, our primary goal wasn’t to achieve a given level of attention as much as to follow common practices—i.e., whatever is typically done in online settings with short surveys.

F. Representativeness of Nudges

We followed relatively strict inclusion criteria, above, for selecting experiments. For the experiments that ended up being selected, we examine here what kind of nudge interventions they represent. To do so, we leverage two prior reviews (the same as above): Szaszi et al., (2018) and Mertens et al., (2022). Both use the same framework for categorizing interventions: Munscher et al., (2015). The framework categories are listed below in SI Table S6. Notably, in Szaszi et al., (2018), interventions could be “kitchen sink” style, i.e., they could utilize more than one psychological mechanism and therefore could be coded across more than one category from the Munscher et al., (2015) framework. In contrast, interventions in Mertens et al., (2022) were only included if they represented a single mechanism and therefore could only be coded under one category.

⁶ Current guidelines can be found here: <https://researcher-help.prolific.com/en/article/fb63bb>

Table S6. A breakdown of the distribution of nudge types across two research syntheses.

From Munscher et al., 2015 (Table 1)		Prevalence of intervention categories:				Where do our five field experiments fall?
		Szasz et al., 2018 (Study N, %)		Mertens et al., 2022 (Effect size N, %)		
		Hyp	Real	Hyp	Real	
Category	Definition					
Translate information	E.g., reframe, simplify	15 (30%)	17 (16%)	20 (16%)	30 (9%)	CONSUMER FINANCE
Make information visible	E.g., make own behavior visible (feedback), make external information visible	4 (8%)	27 (25%)	10 (8%)	21 (6%)	HEALTH TRANSPORTATION
Provide social reference point	E.g., refer to descriptive norm, refer to opinion leader	8 (16%)	11 (10%)	14 (11%)	35 (11%)	TRANSPORTATION
Change choice defaults	E.g., set no-action default, use prompted choice	23 (46%)	34 (32%)	48 (40%)	80 (25%)	-
Change option-related effort	E.g., increase/decrease physical/financial effort	2 (4%)	16 (15%)	1 (1%)	22 (7%)	TRANSPORTATION
Change range or composition of options	E.g., change categories, change grouping of options	10 (20%)	7 (7%)	24 (20%)	29 (10%)	-
Change option consequences	E.g., connect decision to benefit/cost, change social consequences of the decision	1 (2%)	6 (6%)	1 (1%)	18 (6%)	-
Provide reminders	-	2 (4%)	32 (30%)	3 (2%)	66 (20%)	SUSTAINABILITY
Facilitate commitment	E.g., support self-commitment/public-commitment	0 (0%)	9 (8%)	2 (2%)	23 (7%)	-
Totals		50 (32%)	106 (68%)	123 (26%)	324 (72%)	
		Nudges can be in 1+ categories, so % do not sum to 100%		Nudges can only be in 1 category		Nudges can be in 1+ categories

The distributions of intervention prevalence varies somewhat, across hypothetical and real studies and across the two source research syntheses. The interventions in our five field studies fall across five of the nine categories, with the spread particularly driven by the transportation studies aggregate of treatments. The only highly prevalent category not covered is “change choice defaults”.

We can also explore how the interventions from our five field studies cover different cognitive mechanisms, as explored in Luo et al., (2023). The authors describe six cognitive mechanisms: attention, perception, memory, effort, intrinsic motivation, and extrinsic motivation. According to their framework (see the authors’ Table 1), our interventions cover five of the six categories: attention⁷ (transportation), perception (consumer choice, finance), memory (sustainability), effort (transportation), and intrinsic motivation (transportation). The one category they do not cover is extrinsic motivation. By these benchmarks, the nudges in our field studies therefore provide relatively good coverage—though this was not by design.

⁷ All interventions in our field studies required some attention inasmuch as participants need to pay attention to the focal elements of each given treatment stimulus. (For more on this, see the Discussion in the main text.)

IV. Supplementary Methods: Hypothetical Experiments

A. Study Demographics and Balance Checks

Table S7 below shows how long the survey took and the demographics of participants across studies. It also includes a balance check comparing the control and treatment values for these attributes, using an ANOVA F-test for duration and age and Pearson's Chi-Square test for gender and ethnicity.

Table S7. Balance checks across experiments. (See code files for each specific study to reproduce this.)

	Consumer Choice	Finance	Health	Sustainability Original	Updated	Transportation
<u>Duration (minutes)</u>						
Average	3.06	3.43	2.99	3.18	4.65	2.97
Lower-to-Upper Quantile	1.73 to 3.45	2.06 to 3.83	1.67 to 3.28	1.90 to 3.47	2.45 to 5.55	1.90 to 3.30
<i>Balance check (control-treatment)</i>	$p = .888$	$p = .180$	$p = .328$	$p = .039$	$p = .904$	$p = .000$
<u>Age</u>						
Average	39.46	35.91	41.34	38.96	39.57	38.13
Standard deviation	13.91	12.91	13.82	13.32	13.35	11.26
<i>Balance check (control-treatment)</i>	$p = .494$	$p = .975$	$p = .601$	$p = .930$	$p = .378$	$p = .657$
<u>Gender</u>						
Female	0.49	0.56	0.47	0.56	0.52	0.57
Male	0.48	0.41	0.52	0.41	0.47	0.42
Non-binary	0.02	0.03	0.02	0.02	0.01	0.00
Not listed or prefer not to say	0.00	0.00	0.00	0.00	0.00	0.00
<i>Balance check (control-treatment)</i>	$p = .449$	$p = 0.484$	$p = .738$	$p = .719$	$p = .249$	$p = .060$
<u>Ethnicity</u>						
American Indian / Alaska Native	0.00	0.00	0.00	0.01	0.01	0.00
Asian	0.09	0.07	0.08	0.09	0.06	0.04
Black	0.10	0.10	0.11	0.09	0.23	0.02
Hispanic	0.07	0.07	0.06	0.06	0.05	0.00
Native Hawaiian / Other Pacific Islander	0.00	0.00	0.00	0.00	0.00	0.00
White / Caucasian	0.71	0.72	0.72	0.72	0.63	0.91
Other	0.02	0.00	0.02	0.02	0.02	0.02
Prefer not to say	0.00	0.00	0.00	0.00	0.00	0.00
<i>Balance check (control-treatment)</i>	$p = .442$	$p = .217$	$p = .560$	$p = .281$	$p = .562$	$p = .525$

The imbalance in duration for some domains is unsurprising given the complex versions of the hypothetical scenarios had, by definition, more screens and questions for participants. No other significant differences emerged at the $p=.05$ level. P -values are unadjusted. The time to take each survey varied, with the lower quantile averaging around 1.96 minutes (average of 1.67, 2.45, 1.73, 1.90, 2.07) and the upper quantile averaging around 3.88 minutes (3.28, 5.55, 3.45, 3.30, 3.83). These same values, using the original sustainability study rather than the new one were: 1.85 minutes and 3.47 minutes, respectively.

B. Summary Statistics

Table S8 shows summary statistics across studies. While our main text focuses on those with binary outcomes, we include those with continuous outcomes (consumer choice WTP and the original sustainability study) here for transparency.

Table S8. Summary statistics across experiments.

	Consumer Choice - Stop		Consumer Choice - WTP		Finance		Health		Sustainability - Original		Sustainability - Updated		Transportation	
<u>Dependent variable</u>														
Description	Stop or not		\$ per donut		Select delayed savings or not		Choose healthier cookie or not		Thermostat setpoint (F)		Setting thermostat to 68 (F) or lower		Sign up for carpool or not	
Variable type	Binary		Continuous		Binary		Binary		Continuous		Binary		Binary	
	Control	Treat	Control	Treat	Control	Treat	Control	Treat	Control	Treat	Control	Treat	Control	Treat
<u>All Studies</u>														
Count	2694	2705	1430	1349	4957	3324	978	972	1373	1383	1396	1404	41059	18506
Mean (SD)	-	-	1.48 (1.58)	1.50 (1.23)	-	-	-	-	71.03 (3.44)	68.59 (2.83)	.74	.92	-	-
Proportion	.53	.50	-	-	.09	.10	.5	.65	-	-	-	-	.02	.15
<u>Only Field</u>														
Count	882	898	86	82	2595	2100	100	100	159	159	159	160 ⁸	39900	14987
Mean (SD)	-	-	0.65 (0.47)	0.83 (0.55)	-	-	-	-	72.69 (2.83)	71.86 (2.52)	-	-	-	-
Proportion	0.098	0.091	-	-	.03	.04	.02	0.71	-	-	.239	.613	.001	.002
<u>Only Hypothetical*</u>														
Count	1812	1807	1344	1267	2362	1224	878	872	1214	1224	1237	1244	1159	3519
Mean (SD)	-	-	1.54 (1.61)	1.54 (1.25)	-	-	-	-	70.81 (3.46)	68.16 (2.58)	-	-	-	-
Proportion	.74	.70	-	-	.17	.20	.53	.64	-	-	.80	.96	.75	.77
<u>Simple & generic</u>														
Count	453	450	382	351	606	295	217	218	309	304	310	308	289	872
Mean (SD)	-	-	1.52 (1.42)	1.50 (1.01)	-	-	-	-	71.24 (2.93)	68.37 (2.17)	-	-	-	-
Proportion	.84	.78	-	-	.16	.22	.50	.67	-	-	.78	.94	.71	.76

⁸ The counts on sustainability differ slightly because thermostat setting data was not perfectly available for all days across all rooms.

(Table continued)	Consumer Choice - Stop		Consumer Choice - WTP		Finance		Health		Sustainability - Original		Sustainability - Updated		Transportation	
<u>Dependent variable</u>														
Description	Stop or not		\$ per donut		Select delayed savings or not		Choose healthier cookie or not		Thermostat setpoint (F)		Setting thermostat to 68 (F) or lower		Sign up for carpool or not	
Variable type	Binary		Continuous		Binary		Binary		Continuous		Binary		Binary	
	Control	Treat	Control	Treat	Control	Treat	Control	Treat	Control	Treat	Control	Treat	Control	Treat
<u>Simple & specific</u>														
Count	457	452	384	352	575	316	220	219	305	310	310	314	282	879
Mean (SD)	-	-	1.33 (1.20)	1.42 (1.41)	-	-	-	-	71.10 (2.67)	68.28 (2.28)	-	-	-	-
Proportion	.84	.78	-	-	.18	.22	.54	.68	-	-	.80	.97	.82	.83
<u>Complex & generic</u>														
Count	454	455	290	280	587	312	221	216	299	306	308	312	297	878
Mean (SD)	-	-	1.76 (1.86)	1.70 (1.31)	-	-	-	-	70.37 (3.95)	68.03 (3.01)	-	-	-	-
Proportion	.64	.62	-	-	.15	.18	.57	.58	-	-	.79	.96	.66	.69
<u>Complex & specific</u>														
Count	448	450	288	284	594	301	220	219	301	304	309	310	291	890
Mean (SD)	-	-	1.63 (1.98)	1.57 (1.24)	-	-	-	-	70.51 (4.02)	67.96 (2.77)	-	-	-	-
Proportion	.64	.63	-	-	.17	.18	.51	.65	-	-	.84	.98	.83	.81

V. Supplementary Methods: Regressions

Within each domain, we run a series of regressions using individual level data from our experiments and their corresponding field studies, evaluating whether study features impact outcomes. For our binary outcomes, we run both linear and logistic regressions as a robustness check. For our sustainability initial run (continuous), effects should be interpreted in the reverse direction—turning down the temperature more, or a lower beta, is in the target direction of the nudge. Our general finding – that no design feature, beyond condition, significantly influenced hypothetical outcomes in any consistent way across domains – is robust to both specifications.

