**Peer Review Correspondence**

**Ms Title:** When 'Scientists say' coffee is good for you one day and bad for you the next:

Do generic attributions to ‘Scientists’ and ‘Experts’ amplify perceived conflict?

**Author names:** Matthew Haigh and Hope A. Birch

**Manuscript submitted**: April 10, 2020

**Editor First Decision—Revise & Resubmit**

Nov 1, 2020

Dear Dr. Haigh,

I have now received two reviews of your manuscript, “When ‘Scientists say’ coffee is good for you one day and bad for you the next: Do generic attributions to ‘Scientists’ and ‘Experts’ amplify perceived conflict?” from qualified researchers. I also independently read the manuscript before consulting these reviews. I agree that your manuscript has important strengths and also that there are some issues that need to be addressed. I therefore encourage you to submit a revised version for further consideration at Collabra: Psychology.

The reviewers did an outstanding job in their reviews, I won’t discuss each of their points here. Below I add a few other points or elaborate on some of the points raised in the reviews. In your resubmission, please include a document with a point-by-point response to both the points I list here and the reviewers’ comments, outlining each change made in your manuscript or providing a suitable rebuttal.

1. My biggest concern is about how to interpret the null results across your studies. Like Reviewer 1, I am concerned that the non-significant differences are difficult to interpret because of low power/precision. In the case of the medium-sized (but not significant) effects that Reviewer 1 calculated for Experiment 1, and the two interaction effects in Table 1 that are not-quite-significant, I am not sure that that these are necessarily meaningful results that you were underpowered to detect – part of the problem of low power/precision is that we can’t easily distinguish between those and medium/large effects that are not quite significant but are flukes (i.e., the true effect sizes is close to zero, but because of the lack of precision you happened to get a large estimate). My own priors are consistent with Reviewer 1’s (i.e., that these are likely real, meaningful effects), but I don’t think we can assume that. Thus, I worry that the informational value of Experiment 1 is currently quite limited – I am not sure how we can conclude much one way or the other about the effects of the “some” qualifier. To be honest, I don’t see a way to address this without collecting more data. I am open to other ideas you might have, but I suspect the best approach would be to run Experiment 1 again with a larger sample. If you do this, I would like you to also pre-register a plan for how to test for evidence of absence in the case of non-significant results (see my next point about Experiment 2 – to anticipate this possibility in a new study, I think it is a good idea to pre-register an equivalence test or Bayesian analysis to quantify evidence of absence).
2. For Experiment 2, the non-significant results seem clearer (i.e., the effects seem closer to zero), but it would still be important to quantify/test the amount of evidence you have for a zero/trivial effect (i.e., evidence of absence). How much precision do you have for these results? If you use non-informative priors, does a Bayesian analysis suggest these results are strong evidence for the null? There are a variety of ways to address this, but I think you need to provide readers with more information about what these results do or don’t mean.
3. Reviewer 1 mentions that the direction of the response scale is switched for one of the scales in Experiment 2 (Nutritional Backlash) compared to the other scales. I tried to find this information but could not find it (I assume it is in the Qualtrics file on the OSF page, but I don’t know how to open a Qualtrics file, which I know is a little bit ridiculous. However, for other readers like me, I think it would be good if you could include files with the materials that are easy to read within the browser). If this is in fact the case, that presents a major threat to the interpretability of the Experiment 2 results, in my opinion. If the scales were presented in the order that they are reported in Figure 2, that would mean that the scale was in one direction for the first measure (confusion), then switched directions for the second measure (backlash), then switched directions again for the third measure (mistrust). To my mind, this makes me doubt the interpretability of the results for at least the second and third measures, but possibly all measures after that as well. If this is what happened, I am not sure what to do about it. I would not be in favor of file-drawering this study, but I think it would be very difficult to know what to make of its results. I do agree with Reviewer 1 that most of us would probably have relatively low priors for the predicted effects for Experiment 2, because changing people’s views on such broad issues should be harder than just presenting them with 19 headlines read in 8-9 minutes (especially if we assume most people have a lot more than 8-9 minutes of prior experience with nutrition/health science news). So that may make the null results more plausible and less likely to be an artifact of the scale-switching aspect of the design. However, if the scales were switched I think this needs to be very clear to readers and needs to be taken into account in the interpretation of the Experiment 2 results.
4. More broadly, I think it is an interesting question how the results of Experiment 2 (assuming the negative results are interpretable and can be taken as evidence of absence) should impact your overall conclusions. Reviewer 1 argues that you should not put much weight on these results because they are not a particularly good test of your research question, because the outcome measures are too broad. As I mentioned above, I am sympathetic to the view that we should not expect it to be this easy to push around people’s broad, stable views about nutrition research in general. However, you presumably designed and ran the study because you did believe it would speak to your broad research question, and you thought the manipulation might impact the outcome variables. Personally, I felt that your conclusions (the abstract, and the discussion section) did not give much (if any) weight to the results of Experiment 2, and I disagree with Reviewer 1 that you should give these results even less weight in your overall conclusion. For example, in the first paragraph of the discussion, you write that “Across two experiments, we found that relative to those exposed to non-conflicting headlines, those exposed to conflicting headlines felt they were more contradictory, created more confusion and resulted in us knowing less about how to be healthy.” But, unless I am misunderstanding the results of Experiment 2, I don’t think you found this across two studies.

