

Hypothetical nudges provide directional but noisy estimates of real behavior change

Corresponding Author: Ms Linnea Gandhi

This file contains all editorial decision letters in order by version, followed by all author rebuttals in order by version.

Version 0:

Decision Letter:

**** Please ensure you delete the link to your author homepage in this e-mail if you wish to forward it to your coauthors ****

Dear Ms Gandhi,

Thank you for submitting your manuscript titled "Hypothetical nudges provide directional—but noisy—estimates of real behavior change" to Communications Psychology. We have given the paper our careful consideration and find it of potential interest. However, due to the following reasons we are concerned that sending the current manuscript out to review could lead to unnecessary delays and quite possibly an undesirable outcome of the review process.

In particular, we would welcome a revised version of the manuscript with a point-by-point letter addressed to the referees. In all matters of style and report, please follow the journal guidelines (as summarized in the attached document), which provide details on section order, statistical reporting, etc.

We would therefore like to invite you to revise your manuscript to address these concerns before we make a final determination on whether to send your manuscript for external review. We will endeavour to send the work back to the original referees, but may also seek additional feedback if necessary.

We shall hope to receive your revised version as soon as you are able to complete the suggested revisions. If something similar is published in the interim we will have to consider the impact it has on the novelty of a revised manuscript.

If you anticipate a delay of more than four weeks, please let us know. Should your manuscript be substantially delayed without notifying us in advance and your article is eventually published, the received date may be that of the revised, not the original, version.

We also ask that you ensure your manuscript complies with our editorial policies and reporting requirements.

To that end, we require revised manuscripts to be accompanied by two completed items: a reporting summary that collects information on study design and procedure, and an editorial policy checklist that verifies compliance with all required editorial policies.

- <https://www.nature.com/documents/nr-reporting-summary.zip>>Nature Research Reporting Summary
- <https://www.nature.com/documents/nr-editorial-policy-checklist.pdf>>Editorial Policy Checklist

All points on the policy checklist must be addressed. Your revised manuscript can only be sent to referees if these checklists are completed and uploaded with the revision.

If you are not interested in submitting a suitably revised manuscript in the future please let me know immediately so we can close your file. If you have any questions, please contact me.

Please use the link below when you are prepared to resubmit.
Link Redacted

Thank you for your interest in Communications Psychology.

Best regards,
Trobey Lui

Trobey Lui, PhD
Associate Editor
Communications Psychology

Version 1:

Decision Letter:

**** Please ensure you delete the link to your author homepage in this e-mail if you wish to forward it to your coauthors ****

Dear Ms Gandhi,

Thank you for your patience during the peer-review process. Your manuscript titled "Hypothetical nudges provide directional—but noisy—estimates of real behavior change" has now been seen by 2 reviewers, and I include their comments at the end of this message. For consistency with previous communication, these are numbered Reviewers #3 and #4. They find your work of interest but raised some important points. We are interested in the possibility of publishing your study in Communications Psychology, but would like to consider your responses to these concerns and assess a revised manuscript before we make a final decision on publication.

We therefore invite you to revise and resubmit your manuscript, along with a point-by-point response to the reviewers. Please highlight all changes in the manuscript text file.

Editorially, we consider it crucial that reviewer #4's concerns regarding the representativeness and clarity of the use of these nudges in the manuscript are addressed carefully. Please also address reviewer #3's concern on the claim that hypothetical studies can predict which interventions will backfire. We ask that the presentation clearly differentiates between Introduction, Methods, Results, and Discussion, with results of statistical analyses only reported in the Results section.

Please ensure you follow our statistical guidelines when reporting statistics (<https://www.nature.com/commpsychol/submit/submission-guidelines#statistical-guidelines>). Please note in particular our requirements for the reporting and interpretation of null-results. Non-significant findings derived from null-hypotheses significance tests should be reported in full, but may not be interpreted. Where you interpret null results, this interpretation must be based on Bayes Factors or equivalence tests.

We are committed to providing a fair and constructive peer-review process. Please don't hesitate to contact us if you wish to discuss the revision in more detail.

I am attaching an Editorial Requests Table that details critical reporting requirements for the revised manuscript. Please attend to each item and ensure your manuscript is fully compliant. If your revised manuscript is not aligned with these requests on major issues, such as those concerning statistics, it may be returned to you for further revisions without re-review.

Please submit the following items:

- Revised manuscript
- Point-by-point response to the referees' comments
- Cover letter (as a separate document)
- <https://www.nature.com/documents/nr-reporting-summary.pdf> Nature Research Reporting Summary
- Completed Editorial Request Table (attached).

via this link: Link Redacted .