We report here the results of four sets of regressions:

- Pooled across Conditions and Designs (Table S9): $Y \sim \text{Hypothetical}$
- Pooled across Designs (Table S10): $Y \sim \text{Hypothetical} * \text{Condition}$
- Expanded without Field Data (Table S11): $Y \sim \text{Setting} * \text{Condition} + \text{Descriptors} * \text{Condition} + \text{Setting} * \text{Descriptors}$
- Expanded with Field Data (Table S12, pre-registered): $Y \sim \text{Setting} * \text{Condition} + \text{Descriptors} * \text{Condition} + \text{Setting} * \text{Descriptors}$

A. Effect of Hypotheticals on Behavior

Table S9: The effect of hypotheticals, across designs, on outcomes by domain.

	Consumer Choice				Sustainability		
	Stop	WTP	Finance	Health	Original	Updated	Transportation
<i>Linear Regression</i>							
Intercept	0.094*** (0.010)	0.739*** (0.109)	0.032*** (0.004)	0.455*** (0.035)	72.271*** (0.183)	0.426*** (0.019)	0.001* (0.001)
Hypothetical	0.627*** (0.012)	0.799*** (0.112)	0.146*** (0.006)	0.131*** (0.037)	-2.792*** (0.194)	0.457*** (0.021)	0.766*** (0.002)
<i>Logistic Regression</i>							
Intercept	-2.261*** (0.081)	-	-3.404*** (0.083)	-0.180 (0.142)	-	-0.297*** (0.113)	-6.942*** (0.137)
Hypothetical	3.213*** (0.089)	-	1.80.45776* ** (0.094)	0.529*** (0.150)	-	2.319*** (0.129)	8.136*** (0.142)
R-Squared	0.348	0.018	0.061	0.006	0.070	0.15	0.740
Adjusted R-Squared	0.348	0.018	0.061	0.006	0.069	0.15	0.740
AIC (logistic)	16251.9	-	16765.1	72406.8	-	66600.8	5920.3
Observations	5399	2779	8281	1950	2756	2800	59565

Note: Intercept absorbs real-life data. *, **, *** indicates significance at the 90%, 95%, and 99% level, respectively.

B. Effect of Hypotheticals x Conditions on Behavior

Table S10: The effect of hypotheticals, across designs, and condition on outcomes by domain.

	Consumer Choice				Sustainability		
	Stop	WTP	Finance	Health	Original	Updated	Transportation
<i>Linear Regression</i>							
Intercept	0.098*** (0.014)	0.653*** (0.152)	0.026*** (0.006)	0.200*** (0.048)	72.683*** (0.239)	0.239*** (0.026)	0.001 (0.001)
<i>Main Effects</i>							
Hypothetical	0.644*** (0.017)	0.885*** (0.157)	0.141*** (0.008)	0.330*** (0.051)	-1.874*** (0.254)	0.563*** (0.028)	0.752*** (0.004)
Treatment Condition	-0.006* (0.019)	0.176 (0.217)	0.014* (0.008)	0.510*** (0.068)	-0.825** (0.337)	0.374*** (0.037)	0.002 (0.001)
<i>Interaction Effects</i>							
Hypothetical * Treatment Condition	-0.034 (0.023)	-0.175 (0.224)	0.019 (0.013)	-0.396*** (0.072)	-1.825*** (0.359)	-0.212*** (0.039)	0.018*** (0.004)
<i>Logistic Regression</i>							
Intercept	-2.225*** (0.113)	-	-3.630*** (0.124)	-1.386*** (0.250)	-	-1.158*** (0.186)	-7.598*** (0.224)
<i>Main Effects</i>							
Hypothetical	3.280*** (0.126)	-	2.022*** (0.136)	1.505*** (0.259)	-	2.557*** (0.199)	8.709*** (0.234)
Treatment Condition	0.072 (0.162)	-	0.452*** (0.166)	2.282*** (0.333)	-	1.616*** (0.247)	1.482*** (0.284)
<i>Interaction Effects</i>							
Hypothetical * Treatment Condition	-0.130 (0.178)	-	-0.229 (0.189)	-1.810*** (0.347)	-	0.268 (0.298)	-1.371*** (0.294)
R-Squared	0.349	0.018	0.062	0.046	0.070	0.220	0.741
Adjusted R-Squared	0.349	0.017	0.062	0.044	0.069	0.219	0.740
AIC (logistic)	16468.7	-	16004.4	60506.0	-	57067.5	5894.2
Observations	5399	2779	8281	1950	2756	2800	59565

Note: Intercept absorbs real-life data (real setting with specific descriptors) and the control (non-nudge) condition. *, **, *** indicates significance at the 90%, 95%, and 99% level, respectively.

C. Effect of Setup x Descriptors x Conditions on Behavior—Hypothetical Data Only

Table S11. The effect of hypothetical setting, hypothetical descriptors, and condition on outcomes by domain.

<i>Linear Specification</i>	Consumer Choice		Sustainability				
	Stop	WTP	Finance	Health	Original	Updated	Transportation
Intercept	0.844*** (0.019)	1.505*** (0.070)	0.162*** (0.015)	0.518*** (0.031)	71.205*** (0.162)	0.774*** (0.017)	0.705*** (0.022)
<i>Main effects</i>							
Conditions: Treatment	-0.066*** (0.025)	0.009 (0.094)	0.057** (0.023)	0.128*** (0.041)	-2.809*** (0.213)	0.174*** (0.022)	0.057* (0.025)
Setting: Complex	-0.207*** (0.025)	0.267** (0.09)	-0.009 (0.020)	0.028 (0.041)	-0.802*** (0.213)	0.018 (0.022)	-0.047* (0.027)
Descriptors: Specific	-0.005 (0.025)	-0.166 (0.093)	0.018 (0.020)	0.003 (0.041)	-0.069 (0.213)	0.033 (0.022)	0.119*** (0.028)
<i>Interaction effects</i>							
Setting: Complex & Descriptors: Specific	0.012 (0.029)	0.004 (0.114)	0.001 (0.026)	-0.016 (0.047)	0.143 (0.246)	0.010 (0.025)	0.051* (0.024)
Setting: Complex & Conditions: Treatment	0.045 (0.029)	-0.096 (0.114)	-0.029 (0.027)	-0.075 (0.047)	0.404 (0.246)	-0.012 (0.025)	-0.026 (0.028)
Descriptors: Specific & Conditions: Treatment	0.007 (0.029)	0.063 (0.113)	-0.018 (0.027)	0.047 (0.047)	-0.077 (0.246)	-0.013 (0.025)	-0.050* (0.028)
R-Squared	0.042	0.008	0.003	0.016	0.166	0.067	0.020
Adjusted R-Squared	0.041	0.006	0.001	0.013	0.164	0.064	0.019
Observations (Hypothetical)	3619	2611	3586	1750	2438	2481	4678

Note: Intercept absorbs simple setting, generic descriptors, and the control (non-nudge) condition. Standard errors are reported in parentheses. *, **, *** indicates significance at the 90%, 95%, and 99% level, respectively, after applying the Benjamini-Hochberg adjustment.

Table S12. The analyses from Table S11 using logistic regression for studies with binary outcomes.

<i>Logistic Specification</i>	Consumer Choice				Sustainability		
	Stop	WTP	Finance	Health	Original	Updated	Transportation
Intercept	1.689*** (0.118)	- -	-1.642*** (0.105)	0.069 (0.127)	- -	1.248*** (0.134)	0.867*** (0.118)
<i>Main effects</i>							
Conditions: Treatment	-0.427*** (0.143)	- -	0.374* (0.155)	0.532*** (0.170)	- -	0.297* (0.134)	0.297* (0.134)
Setting: Complex	-1.122*** (0.138)	- -	-0.070 (0.143)	0.118 (0.167)	- -	-0.212 (0.149)	-0.212 (0.149)
Descriptors: Specific	-0.034 (0.148)	- -	0.129 (0.140)	0.018 (0.167)	- -	0.680*** (0.159)	0.680*** (0.159)
<i>Interaction effects</i>							
Setting: Complex & Descriptors: Specific	0.059 (0.155)	- -	0.013 (0.175)	-0.073 (0.196)	- -	0.160 (0.263)	0.226 (0.142)
Setting: Complex & Conditions: Treatment	0.334* (0.156)	- -	-0.176 (0.181)	-0.320 (0.196)	- -	0.205 (0.342)	-0.159 (0.160)
Descriptors: Specific & Conditions: Treatment	0.036 (0.152)	- -	-0.131 (0.181)	0.205 (0.196)	- -	0.500 (0.353)	-0.259 (0.163)
AIC	5247.9	-	12342.0	0.016	-	6859.8	0.020
Observations (Hypothetical)	3619	-	3586	1750	-	2481	4678

Note: Intercept absorbs simple setting, generic descriptors, and the control (non-nudge) condition. Standard errors are reported in parentheses. *, **, *** indicates significance at the 90%, 95%, and 99% level, respectively, after applying the Benjamini-Hochberg adjustment.

D. Effect of Setup x Descriptors x Conditions on Behavior—Hypothetical and Real Data

Table S13. The effect of hypothetical setting, hypothetical descriptors, and condition on outcomes by domain—compared to the real data.