How you weight the results of Experiment 2 will depend a lot on how you deal with the issues raised in points 2 and 3 above, but putting those aside (if it turns out the negative results in Experiment 2 are interpretable and provide evidence of absence), I don’t think it is fair to discount them in your conclusion. Even though I think Reviewer 1 makes a good point that Experiment 2 might not be a great test of your broad research question, deciding that post hoc, after seeing the results, puts us at very high risk of bias. If you designed and ran the study thinking it would be informative, and then change your mind after seeing the results, I think you should provide quite an extensive discussion of why you changed your mind about the information value of this study, and clearly flag for readers that this interpretation goes against the perspective you had when you designed and ran the study, and so readers should be warned that this new interpretation could be motivated/biased. At a minimum, I would want you to change your description of the results in the discussion, but I hope you would go beyond this and “steel man” the interpretation of the Experiment 2 results.

1. Related to point 4 above, I think Reviewer 1’s question about whether you’ve conducted any other studies that test this same broad research question is a very important one. Please make sure to be completely transparent about this.
2. Please also make it very clear to readers which analytic (and any design) decisions were not pre-registered or deviated from the pre-registration. You should assume most readers will not check your pre-registration, and it is important to make sure that readers do not misinterpret a design or analysis decision as planned if it was not (or if the plan was ambiguous).
3. Related to some of the points raised above, it seems that you sometimes interpret p-values in the range that is sometimes called “marginal” (i.e., between .05 and .10) as evidence for an effect, and sometimes not. It would be better to be consistent about this, or, if you are going to selectively interpret some as evidence for a positive result and others not (or even as evidence of a negative result), that should explicitly acknowledge the reasons for this inconsistency (e.g., maybe it depends on your subjective priors?) should be articulated, so that readers can decide whether your reasons are compelling.
4. In the framing of your paper (introduction, discussion), there is an underlying issue that I think could be tackled more directly. Specifically, some of your language implies that it is necessarily a good thing to reduce people’s perception of conflict. Similarly, you seem to assume that people should ideally come to see that conflict is the result of real variation (e.g., in context, etc.) rather than the result of error or bias (things that should make them trust the process of science less). I am curious why you believe that this is obviously the case (or if you do not, then I think some of the language in your manuscript would need to be revised). Specifically, it seems possible to me that it is completely reasonable for non-scientists to be frustrated with inconsistency/conflict, it could be completely warranted to interpret that conflict as a sign of error or bias in the scientific process, and therefore for that conflict to make them trust the process of science less. When that is vs. is not the case (i.e., when contradictions/conflict do vs. don’t indicate a problem with the scientific process in a particular field) is a complex question, but it seems very bold to assume that it is not an appropriate/fair reaction to lose trust or be frustrated by conflicting headlines. If you do take this position, I think you should at least acknowledge that reasonable people could disagree, and that you are making some strong assumptions about what is and is not reasonable for a non-scientist to conclude on the basis of variable headlines/results.
5. Some of the intro and methods sections were quite repetitive (e.g., I think there are three almost identical descriptions of the design/summary of methods).
6. Please report effect sizes and confidence intervals throughout.

In summary, I think this is a promising manuscript and, I hope you will revise it for further consideration at Collabra: Psychology. I look forward to receiving your revision.

Please ensure that your revised files adhere to our author guidelines, and that the files are fully copyedited/proofed prior to upload. Please also ensure that all copyright permissions have been obtained. This is the last opportunity for major editing, therefore please fully check your file prior to re-submission.