**** This url links to your confidential home page and associated information about manuscripts you may have submitted or be reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage first ****

Additional guidance is available in our style and formatting guide Communications Psychology formatting guide.

We hope to receive your revised paper within 8 weeks; please let us know if you aren't able to submit it within this time so that we can discuss how best to proceed. If we don't hear from you, and the revision process takes significantly longer, we may close your file. In this event, we will still be happy to reconsider your paper at a later date, provided it still presents a significant contribution to the literature at that stage.

We would appreciate it if you could keep us informed about an estimated timescale for resubmission, to facilitate our planning.

We look forward to seeing the revised manuscript and thank you for the opportunity to review your work.

Best regards,

Troby Lui

Troby Lui, PhD
Associate Editor
Communications Psychology

REVIEWER EXPERTISE:

Reviewer #3: economics, public policy, experiments

Reviewer #4: nudging in psychological research

REVIEWER REPORTS:

Reviewer #3 (Remarks to the Author):

I would like to thank the authors for responding to my suggestions. I particularly appreciate the examples described in Appendix Section 2D on why researchers use hypotheticals. It provides very helpful context for those like me who don't encounter hypothetical studies very often.

Nevertheless, I still have similar concerns about the contributions of the paper. In my first comment, I asked whether there's any evidence that researchers use hypothetical studies for power analyses before running a field experiment. If not, then the contribution of one of the main findings — that hypotheticals are unreliable for power calculations (Table 2) — becomes peripheral. The authors write that such evidence is hard to find, which I'm sympathetic to, but the concern still remains. There may be some evidence from pre-registration records, or a survey of researchers could be helpful here.

My second comment questioned the claim that hypothetical studies can predict which interventions will backfire. I had two concerns here: (a) none of the five selected field studies are cases of backfiring, and (b) there are only five hypothetically replicated studies. To make this claim, one would have to hypothetically replicate a large number of cases where researchers had enough reason (e.g., strong enough priors) to invest resources into running a field experiment on interventions that eventually backfired, and then show that the hypothetical version would have predicted that they would backfire in most of those cases. That would be quite interesting. But here, only a handful of successful non-backfiring cases are hypothetically replicated.

Ultimately, I admit that my perspective on this paper is limited by the lack of value placed on hypothetical work in my field. It may very well be that many researchers in other fields use hypothetical pilots for power analyses or predicting effects before implementing a real field study, but I haven't heard of it. For this journal (Nature Communications Psychology), I believe other reviewers who have a background in psychology and are more familiar with the use of hypotheticals are better suited to judge the contributions.

Reviewer #4 (Remarks to the Author):

This is an interesting manuscript with relevant implications for research and policy. I believe as a whole the manuscript is well-written and reports on influential findings as hypothetical vignette designs are often used in nudging research. Yet, I do have some concerns, which I will address below:

1. I currently believe the manuscript is too vague about the nudges applied in these scenarios. The experiment description in Table 1 only speaks of 'nudged' and 'varying what they said' or 'varying whether or not they sent a letter'. It would be insightful to see what nudge type was applied and how. While one could go to the supplemental materials to check this, I believe this should be presented in the main text in order to judge the implications for the broader field of nudging and the

use of hypotheticals in that research field well.

2. This will also help in understanding to what extent these nudges are representative of the broader field of nudging. The authors, for example, refer to Saszi et al (2018) who revealed that changing choice defaults is the most used type of nudge (and besides, generally is seen as the most prototypical type of nudge). There may be significant differences between different types of nudges in the extent to which they are suitable for study designs with vignettes. For example, there may be differences in how dependent they are on attention, and consequently the use of hypotheticals may turn out differently for these different types of nudges. I believe it is relevant to address this and to further reflect on the role of attention for the nudges used in comparison to the broader field of nudging.

3. Another issue relates to the discussion about attention brought up in the manuscript. While in principle correct, the authors currently seem to ignore existing concerns about attentive participation in online surveys. While this may be a concern for the broader field of online research with hypotheticals, and thus the study may reflect the use of that methodology well, there are also considerable differences between studies in the extent to which they check for data quality. In the supplementary materials, it is reported that the authors used 1 attention check question in the beginning of the study. Can the authors reflect on how they ensured attentive participation in these studies?

4. The interpretation on page 17 that effects from hypotheticals appear less extreme seems incorrect or requires further explanation. What is meant with less extreme? I'm asking because some effects are less large (e.g., the one described in the next sentence for health), while others are larger (e.g., the one described in the next sentence for transportation). I don't see how an effect size being ten times larger can be considered less extreme.