<i>Linear Specification</i>	Consumer Choice		Sustainability				
	Stop	WTP	Finance	Health	Original	Updated	Transportation
Intercept	0.098*** (0.013)	0.653*** (0.151)	0.026*** (0.006)	0.200*** (0.048)	72.683*** (0.238)	0.239*** (0.026)	0.001 (0.001)
<i>Main effects</i>							
Conditions: Treatment	-0.006 (0.019)	0.176 (0.217)	0.014 (0.008)	0.510*** (0.068)	-0.825* (0.336)	0.374*** (0.037)	0.002 (0.001)
Setting: Simple	0.742*** (0.022)	0.686*** (0.166)	0.155*** (0.013)	0.321*** (0.057)	-1.547*** (0.287)	0.568*** (0.032)	0.823*** (0.006)
Setting: Complex	0.547*** (0.022)	0.958*** (0.170)	0.146*** (0.012)	0.333*** (0.057)	-2.207*** (0.287)	0.596*** (0.032)	0.827*** (0.006)
Descriptors: Generic	-0.007 (0.023)	0.161 (0.099)	-0.019 (0.015)	0.013 (0.040)	-0.074 (0.211)	-0.042* (0.023)	-0.170*** (0.008)
<i>Interaction effects</i>							
Setting: Simple & Descriptors: Generic	0.012 (0.026)	0.004 (0.111)	0.001 (0.019)	-0.016 (0.046)	0.143 (0.243)	0.010 (0.027)	0.051*** (0.007)
Setting: Simple & Conditions: Treatment	-0.053 (0.030)	-0.105 (0.235)	0.025 (0.019)	-0.336*** (0.079)	-2.061*** (0.396)	-0.212*** (0.044)	0.005 (0.007)
Setting: Complex & Conditions: Treatment	-0.008 (0.030)	-0.201 (0.238)	-0.005 (0.019)	-0.410*** (0.079)	-1.657*** (0.397)	-0.224*** (0.044)	-0.021** (0.007)
Descriptors: Generic & Conditions: Treatment	-0.007 (0.026)	-0.063 (0.110)	0.018 (0.020)	-0.047 (0.046)	0.077 (0.243)	0.013 (0.027)	0.050*** (0.008)
R-Squared	0.371	0.026	0.063	0.048	0.214	0.222	0.745
Adjusted R-Squared	0.370	0.024	0.062	0.044	0.212	0.220	0.745
Observations (Hypothetical + Real)	5399	2779	8281	1950	2756	2800	59565

Note: Intercept absorbs real-life data (real setting, specific descriptors) and the control (non-nudge) condition. Interaction between setting: complex and descriptors: generic omitted to avoid collinearity. Standard errors are reported in parentheses. *, **, *** indicates significance at the 90%, 95%, and 99% level, respectively, after applying the Benjamini-Hochberg adjustment.

Table S14. The analyses from Table S13 using logistic regression for studies with binary outcomes.

<i>Logistic Specification</i>	Consumer Choice		Sustainability				
	Stop	WTP	Finance	Health	Original	Updated	Transportation
Intercept	-2.225*** (0.113)	-	-3.630*** (0.124)	-1.386*** (0.250)	-	-1.158*** (0.186)	-7.598*** (0.224)
<i>Main effects</i>							
Conditions: Treatment	-0.072 (0.162)	-	0.452** (0.166)	2.282*** (0.333)	-	1.616*** (0.247)	1.482*** (0.284)
Setting: Simple	3.880*** (0.162)	-	2.117*** (0.161)	1.473*** (0.280)	-	2.567*** (0.233)	9.145*** (0.263)
Setting: Complex	2.816*** (0.147)	-	2.060*** (0.161)	1.518*** (0.280)	-	2.800*** (0.239)	9.195*** (0.262)
Descriptors: Generic	-0.025 (0.124)	-	-0.142 (0.142)	0.055 (0.167)	-	-0.321 (0.199)	-0.905*** (0.157)
<i>Interaction effects</i>							
Setting: Simple & Descriptors: Generic	0.059 (0.155)	-	0.013 (0.175)	-0.073 (0.196)	-	0.160 (0.263)	0.226 (0.142)
Setting: Simple & Conditions: Treatment	-0.319 (0.216)	-	-0.210 (0.226)	-1.545** (0.375)	-	0.484 (0.399)	-1.444*** (0.321)
Setting: Complex & Conditions: Treatment	0.015 (0.204)	-	-0.386 (0.230)	-1.865*** (0.374)	-	0.689 (0.427)	-1.603** (0.320)
Descriptors: Generic & Conditions: Treatment	-0.036 (0.152)	-	0.131 (0.181)	-0.205 (0.196)	-	-0.500 (0.353)	0.259 (0.163)
AIC	16549.7	-	16012.0	60486.7	-	57090.9	5812.1
Observations (Hypothetical + Real)	5399	-	8281	1950	-	2800	59565

Note: Intercept absorbs real-life data (real setting, specific descriptors) and the control (non-nudge) condition. Interaction between setting: complex and descriptors: generic omitted to avoid collinearity. Standard errors are reported in parentheses. *, **, *** indicates significance at the 90%, 95%, and 99% level, respectively, after applying the Benjamini-Hochberg adjustment.

VI. Supplementary Methods: Evaluating Treatment Effect Magnitudes

A. Rates of Under- or Overestimation

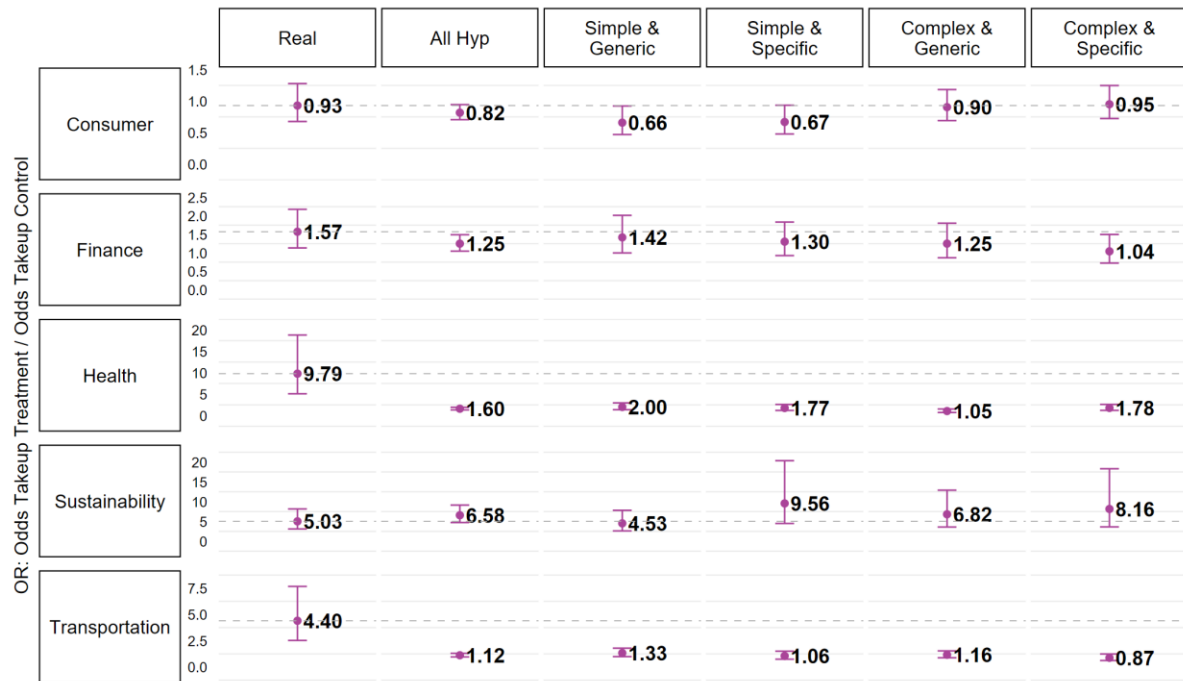
In the main text, we focus on treatment effects measured as a difference (treatment minus control, main text Figure 4). Here, we show the same data presented as a risk ratio (treatment divided by control) and an odds ratio (treatment odds ratio divided by control odds ratio), in panels A and B of Figure S8.

As noted in the main text, effects from hypotheticals were generally less extreme than their field counterparts. The underestimation rate of absolute lift was 45%, $X = 9$, $n = 20$, 95% CI [.231, .685], $p = .824$) Since the consumer choice study yielded uniformly negative lift values, taking the absolute value better captures whether hypotheticals are or less extreme. The underestimation rate of raw lift, without the absolute value, is 70%, $X = 14$, $n = 20$, 95% CI [.457, .881], $p = .115$.

Correspondingly, when measured in ratio form—e.g., an odds ratio—they tended to underestimate effects (80% underestimation rate, $X = 16$, $n = 20$, 95% CI [.563, .943], $p = .012$). This is partly a mathematical artifact: the very small control groups in our studies with very small real world effects (finance and transportation) inflated treatment effects when converted to ratio form (risk ratios and odds ratios).

This same pattern is apparent in our series of regressions, when examining the interaction between a treatment condition dummy and our hypothetical designs—i.e., the incremental impact hypotheticals had on the treatment versus the control condition. A positive coefficient indicates overestimation of treatment effects, a negative coefficient underestimation. When pooling across hypothetical designs (Tables S9 and S10), the linear specification yielded a mix of negative and positive terms (i.e., under- and overestimation); the logistic specification yielded uniformly negative terms (i.e., underestimation).

A Odds Ratio



B Risk Ratio

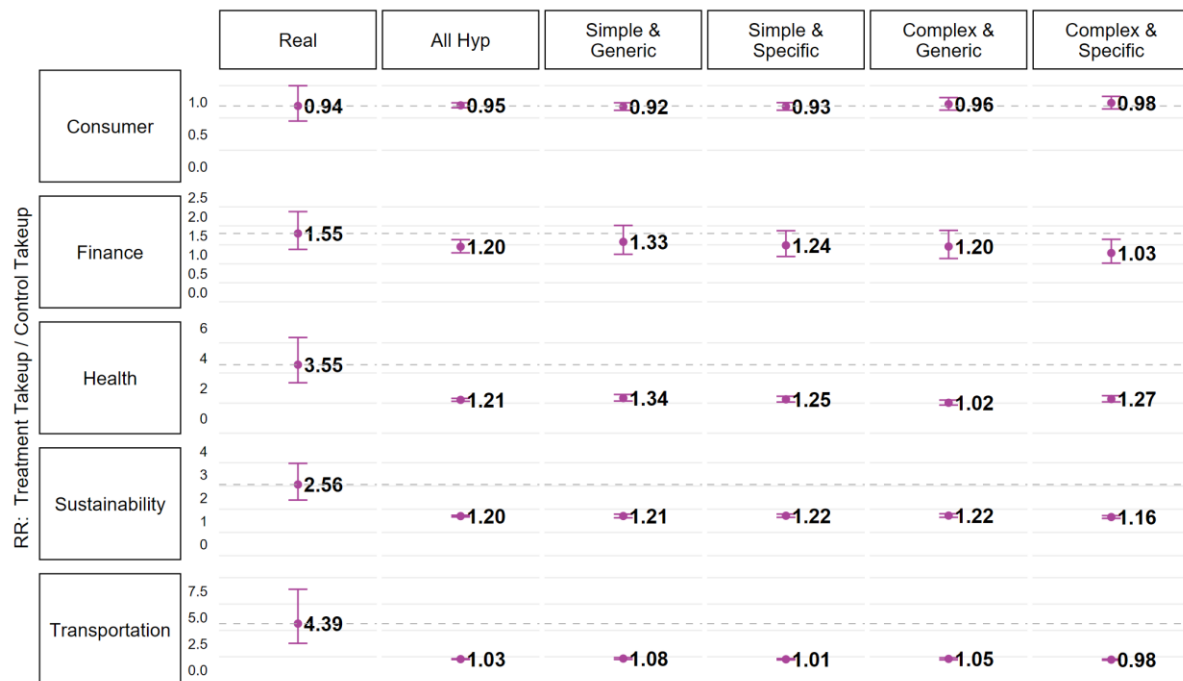


Figure S8. Estimated treatment effects (odds ratio and risk ratio) across all 20 experiments. Panel A shows the *odds ratio* of conditions (treatment odds divided by control odds) while Panel B shows the *risk ratio* of conditions (treatment proportion divided by control proportion). The colored dots show the estimated odds ratio or risk ratio, respectively; the error bars represent ± 2 standard errors around this estimate; the gray line extends the estimate for the field study across facets for comparison with estimates from the four hypothetical studies. Y-axis Difference between Treatment estimate and Control estimate: 0 = No difference. Wherever both real and hypothetical results were *both* below one or *both* above one, they were directionally concordant.