Minor:

1. While I do think the discussion is nicely concise, there is a lot of interpretation of results in the results section. Perhaps this is protocol for this journal, but I would prefer saving these interpretations for the discussion section.

2. Figure 2 itself speaks of a 5 points scale, the description below a 7 point scale. The supplementary material suggest a 5 point scale was used.

3. Page 10 says 'aforementioned regressions'. I may have missed it but I don't think regressions were mentioned before that point.

4. Language use can be slightly more consistent. Page 10 includes 'descriptors'. I believe this is the same as specificity earlier in the manuscript? Would be good to use language consistently.

5. The coloured lines in figure 4 can be made thicker to make the figure more readable.

* **TRANSPARENT PEER REVIEW:** Communications Psychology uses a transparent peer review system. This means that we publish the editorial decision letters including Reviewers' comments to the authors and the author rebuttal letters online as a supplementary peer review file. However, on author request, confidential information and data can be removed from the published reviewer reports and rebuttal letters prior to publication. If your manuscript has been previously reviewed at another journal, those Reviewers' comments would not form part of the published peer review file.

** Visit Nature Research's author and referees' website at <http://www.nature.com/authors> for information about policies, services and author benefits**

Communications Psychology is committed to improving transparency in authorship. As part of our efforts in this direction, we are now requesting that all authors identified as 'corresponding author' create and link their Open Researcher and Contributor Identifier (ORCID) with their account on the Manuscript Tracking System prior to acceptance. ORCID helps the scientific community achieve unambiguous attribution of all scholarly contributions. You can create and link your ORCID from the home page of the Manuscript Tracking System by clicking on 'Modify my Springer Nature account' and following the instructions in the link below. Please also inform all co-authors that they can add their ORCIDs to their accounts and that they must do so prior to acceptance.

<https://www.springernature.com/gp/researchers/orcid/orcid-for-nature-research>

For more information please visit <http://www.springernature.com/orcid>

If you experience problems in linking your ORCID, please contact the [Platform Support Helpdesk](http://platformsupport.nature.com/).

Version 2:

Decision Letter:

** Please ensure you delete the link to your author homepage in this e-mail if you wish to forward it to your coauthors **

Dear Ms Gandhi,

Your manuscript titled "Hypothetical nudges provide directional—but noisy—estimates of real behavior change" has now

been seen by our reviewers, whose comments appear below. In light of their advice I am delighted to say that we are happy, in principle, to publish a suitably revised version in Communications Psychology.

We therefore invite you to revise your paper one last time to address the remaining concerns of our reviewers and a list of editorial requests. At the same time we ask that you edit your manuscript to comply with our format requirements and to maximise the accessibility and therefore the impact of your work.

EDITORIAL REQUESTS:

Please review our specific editorial comments and requests regarding your manuscript in the attached "Editorial Requests Table". Please outline your response to each request in the right hand column. Please upload the completed table with your manuscript files as a Related Manuscript file.

If you have any questions or concerns about any of our requests, please do not hesitate to contact me.

SUBMISSION INFORMATION:

In order to accept your paper, we require the files listed here <https://www.nature.com/documents/commsj-file-checklist.pdf>.

OPEN ACCESS:

Communications Psychology is a fully open access journal. Articles are made freely accessible on publication. For further information about article processing charges, open access funding, and advice and support from Nature Research, please visit <https://www.nature.com/commspsychol/open-access>

At acceptance, you will be provided with instructions for completing the open access licence agreement on behalf of all authors. This grants us the necessary permissions to publish your paper. Additionally, you will be asked to declare that all required third party permissions have been obtained, and to provide billing information in order to pay the article-processing charge (APC).

* **TRANSPARENT PEER REVIEW:** Communications Psychology uses a transparent peer review system. On author request, confidential information and data can be removed from the published reviewer reports and rebuttal letters prior to publication. If you are concerned about the release of confidential data, please let us know specifically what information you would like to have removed. Please note that we cannot incorporate redactions for any other reasons.

* **CODE AVAILABILITY:** All Communications Psychology manuscripts must include a section titled "Code Availability" at the end of the methods section. We require that the custom analysis code supporting your conclusions is made available in a publicly accessible repository at this stage; please choose a repository that generates a digital object identifier (DOI) for the code; the link to the repository and the DOI must be included in the Code Availability statement. Publication as Supplementary Information will not suffice.

* DATA AVAILABILITY:

All Communications Psychology manuscripts must include a section titled "Data Availability" at the end of the Methods section. More information on this policy, is available in the Editorial Requests Table and at <http://www.nature.com/authors/policies/data/data-availability-statements-data-citations.pdf>.

Please use the following link to submit the above items:

Link Redacted

** This url links to your confidential home page and associated information about manuscripts you may have submitted or be reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage first **

We hope to hear from you within two weeks; please let us know if you need more time.

Best regards,

Troby Lui

Troby Lui, PhD
Associate Editor
Communications Psychology

REVIEWERS' COMMENTS:

Reviewer #3 (Remarks to the Author):

I appreciate the additional effort from the authors to address my comments. I think the discussion of the results has become more balanced. I have no additional concerns.

Reviewer #4 (Remarks to the Author):

The authors have successfully addressed my comments

** Visit Nature Research's author and referees' website at www.nature.com/authors for information about policies, services and author benefits**

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

REVIEWER REPORTS:

Reviewer #3 (Remarks to the Author):

I would like to thank the authors for responding to my suggestions. I particularly appreciate the examples described in Appendix Section 2D on why researchers use hypotheticals. It provides very helpful context for those like me who don't encounter hypothetical studies very often.

Thank you in turn for the helpful suggestions in the prior round of feedback!

Nevertheless, I still have similar concerns about the contributions of the paper. In my first comment, I asked whether there's any evidence that researchers use hypothetical studies for power analyses before running a field experiment. If not, then the contribution of one of the main findings — that hypotheticals are unreliable for power calculations (Table 2) — becomes peripheral. The authors write that such evidence is hard to find, which I'm sympathetic to, but the concern still remains. There may be some evidence from pre-registration records, or a survey of researchers could be helpful here.

We appreciate your concern and agree that without more substantial descriptive evidence of their use, the extrapolation to sample size should not be prioritized. It was originally included based on feedback in prior reviews, and we were unable to find reliable publicly-available data (per our earlier notes about publication bias and sparse documentation). Accordingly, we have taken your feedback: We moved the bulk of the analysis of sample sizes to the supplement (section 6b), noted the difficulties finding evidence of their use, and emphasized the need for validation of the latter.

To complement these changes, we moved the core table on replication tests—our focal statistical test of how well hypothetical effect sizes reflect their real-world counterparts—back into the main text. (The latter was also recommended by the Editor.)

Further, we de-prioritize references to sample sizes as an application throughout the text. For example, we delete it from the abstract and now list it last across our three use cases in the final sentence of the introduction as well as in our methods (existence proofs, policy applications, and sample size calculations). We also emphasize (in SI section 2d) that there is greater evidence of the other two use cases. We write: *"Given the greater evidence for the first two applications (see SI section 2d), we focus on the first two in the main text and the third in the supplement (SI section 6b)."* Similarly, in the discussion section, we rephrase the checklist reference to it as "if" hypotheticals are used for sample size calculations, rather than "when".

While we think it is helpful to keep some reference to sample size calculations in the overall manuscript—in order to provide another practical way to think about effect size implications—we agree with you that this analysis should not at all be as focal as the other two.

My second comment questioned the claim that hypothetical studies can predict which interventions will backfire. I had two concerns here: (a) none of the five selected field studies are cases of backfiring, and (b) there are only five hypothetically replicated studies. To make this claim, one would have to hypothetically replicate a large number of cases where researchers had enough reason (e.g., strong enough priors) to invest resources into running a field experiment on interventions that eventually backfired, and then show that the hypothetical version would have predicted that they would backfire in most of those cases. That would be quite interesting. But here, only a handful of successful non-backfiring cases are hypothetically replicated.

Thank you for raising this set of concerns! We had originally tied our hands in selecting experiments, following inclusion criteria that narrowed us to very few viable papers (as noted in SI section 3a); unfortunately this meant we weren't selecting, *ex ante*, for a balanced sample of positive and negative effect sizes. You're absolutely right to note that the effects in the studies we employed were predominantly positive; only one was negative (Consumer Choice), and it was statistically indistinguishable from zero at the sample size of the original field study.

We take your concerns very seriously, so the below is a somewhat longer response to ensure we have appropriately addressed them. We've divided it up into a few subsections for ease of your review.

Backfiring

We implemented a principled approach to go back into our source database and two others for studies with effects that clearly backfired (i.e., negative and statistically significant effects), in field settings, with binary outcomes, with a population we could sufficiently source online (Prolific), and with a context transferable to an online setting.