B. Sample Size Calculations

Researchers in academic or industry organizations might use hypothetical scenarios to estimate treatment effects as an input to power calculations for a costlier, focal field study. (How often they do so is, unfortunately, empirically unclear, as noted in SI section 2d.) Accordingly, one way of evaluating how well hypotheticals estimate treatment effects is to ask whether they would yield appropriate sample size estimates—i.e., enough to detect the real-world effect.

This question can be investigated in several ways, taking into account practical situations. First, when researchers lack real-world data, they may use both the hypothetical control and treatment results to estimate effects. Second, when they have an approximate real-world base rate, they then may use this for the control result and the hypothetical lift to impute the treatment results.

Consider the first situation. Say a researcher runs a hypothetical study as a pilot for an intended field experiment with the goal of estimating the minimum sample size necessary to detect the real effect. Assuming the researcher treats the hypothetical effect size as the “true” effect size, then if this effect is much larger than the real effect the corresponding power calculation will result in an *underpowered* design, potentially resulting in a false negative finding. In contrast, if the hypothetical effect size is smaller than the real effect, the field experiment will be *overpowered*, thereby wasting scarce resources. Table S15 illustrates the results of such an exercise, showing that in many cases, hypothetical estimates would lead to sample sizes that are often off by an order of magnitude or more—sometimes far too high and sometimes far too low (underpowered rate 50%, $X=10$, $n=20$, $p=1.00$). Studies where the real-world effects were very large—health and sustainability studies—would have been substantially overpowered using hypothetical estimates (in the health case, egregiously so), whereas studies where the real-world effects were smaller—consumer choice, finance, and transportation—would have been nearly always underpowered.

Sample (Per Condition) Required to Detect Treatment Estimates From Each Source

Data Source	Consumer	Finance	Health	Sustainability	Transportation
Field	35007	2483	14	26	7325
Pooled Designs	1916	2113	293	61	7104
Simple & Generic	598	841	135	71	992
Simple & Specific	639	1398	204	50	37038
Complex & Generic	6713	2262	22415	55	3270
Complex & Specific	26308	63322	195	69	5662

Note: For each calculation, we used a priori power calculations for a two-sided z-test of two independent proportions, alpha = 0.05, power = 0.80.

Table S15. Required sample size to detect the effect estimated by hypothetical inputs. The size required for the estimate from the real-world study is the first row, in gray. White cells are underpowered; blue cells are sufficiently (or, over-) powered.

In Table S16, we extend to the second situation, showing what sample sizes might result if researchers started with the real world base rate (the control from the field data) and imputed the treatment estimate from either the lift or the odds ratio in the hypothetical data.

Sample (Per Condition) Required to Detect Treatment Estimates From Each Source							
Data Source	Control Source	Treatment Source	Consumer	Finance	Health	Sustainability	Transportation
Field	Real	Real	35007	2483	14	26	7325
Pooled Designs	Hyp	Hyp Lift	1916	2113	293	61	7104
	Real	Hyp Lift	679	573	230	129	406
	Real	Hyp OR	4749	11301	389	19	2430289
Simple & Generic	Hyp	Hyp Lift	598	841	135	71	992
	Real	Hyp Lift	240	271	113	131	138
	Real	Hyp OR	1225	4344	170	30	335003
Simple & Specific	Hyp	Hyp Lift	639	1398	204	50	37038
	Real	Hyp Lift	257	390	168	112	908
	Real	Hyp OR	1311	7837	260	14	10311323
Complex & Generic	Hyp	Hyp Lift	6713	2262	22415	55	3270
	Real	Hyp Lift	2251	643	15006	111	244
	Real	Hyp OR	18578	11273	33753	19	1334779
Complex & Specific	Hyp	Hyp Lift	26308	63322	195	69	5662
	Real	Hyp Lift	9469	12198	158	179	908
	Real	Hyp OR	70563	358640	253	16	1799756

Note: For each calculation, we used a priori power calculations for a two-sided z-test of two independent proportions, alpha = 0.05, power = 0.80.

Table S16. Required sample size to detect the effect estimated by hypothetical inputs. The size required for the estimate from the real-world study is the first row, in grey. “Hyp Lift” means imputing it by adding the raw difference between the treatment and control from the hypothetical experiment, to the control proportion from the real experiment. “Hyp OR” means imputing it by using the odds ratio from the hypothetical experiment, applied to the control proportion from the real experiment. White cells are underpowered; blue cells are sufficiently or overpowered.

Sample estimates varied dramatically depending on the field study and the approach chosen, highlighting the importance of researcher degrees of freedom in something as simple as power calculations. Across specifications, under- or over-sampling rates did not differ more than expected by chance. (All hypothetical inputs: 50%, $X = 10$, $n = 20$, $p = 1.00$; real control and hypothetical lift: 55%, $X = 11$, $n = 20$, $p = .824$; real control and hypothetical odds ratio: 30%, $X = 6$, $n = 20$, $p = .115$).

C. Replication Tests

Following our pre-registration, we ran a series of minimal effect tests to statistically evaluate whether hypothetical and real-world effect sizes were, indeed, as different as they appeared. Because neither of our studies represents a population-level effect, we used the replication test provided by the TOSTER package in R (Caldwell, 2022; Lakens, 2017) and set the hypothesis to “minimal effects”. This test is effectively a two-sided test of the null hypothesis that our sampled hypothetical effect size is equivalent

to—or a “replication” of—the sampled effect size from the original study, within some predefined bounds.

To illustrate in our health study, the target treatment effect (as a log odds ratio) from the field study was 2.282 (95% *CI*: 1.628, 2.935). Collapsing across hypothetical designs, we evaluate the null hypothesis that the hypothetical effect size we found (.471; 95% *CI*: .280, .663) falls between the bounds of the target effect size. Figure S9 shows this comparison graphically.

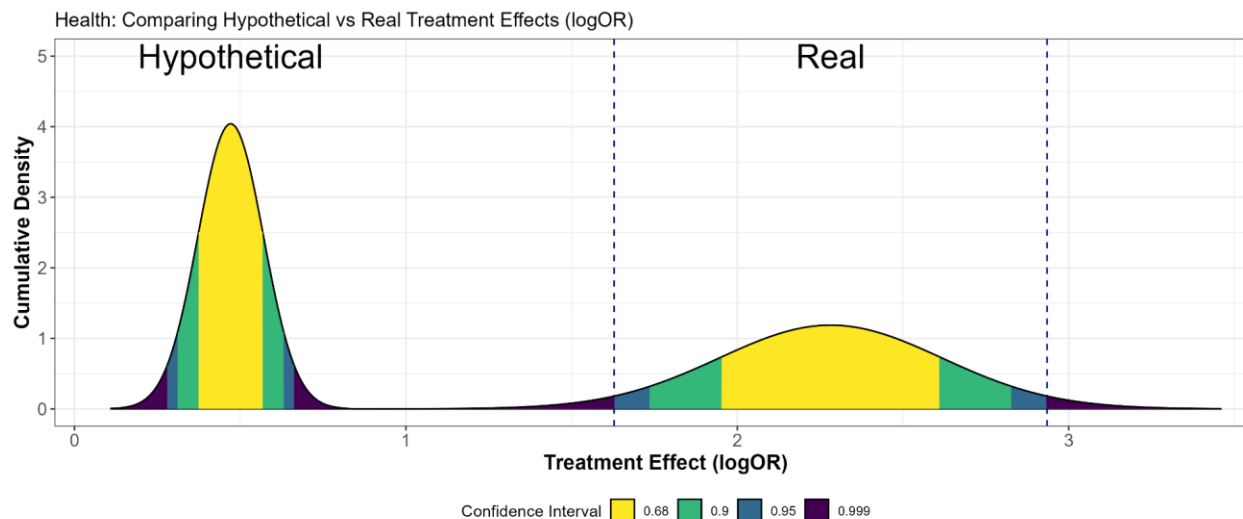


Figure S9. We visualize the replication test (hypothesis = minimal effects) examining whether the hypothetical effect size falls within the plus or minus two standard errors (~95% confidence interval) of the target effect size for our consumer choice study. (Here, a log odds ratio.) The distribution to the right represents the hypothetical effect size with its sampling variability; the distribution to the left represents the same for the real effect size; and the dashed lines represent the 95% confidence interval of the target effect size (+/- 2SE). The hypothetical distribution lies well outside the bounds of the dashed interval, providing evidence with which to reject the null that the two effects are effectively equivalent or replications of each other.

As shown in Table 4 in the main text, our results vary based on which type of standardized effect we use. When using the difference in lift, we reject the null that the hypothetical effect is a “replication” of the real-world effect for health and sustainability. When using log odds ratios, we reject that null for health and transportation. The results from the main text represent pooled estimates; the same conclusions carry over when broken out by designs, in Table S17. Here, the *p*-values for each of the four designs are adjusted using the Benjamini-Hochberg method, to be conservative. (Effects, test statistics, degrees of freedom, and additional details can be reproduced via the main code file; we showcase *p*-values here for concision.)

Study Set	Measure	Results from Replication Test (“MET”) <i>BH-adjusted p-value of two-sided test</i>				
		Pooled Designs	Simple & Generic	Simple & Specific	Complex & Generic	Complex & Specific
Consumer choice	<i>Difference</i>	.372	.436	.436	.737	.737
	<i>Log odds ratio</i>	.862	.922	.922	.922	.922
Finance	<i>Difference</i>	.280	.467	.467	.509	.535
	<i>Log odds ratio</i>	.708	.827	.827	.827	.827
Health	<i>Difference</i>	<.001***	.002***	<.001***	<.001***	<.001***
	<i>Log odds ratio</i>	<.001***	.009***	.005***	<.001***	.005***
Sustainability	<i>Difference</i>	.019**	.048**	.049**	.049**	.040**
	<i>Log odds ratio</i>	.775	.851	.851	.851	.851
Transportation	<i>Difference</i>	.114	.203	.419	.256	.256
	<i>Log odds ratio</i>	.003***	.025**	.009***	.010**	.004***

Notes: *, **, *** indicates significance at the 90%, 95%, and 99% level, respectively. Benjamini-Hochberg adjustments were applied to yield the reported *p*-values of each set of four designs, but not the pooled estimate. Log odds ratios are more appropriate here, than odds ratios, given their better statistical properties for hypothesis testing; they tend to be more normally distributed and symmetric around zero compared to raw odds ratios.