- First, we revisited our original search criteria (UK & US, $n > 99$), and then expanded from 2018-2023 (rather than 2022) plus focused on binary outcomes per our current paper structure. Only four effect sizes from three distinct studies were significant and negative. One study dealt with EITC recipients in California; we could find a Prolific screener for a UK low-income program but not for similar US programs. (Further, only ~1100 participants in California reported having an income low enough to qualify; we typically aim for more, see SI Table S5, to ensure a sufficient number of participants actually take the survey.) The other two were partnerships with TurboTax, both encouraging savings, and both with the same lead author—suggesting it would be best to only use one for independence. For reference, here are the citations of the three studies we identified.
 - Linos, E., Prohovsky, A., Ramesh, A., Rothstein, J. and Unrath, M., 2022. *Can nudges increase take-up of the EITC? Evidence from multiple field experiments*. *American Economic Journal: Economic Policy*, 14(4), pp.432-452. (Tough to find appropriate pop)
 - 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, pp.357-380. (Viable)
 - 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), pp.87-125. (Viable)

- We then extended our search to a popular database of nudge studies, [Mertens et al., \(2022\)](#), filtering for natural field experiments and negative Cohen's d (after these two filters, all were already binary and from the US). We identified 12 negative effect sizes from 7 distinct studies. None, however, were significant at the $p < 0.05$ level using both a Chi-square and a T-test approach to two-sided significance testing. For robustness, we did both tests, given the sensitivities we noted to specifications when working with the data in this paper. For reference, here is the one study (with three effects) that was significant under the Chi-square method but not the t-test method; its population of low-income immigrants seeking citizenship would unfortunately also be challenging to identify at scale on Prolific.
 - [Hainmueller, J., Lawrence, D., Gest, J., Hotard, M., Koslowski, R. and Laitin, D.D., 2018.](#) *A randomized controlled design reveals barriers to citizenship for low-income immigrants. Proceedings of the National Academy of Sciences*, 115(5), pp.939-944.
- Finally, we investigated [Osman et al., 2020](#) a review of "Behaviour Changes that Fail". Here, we found three cases of significantly negative effect sizes. One was the Finance study that we are already using, but looking at a secondary variable. Another—coincidentally with the same lead and anchor authors as our Finance study—relied on a sub-group to find heterogeneity, in a way we could not authentically induce in an online lab setting (401k eligibility). This left a single study in the realm of finance (nudging tax payments in the UK). For reference, here are the citations of the two new studies we considered.
 - [Beshears, J., Choi, J.J., Laibson, D., Madrian, B.C. and Milkman, K.L., 2015.](#) *The effect of providing peer information on retirement savings decisions. The Journal of finance*, 70(3), pp.1161-1201. (Negative effect with endogenous subgroup.)
 - [John, P.C.H., 2018.](#) *How best to nudge taxpayers?: The impact of message simplification and descriptive social norms on payment rates in a central London local authority. Journal of Behavioral Public Administration*, 1(1), pp.1-11. (Viable)

Given there were only a handful of viable studies and all but one were in the domain of Finance (and two from the same context and lead author), this did not seem like a convincing enough set of negative effects to invest in further data collection to more robustly assess whether hypotheticals would correctly detect the "backfiring" of nudge interventions. Given your second comment about there being only five field experiments to begin with, and wanting a "large number" of additional negative cases, we assume you will agree!

This dearth of negative evidence is unfortunately a symptom of broader bias in the literature, on which many others have prior commented with respect to psychology and other disciplines (c.f. [Ferguson & Heene, 2012](#) and [Franco et al., 2014](#)). Journals may be averse to sharing negative results, and researchers in turn may be averse to submitting interventions that did not "work". Further, given many control conditions are "do-nothing" it may be that in the field nudging, interventions tend to have even a small positive effect; doing something may often work at least a little better than doing nothing.

Number of field studies:

We also appreciate your second point about there being only five field experiments represented in the current manuscript. While these do represent over 16,000 individuals participating in our studies (since there are four versions of each, for 20 experiments), we understand that the results cannot nearly represent “all” the hypothetical literature out there. We acknowledge this in the limitations section of the discussion.

We spent considerable time investigating additional studies to add, but were ultimately dissuaded by the dearth of examples with negative effects (above). We also were concerned that there is no “magic” number at which representativeness is assured or even convincing, *ex ante* of seeing results. No matter what set of studies we added, they would not be fully representative across domains, behaviors, interventions, and effect sizes. This is necessarily a limitation of any research that samples from a broader population of participants or, in our case, studies. (This extends even to meta-analysis, *c.f.*, [Borenstein et al., 2009](#); [Valentine et al., 2010](#).)