Table S17. Replication tests. We ran two-sided tests comparing each hypothetical estimate to its corresponding real-world estimate, adjusting for multiple comparisons (Benjamini-Hochberg). Each test asks whether the hypothetical estimate might be considered a “replication” of the target estimate, or not. Our preregistered bounds were plus or minus two standard errors from the original effect size. Stars visually indicate where the test suggests hypothetical estimates were significantly different from real world estimates. The precision of these tests is bounded by the sample size of the hypothetical and field studies. We can reject the null hypothesis of equivalence at the $p=.05$ level for cells highlighted in red.

Beyond the choice of how to calculate effects (e.g., lift versus log odds ratios), the choice of bounds is also subjective and worth examining across values (Lakens et al., 2018). We originally pre-registered using two standard errors around the target, actual effect size to tether our analysis to the precision of the original results, rather than choose an arbitrary effect size cutoff. Given this strategy problematically depends on the design decisions of the original study, we also present replication tests across a series of bounds—0, 0.05, and 0.1 (units of lift or units of log odds ratios)—and across both measures of treatment effects (lift and log odds ratios) in Figure S10.

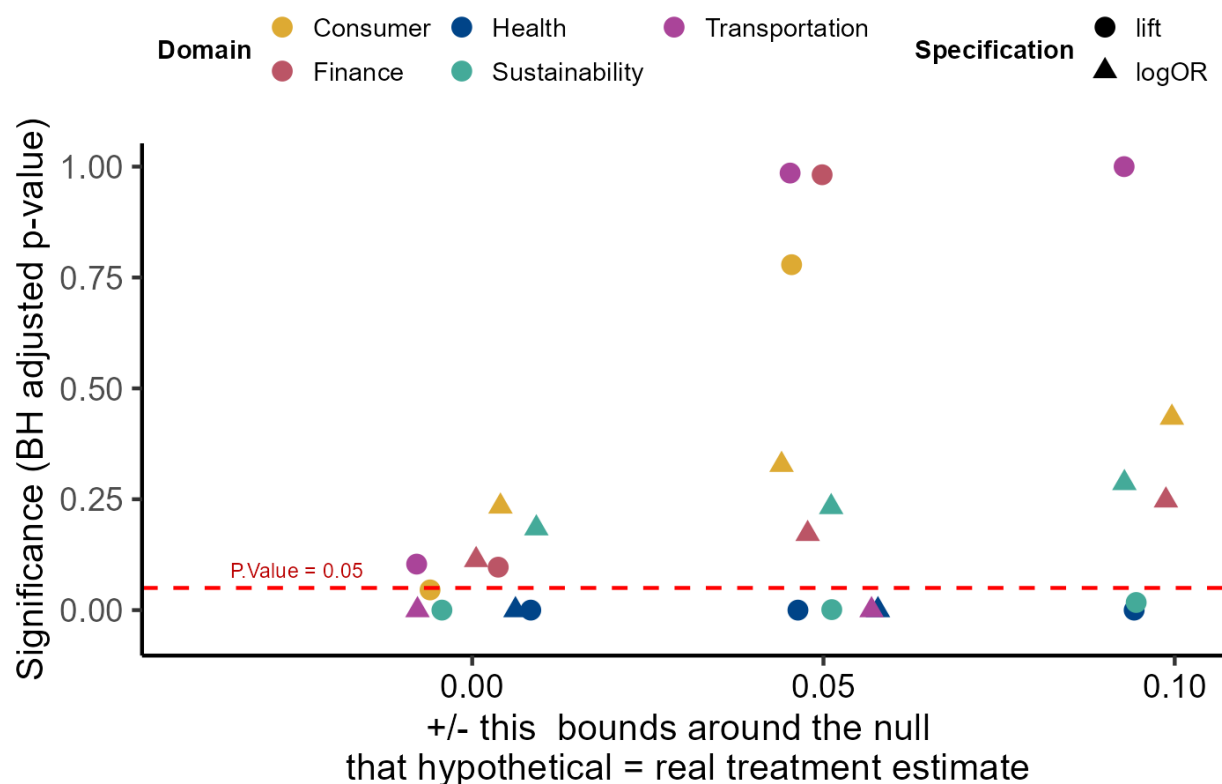


Figure S10. Replication test (hypothesis = minimal effects) comparing each hypothetical effect size to its real counterpart. Each test evaluates the null hypothesis that their difference is zero +/- a given set of bounds (0, 0.05, and 0.1); we adjust for multiple comparisons within each set of tests, by domain. We show results using lift and log odds ratios, which are represented as circles and triangles, respectively. Falling below the significance line of $p=0.05$ suggests that the real and hypothetical effect sizes are significantly different (within each given bounds). Those which almost consistently fall below the line are the health studies (both specifications), the sustainability study (linear specification only), and the transportation study (log odds ratio specification only). Colors represent domain; shapes represent ways of measuring treatment effects (lift or log odds ratio). Points jittered slightly along the x-axis to enable visibility.

D. Hypothetical effects as “less extreme” than real-world counterparts

As an exploratory finding, we observe that hypothetical effects are “less extreme” than their real-world counterparts in the sense that they are neither very small nor very large in terms of absolute lift over (or under) the control group. When effect sizes are very large in the real-world field studies (e.g., Health and Sustainability), corresponding hypothetical effect sizes are smaller; when the former is very small and near zero (e.g., Transportation), corresponding hypothetical effect sizes are larger.

What makes lift “very large” and “very small” is subjective, and there are no formally established benchmarks when it comes to considering the lift of an intervention over a control. For a non-pre-registered example, we borrow suggestive benchmarks from Dellavigna & Linos (2022), a review of field experiments of nudges also focusing on lift, with some commentary on sizes. In their abstract, they note that in their database of field studies, the average pp. lift of a nudge intervention in academic field studies was 8.7 pp, which they refer to as “very large”; they likewise refer to a 1.7 pp lift as “sizeable but

smaller”. We can map our effects visually (pooled hypothetical lift in red; field study real lift in blue) and examine where they fall across these thresholds: 1.7pp (-1.7pp for our negative effects) and 8.7pp. See Figure S11, below.

We see that the effect sizes from real-world data are always more extreme: they’re the smallest (under 1.7pp) and biggest (well over 8.7pp). In contrast, their hypothetical counterparts are neither small nor very large—at least by the standards inferred from Dellavigna & Linos (2022).

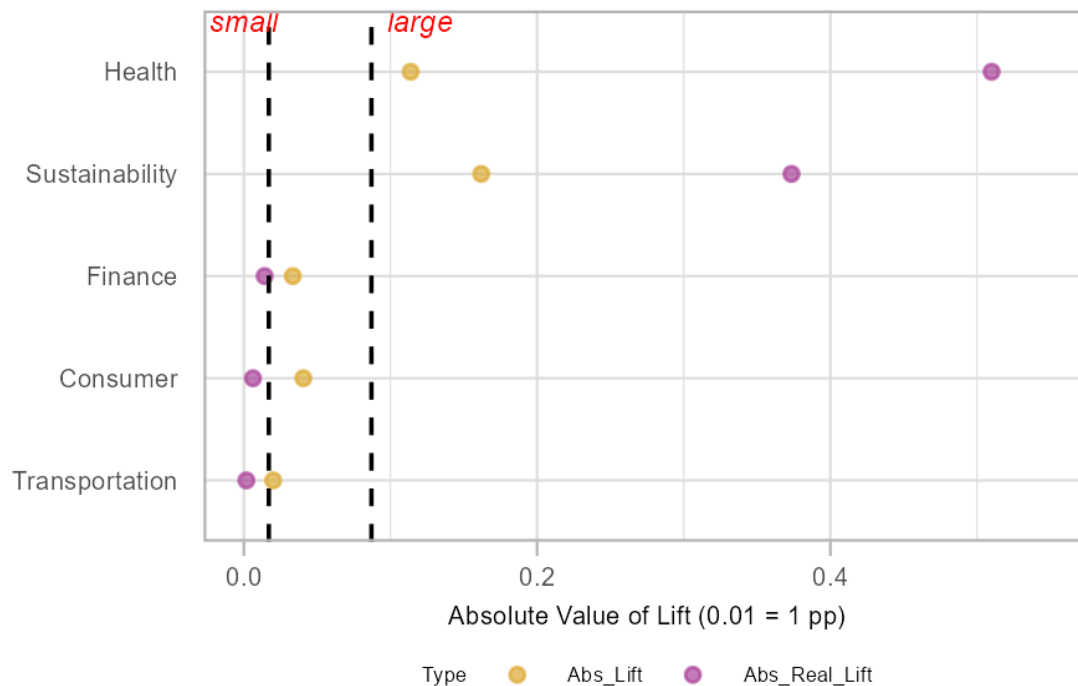


Figure S11. Absolute value of lift (all are originally positive except for Consumer) in relation to the control group, by domain. Hypotheticals are pooled for easier comparison; colors indicate whether the dot represents the hypothetical (yellow) or real-world lift (purple). Dashed lines are at 1.7pp and 8.7pp respectively.

This pattern would require a far greater number of studies than the 20 run here—and across more than 5 underlying field studies—to statistically validate. While a promising opportunity for future research, we note that its pursuit requires considerable financial investment (for sufficient sample sizes in online data collection).

VII. Supplementary Methods: Imaginability

We collected reported imaginability at the end of every survey for exploratory consideration.

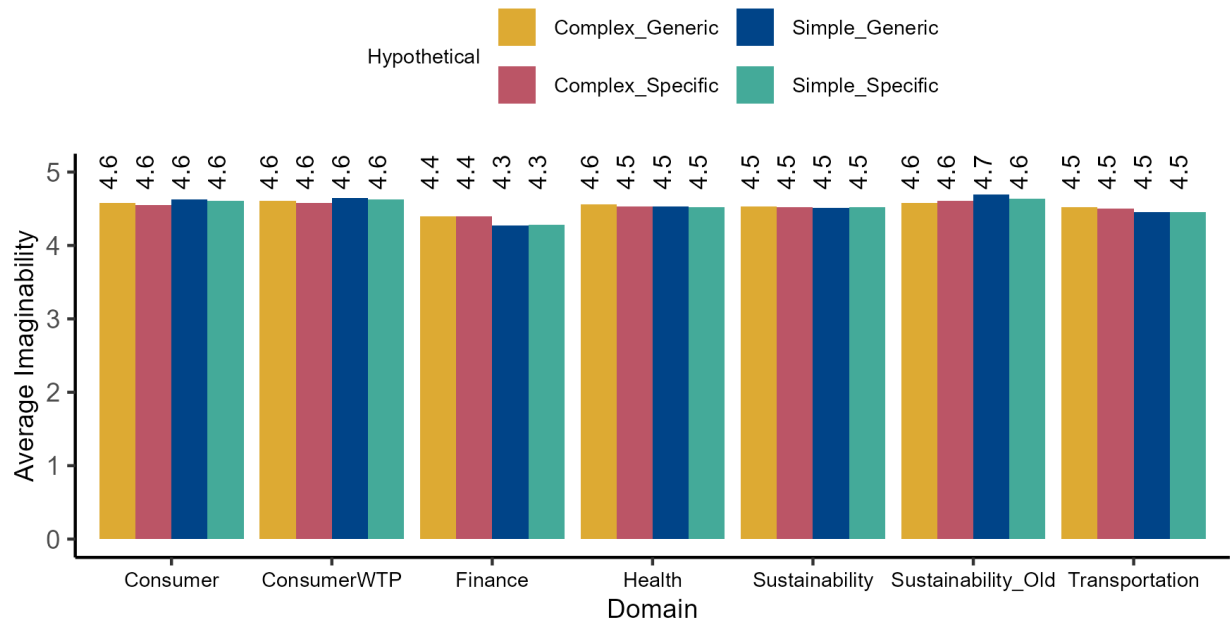


Figure S12. Average imaginability rating by hypothetical design across domains. ConsumerWTP is a subset of participants in Consumer—those who decided to stop and registered what they would pay.