We venture to point out that this limitation similarly holds for any given hypothetical study (see SI section 2d for evidence that publications relying on these single-context, hypothetical studies often claim to contain knowledge that generalizes more broadly, to non-hypothetical contexts). Our goal in this paper is simply to put that assumption to an initial empirical test, for a set of five instances x four hypothetical designs. This set is diverse across domains, by design; and fortunately diverse across interventions, by happenstance (see response to reviewer 4 and the new SI section 3F). We believe additional studies would be suitable as follow-on work with a more targeted scope than this initial investigation—especially if more field studies with significant negative results could be found in the public domain.

Changes made

All this considered, we ultimately did not collect new data during this round of review. However, we have modified the language of the main paper to be as clear as possible about what our evidence does and does not show, given the above limitations. Specifically: (i) We ensured the limitations section (as written) appropriately acknowledges that our study sample may not be representative. (ii) We also removed the use of the term “backfire” in all but one instance in the paper, including the abstract. The instance we retain is in the general discussion section, and we explicitly add the caveat that stronger negative cases are required to fully validate this. We write: *“This consistency suggests they may serve as an effective low-cost tool for existence proofs or to detect nudges that are likely to backfire when rolled out in the field—although additional field studies with strong, negative effects would be necessary to fully validate the latter.”*

Ultimately, I admit that my perspective on this paper is limited by the lack of value placed on hypothetical work in my field. It may very well be that many researchers in other fields use hypothetical pilots for power analyses or predicting effects before implementing a real field study, but I haven’t heard of it. For this journal (Nature Communications Psychology), I believe other reviewers who have a background in psychology and are more familiar with the use of hypotheticals are better suited to judge the contributions.

Thank you for your thoughtful comments on the manuscript. Even if hypothetical studies are not part of your background, we believe your critiques have served to strengthen the manuscript, and we hope you will agree!

Reviewer #4 (Remarks to the Author):

This is an interesting manuscript with relevant implications for research and policy. I believe as a whole the manuscript is well-written and reports on influential findings as hypothetical vignette designs are often used in nudging research. Yet, I do have some concerns, which I will address below:

Thank you for finding the manuscript of interest, and for your below suggestions for improvement!

1. I currently believe the manuscript is too vague about the nudges applied in these scenarios. The experiment description in Table 1 only speaks of 'nudged' and 'varying what they said' or 'varying whether or not they sent a letter'. It would be insightful to see what nudge type was applied and how. While one could go to the supplemental materials to check this, I believe this should be presented in the main text in order to judge the implications for the broader field of nudging and the use of hypotheticals in that research field well.

This is a great suggestion! We would be happy to add this additional detail in the main text. To directly address your comment, we have added a column in Table 1 as well as supplement Table S4. Relatedly, please see our responses to your below comment to ensure readers get a sense of how these specific interventions situate in the literature more broadly.

2. This will also help in understanding to what extent these nudges are representative of the broader field of nudging. The authors, for example, refer to Saszi et al (2018) who revealed that changing choice defaults is the most used type of nudge (and besides, generally is seen as the most prototypical type of nudge). There may be significant differences between different types of nudges in the extent to which they are suitable for study designs with vignettes. For example, there may be differences in how dependent they are on attention, and consequently the use of hypotheticals may turn out differently for these different types of nudges. I believe it is relevant to address this and to further reflect on the role of attention for the nudges used in comparison to the broader field of nudging.

This is a really nice suggestion that we think will strengthen readers' understanding and better situate the paper overall. While we selected the experiments in our paper based on the representativeness of the behavioral domain, it is important to show how representative the interventions (nudges) themselves are. To do so, we have added a new section to the supplement (Section 3F).

In this section, we do three things. First, we show the distribution of interventions in two recent reviews: [Szasz et al., \(2018\)](#) and [Mertens et al., \(2022\)](#), both of which use the [Munscher et al., \(2015\)](#) framework for categorizing interventions. Second, we show where the interventions from the present work lie along these distributions: they cover five of the nine categories (notably, they do *not* provide an example of defaults). Third, we reference a more recent review and categorization of cognitive mechanisms by [Luo et al., \(2023\)](#): the interventions from the present work cover five of the six categories. The interventions therefore provided good coverage and a diverse set of concepts, even though this was not by design.

We also add a footnote in that section (SI section 3F) to clarify the broader role of attention that we see in these studies: Each of the interventions in the experiments of this paper relied on attention – by the mere fact of having to visually notice and then engage with the stimulus – but the interventions *within* the stimuli themselves varied in their use of attention as a cognitive mechanisms (per Luo et al., 2023’s framework). When it comes to the point we bring up in the main text – that attention works differently in the field than in online lab studies, like vignette/hypothetical studies – we mean the former point (noticing the stimulus) rather than the latter.

We think this additional section better situates the studies and their interventions for readers – thank you for suggesting it!

3. Another issue relates to the discussion about attention brought up in the manuscript. While in principle correct, the authors currently seem to ignore existing concerns about attentive participation in online surveys. While this may be a concern for the broader field of online research with hypotheticals, and thus the study may reflect the use of that methodology well, there are also considerable differences between studies in the extent to which they check for data quality. In the supplementary materials, it is reported that the authors used 1 attention check question in the beginning of the study. Can the authors reflect on how they ensured attentive participation in these studies?

Thank you for asking about this! You’re entirely right: our aim was to follow general practices rather than do anything to heighten (or reduce) attention to stimuli. We also followed Prolific’s guidelines (<https://researcher-help.prolific.com/en/article/fb63bb>).

Specifically, we used a single attention check given each survey was designed to only be about three minutes long. Further, Prolific’s guidelines require one attention check to be failed when reviewing responses for quality when a survey is less than five minutes; and two attention checks when a survey is five minutes or longer. Then, from Prolific’s options, we used an instructional attention check rather than a comprehension check, since there was no additional information to comprehend beyond imagining oneself in a hypothetical scenario. We also opted not to use nonsensical questions as, in our own experience in past projects, some participants have left comments saying that they found these to be “insulting to their intelligence”. [Roth & Yakobi \(2024\)](#) also find that typically one attention check is sufficient to differentiate between

attentive and non-attentive participants. In case other readers share your concern, we have also added a brief discussion of this in the Supplement in a new section (section 3E).

We very much appreciate the question you ask about how we ensured participants paid attention. Ensuring a particular level of attention wasn't a primary goal for us; more important was to try to reflect the practices typically engaged in with online lab studies.

4. The interpretation on page 17 that effects from hypotheticals appear less extreme seems incorrect or requires further explanation. What is meant with less extreme? I'm asking because some effects are less large (e.g., the one described in the next sentence for health), while others are larger (e.g., the one described in the next sentence for transportation). I don't see how an effect size being ten times larger can be considered less extreme.

Our apologies for the lack of clarity here. We meant that they are less extreme 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. We realize what makes lift "very large" and "very small" is a subjective one, and there are no formally established benchmarks when it comes to considering the lift of an intervention over a control.

For example, we could take candidate benchmarks from [Dellavigna & Linos \(2022\)](#), a review of field experiments of nudges also focusing on lift, with some commentary on sizes. (Note: we do not use this database in our review of nudges above because they use a distinct framework and also do not have publicly available details on most studies.) 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 (pooled hypothetical lift in red; real lift in blue) visually and look at where they fall across these thresholds: 1.7pp (-1.7pp for our negative effects) and 8.7pp. Real effect sizes are always what we'd consider more extreme, on this mapping: they're the smallest (under 1.7pp) and biggest (well over 8.7pp). In contrast, their hypothetical counterparts are neither small nor very large (by the standards from Dellavigna & Linos, 2022 at least).

A visual of this contrast is directly below (and included now in the Supplement as Fig S11).