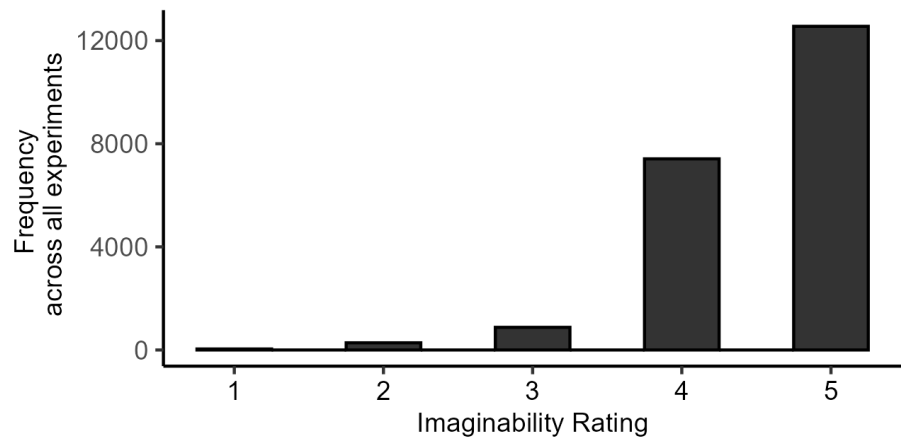


Figure S13. Distribution of responses across all studies. 94.36% of participants across all studies rated imaginability as a four or a five.

VIII. Supplementary Methods: Survey Materials

Note that original QSF files and images for every survey are available on our ResearchBox here:

<https://researchbox.org/1992>

A. Common Consent and Attention Check for All Five Surveys

Given this content is exactly the same across surveys, we include it here for concision.

Consent - all conditions, all surveys

BASICS:

Welcome! In this survey you'll be asked to imagine yourself in an everyday scenario and share what you would do in it.

ANONYMITY:

This study is anonymous, and the results will be used solely for research purposes. The research team will make every effort to keep all the information you tell us during the study strictly confidential, as required by law. The Institutional Review Board (IRB) at the [university anonymized for peer review] is responsible for protecting the rights and welfare of research volunteers like you. The information you provide for this study will not be stored for future research.

VOLUNTARY PARTICIPATION:

Your participation is voluntary which means you can choose whether or not to participate. You may choose not to participate by exiting the survey at any point. If you do so, the survey responses that have already been provided will be retained only until the data collection phase of the study is complete.

PAYMENT:

There are no known costs to you for participating in this research study except for your time. Upon completion of the survey, you will be given a code that you can submit to Prolific so that you can receive your payment. You will be paid \$0.54 for completing the 3-minute survey today.

CONTACT:

Please feel free to contact us if you have any questions, concerns, or complaints about this study. You may contact a member of the research team at [email anonymized for peer review]. If a member of the research team cannot be reached or you want to talk to someone other than those working on the study, you may contact the Office of Regulatory Affairs with any questions, concerns or complaints at the [university anonymized for peer review] by calling [phone number anonymized for peer review]. You may also contact the IRB at the [university anonymized for peer review] by calling [phone number anonymized for peer review] or emailing [email anonymized for peer review].

Please click the captcha below so we know you're human.

[CAPTCHA]

By clicking forward, you acknowledge that your participation in the study is voluntary, that you are 18 years old or older, and that you are aware that you may choose to terminate your participation in the study at any time and for any reason.

Attention check - all conditions, all surveys

Below are the instructions for this survey.

Please read them carefully and answer the question about them below. If you answer that question incorrectly, you will not be able to continue in this survey.

In this survey, you will imagine yourself in an everyday scenario and answer questions about what you would do.

- Got it

Please note that you must answer this question correctly to move forward in the survey.

Based on the description above, what will you do in this survey?

- Imagine myself in an everyday scenario
- Make up my own everyday scenario
- Imagine myself in a rare, odd scenario
- Rank scenarios for how “everyday” or normal they are

If did not answer correctly, survey ends - all conditions, all surveys

If did answer correctly, survey continues - all conditions, all surveys

Great!

What is your Prolific ID?

{free text}

B. Common Ending Questions (Demographics) for All Five Surveys

Given this content is exactly the same across surveys, we include it here for concision.

Final section of each survey - all conditions, all surveys

Thank you for your responses.

The rest of the survey will ask a few debrief questions and then redirect you to Prolific for payment.

<break>

In this hypothetical scenario, you decided {piped}. [Why / Why not?]

<break>

How easy was it to imagine yourself in this scenario?

- Very easy
- Easy
- Neither easy nor difficult
- Difficult
- Very difficult

<break>

In the final part of this survey, we have a few general background questions.

How old are you?

{drop down from 18-99}

With which gender do you most identify?

- Male
- Female

- Non-binary
- Not listed or prefer not to say

With which ethnicity do you most identify?

- American Indian / Alaska Native
- Asian
- Native Hawaiian / Other Pacific Islander
- Black
- Hispanic
- White / Caucasian
- Other (please specify) {free text}
- Prefer not to say

(Optional) If you have any comments or suggestions for the research team, please leave them here. Thank you!
{free text}

<break>

Thank you for completing this survey!

Please click the arrow below to be redirected to the payment page on Prolific.

C. Survey for Consumer Choice Study

Individuals randomized post-attention check into 2 x 2 x 2 conditions

- Simple or Complex hypothetical styles
- Generic or Specific descriptors
- Control or Treatment condition
- *We included a fifth style of hypothetical, not analyzed in this paper, mimicking a hypothetical run by the authors themselves. The materials for this hypothetical (along with the other hypotheticals) are shared in the original QSF file saved in ResearchBox.*

For participants in the SIMPLE x SPECIFIC hypothetical

Imagine that you are walking across the campus of the University of California, San Diego, and you see a stand with a banner, reading the following:

For participants in the CONTROL

"Get a Krispy Kreme Donut, Pay What You Want"

For participants in the TREATMENT

"Get a Krispy Kreme Donut, Pay What You Can"

As you approach, a student working at the stand says:

'Hi! How are you today? Are you interested in getting a donut? Take a donut and pay what you {piped → condition}.

Do you stop at the stand for a donut?

- Yes
- No

[If yes] How much do you pay for the donut? (Please enter as a numeric value, assuming the currency is US dollars.)
{free text, numeric, minimum 0, no maximum, up to 2 decimals allowed}

For participants in the SIMPLE x GENERIC hypothetical

Imagine that you are walking across the campus of a large university, and you see a stand with a banner, reading the following:

For participants in the CONTROL

"Get a Donut, Pay What You Want"

For participants in the TREATMENT

"Get a Donut, Pay What You Can"

As you approach, a student working at the stand says:

'Hi! How are you today? Are you interested in getting a donut? Take a donut and pay what you {piped → condition}.

Do you stop at the stand for a donut?

- Yes
- No

[If yes] How much do you pay for the donut? (Please enter as a numeric value, assuming the currency is US dollars.)
{free text, numeric, minimum 0, no maximum, up to 2 decimals allowed}

For participants in the COMPLEX x SPECIFIC hypothetical

How do you typically spend time on a weekday morning? (Choose whichever option fits best.)

- Working
- Relaxing
- Exercising
- Doing chores

<break>

When you're {piped → activity}, is it on your own or with others?

- On your own
- With your colleagues
- With your friends
- With your family

<break>

Imagine that you live in San Diego, California. You just spent some time {pipped → activity} {piped → others}.

<break>

Now you're taking a walk across the campus of the University of California, San Diego.



<break>

There are students around you, walking to class, carrying backpacks and talking on cell phones.



<break>

On your walk, you see a stand with some Krispy Kreme donuts on the table.



<break>

You walk closer and see the following banner.

For participants in the CONTROL

"Get a Krispy Kreme Donut, Pay What You Want"

For participants in the TREATMENT

"Get a Krispy Kreme Donut, Pay What You Can"

As you approach, a student working at the stand says:

'Hi! How are you today? Are you interested in getting a donut? Take a donut and pay what you {piped → condition}.

Do you stop at the stand for a donut?

- Yes
- No

[If yes] How much do you pay for the donut? (Please enter as a numeric value, assuming the currency is US dollars.)
{free text, numeric, minimum 0, no maximum, up to 2 decimals allowed}

For participants in the COMPLEX x GENERIC hypothetical

How do you typically spend time on a weekday morning? (Choose whichever option fits best.)

- Working
- Relaxing
- Exercising
- Doing chores

<break>

When you're {piped → activity}, is it on your own or with others?

- On your own
- With your colleagues
- With your friends
- With your family

<break>

Imagine that you live in San Diego, California. You just spent some time {pipped → activity} {piped → others}.

<break>

Now you're taking a walk across a nearby university campus.



<break>

There are students around you, walking to class, carrying backpacks and talking on cell phones.



<break>

On your walk, you see a stand with some donuts on the table.



<break>

You walk closer and see the following banner.

For participants in the CONTROL
"Get a Donut, Pay What You Want"

For participants in the TREATMENT
"Get a Donut, Pay What You Can"

As you approach, a student working at the stand says:

'Hi! How are you today? Are you interested in getting a donut? Take a donut and pay what you {piped → condition}.

Do you stop at the stand for a donut?

- Yes
- No

[If yes] How much do you pay for the donut? (Please enter as a numeric value, assuming the currency is US dollars.)
{free text, numeric, minimum 0, no maximum, up to 2 decimals allowed}

D. Survey for Finance Study

Individuals randomized post-attention check into 2 x 2 x 2 conditions

- Simple or Complex hypothetical styles
- Generic or Specific descriptors
- Control or Treatment condition → following the original study, we operationalized it this way:
 - Cohort 1: If your birthday is November, December, January, February, March, you are randomized to conditions contingent on your birthday:
 - Control (“in X months” - 2-6 based on birthday month)
 - Birthday (“after your next birthday”)
 - Holiday (“after Thanksgiving” / “after New Year’s Day”, “after MLK Day”, “after Valentine’s Day”, “after the First Day of Spring”)
 - Cohort 2: If your birthday is April, May, June, July, August, September, October, you are randomized to:
 - Control (all the above, randomized), with 2x in 3-month delay
 - Holiday (all the above, randomized), with 2x in “New Year’s Day”
 - Fresh start, or treatment = Cohort 1 all birthday, Cohort 1 and 2 New Year’s and First Day of Spring holiday
 - Control = Everyone else
 - The original study also breaks these out by holidays and delays main effects; we keep the analysis simpler (Fresh Start versus not) in our analysis

For all participants, used for above randomization

In what month is your birthday?

- January
- February
- March
- April
- May
- June
- July
- August
- September
- October
- November
- December

For participants in the SIMPLE x SPECIFIC hypothetical

Imagine that you work for the University of Pennsylvania.

It is early October and the trees are starting to turn colors.

You’ve received a letter about your employer’s tax-deferred retirement plan, in which you’re not currently enrolled. You open it, below.

For participants in the CONTROL

[We rotated through the original letters, based on the condition they were allocated to. We customized the letters to reflect the # of months / holidays indicated in their condition. See ResearchBox for the QSF file including stimuli.]

For participants in the TREATMENT

[We rotated through the original letters, based on the condition they were allocated to. We customized the letters to reflect the birthdays / holidays indicated in their condition. See ResearchBox for the QSF file including stimuli.]

It's early October, and you need to reply by November 1.

What do you decide to do?

- Enroll now
- Enroll {a later date: piped # of months / after holiday / after birthday based on condition}
- Not enroll

For participants in the SIMPLE x GENERIC hypothetical

Imagine that you work for a large university.

It is early October and the trees are starting to turn colors.

You've received a letter about your employer's tax-deferred retirement plan, in which you're not currently enrolled. You open it, below.

For participants in the CONTROL

[We rotated through the original letters, based on the condition they were allocated to. We customized the letters to reflect the # of months / holidays indicated in their condition. See ResearchBox for the QSF file including stimuli. We also removed all specifics about the university.]

For participants in the TREATMENT

[We rotated through the original letters, based on the condition they were allocated to. We customized the letters to reflect the birthdays / holidays indicated in their condition. See ResearchBox for the QSF file including stimuli. We also removed all specifics about the university.]

It's early October, and you need to reply by November 1.

What do you decide to do?

- Enroll now
- Enroll {a later date: piped # of months / after holiday / after birthday based on condition}
- Not enroll

For participants in the COMPLEX x SPECIFIC hypothetical

Which mode of transit do you tend to take when traveling around town during the week? Pick whichever fits best.

- Your car
- A taxi
- A rideshare
- The subway
- The bus

<break>

What kind of activity do you engage in to relax on an average weeknight? Pick whichever fits best.

- Reading a book
- Watching tv
- Browsing the internet
- Listening to music
- Reading the newspaper
- Working out
- Browsing social media

<break>

Imagine that you work for the University of Pennsylvania.

It is early October and the trees are starting to turn colors.



<break>

You've finished the workday and have taken {piped → transit} home.

You're now relaxing, {piped → activity}.



<break>

You then check your mail.



<break>

You've received a letter about your employer's tax-deferred retirement plan, in which you're not currently enrolled. You open it, below.

For participants in the CONTROL

[We rotated through the original letters, based on the condition they were allocated to. We customized the letters to reflect the # of months / holidays indicated in their condition. See ResearchBox for the QSF file including stimuli.]

For participants in the TREATMENT

[We rotated through the original letters, based on the condition they were allocated to. We customized the letters to reflect the birthdays / holidays indicated in their condition. See ResearchBox for the QSF file including stimuli.]

It's early October, and you need to reply by November 1.

What do you decide to do?

- Enroll now
- Enroll {a later date: piped # of months / after holiday / after birthday based on condition}
- Not enroll

For participants in the COMPLEX x GENERIC hypothetical

Which mode of transit do you tend to take when traveling around town during the week? Pick whichever fits best.

- Your car
- A taxi
- A rideshare

- The subway
- The bus

<break>

What kind of activity do you engage in to relax on an average weeknight? Pick whichever fits best.

- Reading a book
- Watching tv
- Browsing the internet
- Listening to music
- Reading the newspaper
- Working out
- Browsing social media

<break>

Imagine that you work for a large university.

It is early October and the trees are starting to turn colors.



<break>

You've finished the workday and have taken {piped → transit} home.

You're now relaxing, {piped → activity}.



<break>

You then check your mail.



<break>

You've received a letter about your employer's tax-deferred retirement plan, in which you're not currently enrolled. You open it, below.

For participants in the CONTROL

[We rotated through the original letters, based on the condition they were allocated to. We customized the letters to reflect the # of months / holidays indicated in their condition. See ResearchBox for the QSF file including stimuli. We also removed all specifics about the university.]

For participants in the TREATMENT

[We rotated through the original letters, based on the condition they were allocated to. We customized the letters to reflect the birthdays / holidays indicated in their condition. See ResearchBox for the QSF file including stimuli. We also removed all specifics about the university.]

It's early October, and you need to reply by November 1.

What do you decide to do?

- Enroll now
- Enroll {a later date: piped # of months / after holiday / after birthday based on condition}
- Not enroll

E. Survey for Health Study

Individuals randomized post-attention check into 2 x 2 x 2 conditions

- Simple or Complex hypothetical styles
- Generic or Specific descriptors
- Control or Treatment condition

For participants in the SIMPLE x SPECIFIC hypothetical

Imagine that you are sitting in the library of the University of Wisconsin-Madison.

A student nearby is quietly speaking with another library guest. The student ends their conversation, then approaches you.

The student shows you two fortune cookies.



For participants in the CONTROL

They ask you to choose one.

For participants in the TREATMENT

They ask you to choose one, noting that if you choose the plain cookie, the fortune inside the cookie will tell you something that they know about you.

Which fortune cookie do you choose?



{they click one}

For participants in the SIMPLE x GENERIC hypothetical

Imagine that you are sitting in the library of a large university.

A student nearby is quietly speaking with another library guest. The student ends their conversation, then approaches you.

The student shows you two fortune cookies.



For participants in the CONTROL

They ask you to choose one.

For participants in the TREATMENT

They ask you to choose one, noting that if you choose the plain cookie, the fortune inside the cookie will tell you something that they know about you.

Which fortune cookie do you choose?



{they click one}

For participants in the COMPLEX x SPECIFIC hypothetical

Which of the below do you typically prefer to read?

- Nonfiction book
- Fiction book
- Poetry book
- Sports magazine
- Lifestyle magazine
- News magazine
- Gossip magazine

<break>

What type of device do you typically use to tell the time, when not at home?

- Phone
- Tablet
- Watch

<break>

Imagine that you are sitting in the library of the University of Wisconsin-Madison.



<break>

You're reading a {piped → reading} that you picked out from the library stacks.



<break>

You check the time on your {piped → time} then go back to reading.



<break>

A student nearby is quietly speaking with another library guest. The student ends their conversation, then approaches you.

The student shows you two fortune cookies.



For participants in the CONTROL

They ask you to choose one.

For participants in the TREATMENT

They ask you to choose one, noting that if you choose the plain cookie, the fortune inside the cookie will tell you something that they know about you.

Which fortune cookie do you choose?



{they click one}

For participants in the COMPLEX x GENERIC hypothetical

Which of the below do you typically prefer to read?

- Nonfiction book
- Fiction book
- Poetry book
- Sports magazine
- Lifestyle magazine
- News magazine
- Gossip magazine

<break>

What type of device do you typically use to tell the time, when not at home?

- Phone
- Tablet
- Watch

<break>

Imagine that you are sitting in the library of a large university.



<break>

You're reading a {piped → reading} that you picked out from the library stacks.



<break>

You check the time on your {piped → time} then go back to reading.



<break>

A student nearby is quietly speaking with another library guest. The student ends their conversation, then approaches you.

The student shows you two fortune cookies.



For participants in the CONTROL

They ask you to choose one.

For participants in the TREATMENT

They ask you to choose one, noting that if you choose the plain cookie, the fortune inside the cookie will tell you something that they know about you.

Which fortune cookie do you choose?



{they click one}

F. Survey for Sustainability Study (two versions)

Content added when we ran the study the second time is in yellow; deleted is ~~struck through~~.

Individuals randomized post-attention check into 2 x 2 x 2 conditions

- Simple or Complex hypothetical styles
- Generic or Specific descriptors
- Control or Treatment condition

For participants in the SIMPLE x SPECIFIC hypothetical

Imagine that you live on the campus of the University of Illinois at Urbana-Champaign.

It is mid-December – final exams period – just before winter break. You'll leave for break sometime in the next week or so.

Outside, it's a typical December day: about 29 degrees Fahrenheit and almost humid enough for some flurries to show. Your thermostat is currently set at 72 degrees Fahrenheit.

For participants in the CONTROL

<nothing>

For participants in the TREATMENT

You check your email and notice the following message from the university:



Over the next week, you receive this same email message two more times – once in three days, and again two days after that.

~~Your thermostat is currently set at 72 degrees Fahrenheit.~~

~~If you'd like to change it to another temperature, please write the number (in degrees Fahrenheit) below. Otherwise, write 72 to keep it at 72.~~

Fast forward to the end of the week...

You're done with exams and about to leave for winter break. You won't return until the new year.

When you leave campus, to what temperature do you set your thermostat?
{free text, numeric, min 50, max 100, 0 decimals allowed}

For participants in the SIMPLE x GENERIC hypothetical

Imagine that you live on the campus of a large university.

It is mid-December – final exams period – just before winter break. You'll leave for break sometime in the next week or so.

Outside, it's a typical December day: about 29 degrees Fahrenheit and almost humid enough for some flurries to show. **Your thermostat is currently set at 72 degrees Fahrenheit.**"

For participants in the CONTROL

<nothing>

For participants in the TREATMENT

You check your email and notice the following message from the university:



Over the next week, you receive this same email message two more times – once in three days, and again two days after that.

~~Your thermostat is currently set at 72 degrees Fahrenheit.~~

~~If you'd like to change it to another temperature, please write the number (in degrees Fahrenheit) below. Otherwise, write 72 to keep it at 72.~~

Fast forward to the end of the week...

You're done with exams and about to leave for winter break. You won't return until the new year.

When you leave campus, to what temperature do you set your thermostat?

{free text, numeric, min 50, max 100, 0 decimals allowed}

For participants in the COMPLEX x SPECIFIC hypothetical

What's a course you've had a final exam or assessment in, when attending your higher education institution (technical school, university, or college)?

Please select one of the below examples that best fits, even if the course title doesn't perfectly match what you took.

- Economics
- Psychology
- Biology
- History
- Chemistry
- English
- Engineering
- Computer Technology
- Automotive Technology
- Culinary Arts
- Business Administration
- Marketing
- Cosmetology
- Fashion

<break>

To what temperature (in degrees Fahrenheit) do you normally set your living space in December?
{free text, numeric, min 50, max 100, 0 decimals allowed}

<break>

Imagine that you live on the campus of the University of Illinois at Urbana-Champaign.

It is mid-December – final exams period – just before winter break. You'll leave for break sometime in the next week or so.



<break>

You're in the middle of studying for your exams, including the one for {piped → course}.