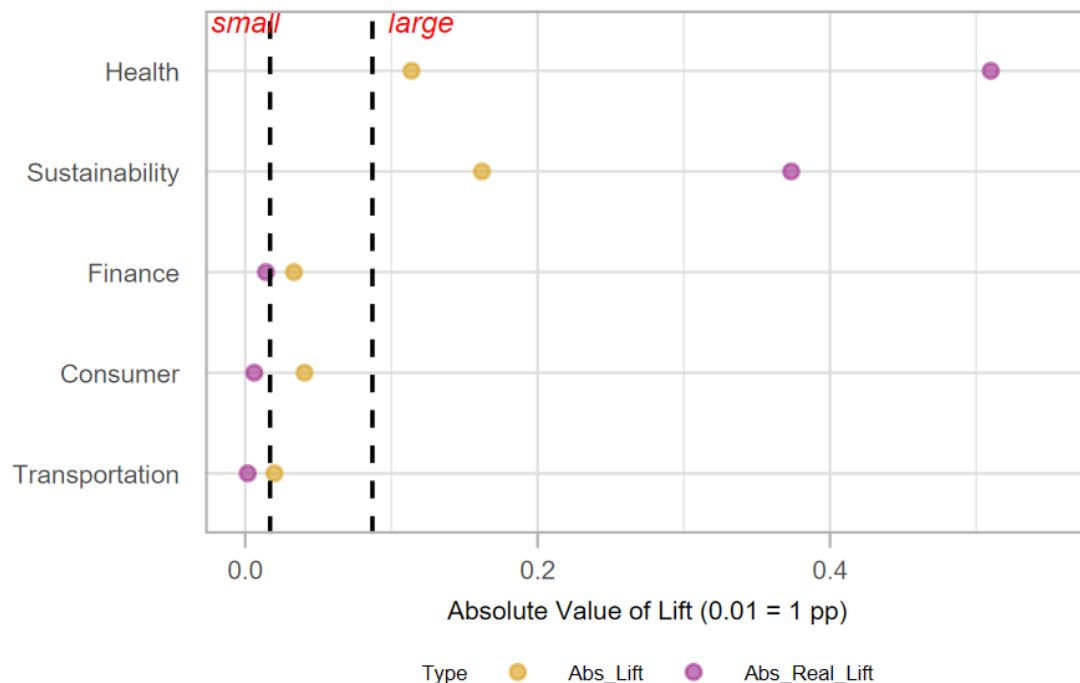


Fig (S11). Absolute value of lift (all are originally positive except for Consumer Choice) in relation to the control group, by domain. Hypotheticals are pooled for easier comparison; colors indicate whether the dot represents the hypothetical (yellow) or real-world lift (purple). Dashed lines are at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

We have added a version of this discussion to the supplement text (see new section 6d) to help clarify what we meant by “less extreme”, while also making sure the reader knows that these benchmarks were not pre-registered comparators—we only use them for ex post exploratory interpretation. Further, we remove the confusing ratios that you noted in the main text, replacing them with a clearer statement:

- OLD, deleted: “For example, when expressed in terms of the difference between treatment and control, hypothetical effect sizes vary from one-fourth of the largest real effect (health) to ten times the smallest real effect (transportation).”
- NEW, added: “Where field effects are small (e.g., $abs|lift| < 2pp$), hypothetical effects are larger; where field effects are large (e.g., $abs|lift| > 20pp$) hypothetical effects are smaller. Or, more simply put, if we rank the absolute value of all effects across domains from smallest to largest, the field effects make up the two endpoints with hypothetical effects in the middle (see SI Figure S11 for visualization).”

Minor:

1. While I do think the discussion is nicely concise, there is a lot of interpretation of results in the results section. Perhaps this is protocol for this journal, but I would prefer saving these interpretations for the discussion section.

Thank you for noting this! This is a great suggestion. We have integrated the last 6 paragraphs from the Results section—where the interpretation occurred—into the Discussion section. We agree with you that this more cleanly and appropriately separates the statistical results from

their interpretation. To better organize the Discussion section with this additional material, we have also added a sub-header called “Implications” where we pull together our interpretation of the results into a working list of guidelines for the reader.

2. Figure 2 itself speaks of a 5 points scale, the description below a 7 point scale. The supplementary material suggest a 5 point scale was used.

Our apologies! This was a typo. You are correct, we used a 5-point scale. This is corrected in the main text. Thank you for your attention to detail!

3. Page 10 says 'aforementioned regressions'. I may have missed it but I don't think regressions were mentioned before that point.

Our apologies! This was another typo from an earlier draft where we had discussed regressions at a high level before discussing their detail. Thank you for catching this! It is now corrected in the main text.

4. Language use can be slightly more consistent. Page 10 includes 'descriptors'. I believe this is the same as specificity earlier in the manuscript? Would be good to use language consistently.

Thank you for this suggestion to improve the readability of the paper! Yes, these are related: the descriptors in a study can be more or less specific. We’ve reviewed the language in both the main and supplemental text to make sure to pull these together more clearly and consistently. The main change we made was adding a parenthetical with “generic / specific” after the word “descriptors” on page 11 (old page 10); all other instances had the word specific and/or generic directly next to it (or on an image below it).

5. The coloured lines in figure 4 can be made thicker to make the figure more readable.

Thank you for pointing this out! In addition to increasing the bar thickness (and bolding the numbers, to better contrast), we think there was an issue when exporting images as .png files, that distorted their pixelation. We have now updated all figures in the manuscript—not just figure 4—by exporting as .pdfs and then inserting them, to preserve the quality of their pixelation. (Further, following publication guidelines, we have also adjusted the colors to better accommodate any readers with colorblindness or who wish to print the paper in grayscale.) We appreciate the feedback to ensure readers can better engage with the visuals in the paper!