<break>

You look out the window. Outside, it's a typical December day: about 29 degrees Fahrenheit and almost humid enough for some flurries to show. Your thermostat is currently set at {piped december temp} Fahrenheit."



<break>

For participants in the CONTROL

<nothing>

For participants in the TREATMENT

You check your email and notice the following message from the university:



Please, remember to **lower your thermostat to 68 degrees** before leaving for Winter Break. This is a very simple task that can **help save a lot of energy!**

Thank You!

Over the next week, you receive this same email message two more times – once in three days, and again two days after that.

~~Your thermostat is currently set at {piped december temp} degrees Fahrenheit.~~

~~If you'd like to change it to another temperature, please write the number (in degrees Fahrenheit) below. Otherwise, write {piped december temp} to keep it at {piped december temp}.~~

Fast forward to the end of the week...

You're done with exams and about to leave for winter break. You won't return until the new year.

When you leave campus, to what temperature do you set your thermostat?

{free text, numeric, min 50, max 100, 0 decimals allowed}

For participants in the COMPLEX x GENERIC hypothetical

What's a course you've had a final exam or assessment in, when attending your higher education institution (technical school, university, or college)?

Please select one of the below examples that best fits, even if the course title doesn't perfectly match what you took.

- Economics
- Psychology
- Biology
- History
- Chemistry
- English
- Engineering
- Computer Technology
- Automotive Technology
- Culinary Arts
- Business Administration
- Marketing
- Cosmetology
- Fashion

<break>

To what temperature (in degrees Fahrenheit) do you normally set your living space in December?

{free text, numeric, min 50, max 100, 0 decimals allowed}

<break>

Imagine that you live on the campus of a large university.

It is mid-December – final exams period – just before winter break. You'll leave for break sometime in the next week or so.



<break>

You're in the middle of studying for your exams, including the one for {piped → course}.



<break>

You look out the window. Outside, it's a typical December day: about 29 degrees Fahrenheit and almost humid enough for some flurries to show. Your thermostat is currently set at {piped december temp} Fahrenheit."



<break>

For participants in the CONTROL

<nothing>

For participants in the TREATMENT

You check your email and notice the following message from the university:



Over the next week, you receive this same email message two more times – once in three days, and again two days after that.

Your thermostat is currently set at {piped december temp} degrees Fahrenheit.

~~If you'd like to change it to another temperature, please write the number (in degrees Fahrenheit) below. Otherwise, write {piped december temp} to keep it at {piped december temp}.~~

Fast forward to the end of the week...

You're done with exams and about to leave for winter break. You won't return until the new year.

When you leave campus, to what temperature do you set your thermostat?
{free text, numeric, min 50, max 100, 0 decimals allowed}

G. Survey for Transportation Study

Note that the images and organization name for the “specific descriptors” version of this study are redacted to respect the privacy concerns in the original study.

Individuals randomized post-attention check into 2 x 2 x 2 conditions

- Simple or Complex hypothetical styles
- Generic or Specific descriptors
- Control or Treatment condition

For participants in the SIMPLE x SPECIFIC hypothetical

Imagine that you live outside of [REDACTED], with no convenient public transit options.

Every workday, you drive yourself to your job at [REDACTED].

Your company has a carpool program (called "[REDACTED]") to help match employees with similar commutes, but you're not currently participating in it.

For participants in the CONTROL

<nothing>

For participants in the TREATMENT

You've received a letter in the mail about your company's carpool program. You open it, below:
[REDACTED - randomized to get 1 of the 3 original treatments; these are collapsed in analysis]

Sometime in the next two months, do you register for your company's carpool program?

- Yes
- No

For participants in the SIMPLE x GENERIC hypothetical

Imagine that you live outside of a large European city, with no convenient public transit options.

Every workday, you drive yourself to your job.

Your company has a carpool program (called "Company Commuter") to help match employees with similar commutes, but you're not currently participating in it.

For participants in the CONTROL

<nothing>

For participants in the TREATMENT

You've received a letter in the mail about your company's carpool program. You open it, below:
[REDACTED - randomized to get 'genericized' versions of 1 of the 3 original treatments; these are collapsed in analysis]

Sometime in the next two months, do you register for your company's carpool program?

- Yes
- No

For participants in the COMPLEX x SPECIFIC hypothetical

What type of car do you primarily drive?

- Sedan
- SUV
- Sports car
- Minivan

- Hatchback
- Crossover
- Station wagon
- Pickup truck

<break>

In what kind of dwelling do you live?

- Apartment
- Condominium
- House
- Townhouse
- Cottage
- Duplex
- Mobile home
- Houseboat

<break>

Imagine that you live outside of [REDACTED], with no convenient public transit options.



<break>

Every workday, you use your {piped → car} to drive yourself from your {piped → dwelling} to your job at [REDACTED].



<break>

Your company has a carpool program (called [REDACTED]) to help match employees with similar commutes, but you're not currently participating in it.

<break>

For participants in the CONTROL

<nothing>

For participants in the TREATMENT

You've received a letter in the mail about your company's carpool program. You open it, below:
[REDACTED - randomized to get 1 of the 3 original treatments; these are collapsed in analysis]

Sometime in the next two months, do you register for your company's carpool program?

- Yes
- No

For participants in the COMPLEX x GENERIC hypothetical

What type of car do you primarily drive?

- Sedan
- SUV
- Sports car
- Minivan
- Hatchback
- Crossover
- Station wagon
- Pickup truck

<break>

In what kind of dwelling do you live?

- Apartment
- Condominium
- House
- Townhouse
- Cottage
- Duplex
- Mobile home
- Houseboat

<break>

Imagine that you live outside of a large European city, with no convenient public transit options.



<break>

Every workday, you use your {piped → car} to drive yourself from your {piped → dwelling} to your job.



<break>

Your company has a carpool program (called “Company Commuter”) to help match employees with similar commutes, but you’re not currently participating in it.

<break>

For participants in the CONTROL

<nothing>

For participants in the TREATMENT

You've received a letter in the mail about your company's carpool program. You open it, below:

[REDACTED - randomized to get 'genericized' versions of 1 of the 3 original treatments; these are collapsed in analysis]

Sometime in the next two months, do you register for your company's carpool program?

- Yes
- No

IX. Supplementary References

- Albers, C. and Lakens, D. (2018) When power analyses based on pilot data are biased: Inaccurate effect size estimators and follow-up bias. *Journal of Experimental Social Psychology*, 74, pp.187-195.
- Beshears, J., Dai, H., Milkman, K. L., and Benartzi, S. (2021) Using fresh starts to nudge increased retirement savings. *Organizational Behavior and Human Decision Processes*, 167, 72-87.
- Boyce, V., Mathur, M. and Frank, M.C. (2023) Eleven years of student replication projects provide evidence on the correlates of replicability in psychology. *Royal Society Open Science*, 10(11), 231240.
- Caldwell, A.R. (2022) Exploring Equivalence Testing with the Updated TOSTER R Package. *PsyArXiv*. Accessible at <https://doi.org/10.31234/osf.io/ty8de>.
- Champ, P.A. , Moore, R., and Bishop, R.C. (2009) A comparison of approaches to mitigate hypothetical bias. *Agricultural and Resource Economics Review*, 38(2), 166-180.
- Cohen, J. (1962) The statistical power of abnormal-social psychological research: a review. *The Journal of Abnormal and Social Psychology*, 65(3), 145.
- DellaVigna, S. and Linos, E. (2022) RCTs to scale: Comprehensive evidence from two nudge units. *Econometrica*, 90(1), 81-116.
- Gandhi, L., Kiyawat, A., Tipton, E. and Watts, D. (in prep) Beyond systematic reviews: Building a detailed map to navigate the choice architecture literature.
- Gomila, R. (2021) Logistic or linear? Estimating causal effects of experimental treatments on binary outcomes using regression analysis. *Journal of Experimental Psychology: General*, 150 (4), 700-710.
- Higgins, JPT et al. (editors) (2023) Cochrane Handbook for Systematic Reviews of Interventions version 6.4 (updated August 2023). Cochrane. Available from www.training.cochrane.org/handbook. Accessed November 26, 2023.
- Kristal, A.S. and Whillans, A.V. (2020) What we can learn from five naturalistic field experiments that failed to shift commuter behaviour. *Nature Human Behaviour*, 4 (2), 169-176.
- Lakens, D. (2017) Equivalence tests: A practical primer for t-tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, 1, 1–8.
- Lakens, D., Scheel, A.M., and Isager, P.M. (2018) Equivalence Testing for Psychological Research: A Tutorial. *Advances in Methods and Practices in Psychological Science*, 1(2), 259-269x.
- Luo, Y., Li, A., Soman, D. and Zhao, J., 2023. A meta-analytic cognitive framework of nudge and sludge. *Royal Society Open Science*, 10(11), p.230053.
- Mertens, S., Herberz, M., Hahnel, U.J. and Brosch, T., 2022. The effectiveness of nudging: A meta-analysis of choice architecture interventions across behavioral domains. *Proceedings of the National Academy of Sciences*, 119(1), p.e2107346118.
- Münscher, R., Vetter, M. and Scheuerle, T., 2016. A review and taxonomy of choice architecture techniques. *Journal of Behavioral Decision Making*, 29(5), pp.511-524.

Myers, E. and Souza, M. (2020) Social comparison nudges without monetary incentives: Evidence from home energy reports. *Journal of Environmental Economics and Management*, 101 (102315), 1-19.

Open Science Collaboration (2015) Estimating the reproducibility of psychological science. *Science*, 349 (6251), aac4716.

Polman, E., Ruttan, R.L., Peck, J. (2022) Using curiosity to incentivize the choice of “should” options. *Organizational Behavior and Human Decision Processes*, 173, 104192.

Prolific, researcher-help.prolific.co/hc/en-gb/articles/360009221093 accessed November 26, 2023.

Roll, S., Grinstein-Weiss, M., Gallagher, E. and Cryder, C. (2020) Can pre-commitment increase savings deposits? Evidence from a tax-time field experiment. *Journal of Economic Behavior & Organization*, 180, 357-380.

Roll, S.P., Russell, B.D., Perantie, D.C., and Grinstein-Weiss, M. (2019) Encouraging tax-time savings with a low-touch, large-scale intervention: evidence from the refund to savings experiment. *Journal of Consumer Affairs*, 53(1), 87-125.

Roth, Y. and Yakobi, O., 2024. Attention! Do We Really Need Attention Checks?. *Journal of Behavioral Decision Making*, 37(2), p.e2377.

Saccardo, S. , Li, C. X., Samek, A., and Gneezy, A. (2021) Nudging generosity in consumer elective pricing. *Organizational Behavior and Human Decision Processes*, 163, 91-104.

Simonsohn, U., Simmons, J., and Nelson, L.D. (2022) Above averaging in literature reviews. *Nature Reviews Psychology*, 1 (10), 551-552.

Szaszi, B., Palinkas, A., Palfi, B., Szollosi, A., B. Aczel, B. (2018) A systematic scoping review of the choice architecture movement: Toward understanding when and why nudges work. *Journal of Behavioral Decision Making*, 31 (3), 355-366.