If you have any questions or difficulties during this process, please contact the editorial office at editorialoffice@collabra.org.

We hope you can submit your revision within the next six weeks. If you cannot make this deadline, please let us know as early as possible.

Sincerely,

Simine Vazire Editor in Chief Collabra: Psychology

**Reviewer 1**

**Open response questions**

Please write your review here. The author(s) will see this review. Your identity will not be revealed to the authors unless you also include your name (i.e., sign your review) in this box. It is up to you whether to reveal your identity or not, either is fine.

This is a really interesting study, and I’m glad to be invited to review it. I think the idea is great and very intuitive and this is a pretty interesting finding. My main concern is that the authors are missing some effects because they are very underpowered.

I’ll go through some points below. I’ve also included some snippets of code here and there and some code at the end after the references which I used in my reanalysis of the data.

I’m very happy that the authors made their code and data public and that the studies were preregistered. This made evaluation pretty easy, but I do also have some comments regarding this.

Abstract

1. In the abstract, you call the formats Generic and Quantified format, but do you perhaps mean qualified format? It’s nit-picky, but quantified brings to mind something like “5 researchers say…” whereas you are “qualifying” the statement by indicating that “not all researchers say…”.
2. One quick note about the OSF page. In the future, I suggest creating a single project page and then creating multiple registrations on that page. This keeps everything organised in one place. Creating new pages for each ‘study’ can be confusing. This leads to my next comment…
3. Based on the data files on the OSF page, this also is called Study 7 and Study 8, but in the manuscript, this is referred to as Experiment 1 and 2. I’m wondering why this discrepancy occurs. Further, experiment 1 links to the prereg for ‘Study 8’ and Experiment 2 links to the prereg for ‘Study 7’. Please include a statement describing (and ideally linking to) other experiments that have been done in this line of work so that any potential overlap and file drawer effects and such can be evaluated.

Power analysis

1. In Study 1, you explain how you determined the sample size, but I believe there is an error here. It seems you calculated power for the interaction term rather than for the simple comparisons (ie using the “ANOVA: special effects, main effects, and interactions options”). The interaction term itself always requires less power, but the simple comparisons are where the power is needed. For example, if you have a 2x2 with 300 people, that’s 75 per cell; you would need to determine the smallest effect you could detect in a comparison of two cells with n = 75 (e.g. N = 150). This is a d = .70/f = .35, 99% of the time with a two-tailed test, which is much larger than any social psychological effect on average. To detect an f = .25/d=.50 99% of the time (still much larger than most effects), you would need 148 in each condition, so 148\*4 = 592 total, as per the below code:

library(pwr)

pwr.t.test(d = .50, sig.level = .05, power = .99)

This is for your future reference, but also has important bearing on the results you report here.

Because, for example, in Table 1, two of your interactions are around p = .06, suggesting that you might be a little low on power. Some of your main effects are significant, because these comparisons are pooled across two of the cells, yielding increased power. However, you don’t report effect sizes here. More importantly, the lack of power resulted in you not comparing the simple effects, which I show below have some fairly large effects but you didn’t pick them up because of low power. See point 9.

1. Please report effect sizes throughout the manuscript. R and/or cohen’s d or standardized mean differences would be preferred over eta-squared.
2. The nutritional backlash items (experiment 2) seem to be scaled in the opposite direction (eg 1 is strongly agree and 5 is strongly disagree). In the past, I’ve encountered issues where subjects don’t notice the scale labels when they are switched. How can you address this possibility?
3. The analysis plan in the preregistration does not completely outline the analyses to be conducted. It simply says “2x2 Independent groups ANOVA. If there is a significant interaction this will be decomposed (simple main effects). ” and specifies and alpha level for the interaction and “simple main effects”. So, it’s not clear what actually counts as a preregistered analysis- for example, the one-sample t-tests on the ‘scientific advancement’ measure were not preregistered. I’m not sure I agree with that analysis, anyways: I want to know whether they differ from each other, not whether they are different from 0 on their own.
4. There were virtually no significant effects in Experiment 2 and the graphs clearly show that the means across the four conditions are basically identical. However, I am concerned that these measures go a step beyond what would normally be influenced by these kinds of materials and manipulations. The questions (mistrust of expertise, confidence in science, epistemic beliefs) seem to be really general and broad evaluations of science and knowledge. I have serious doubts that a one-shot survey manipulation would change people’s evalautions of knowledge. I would not conclude that the headline manipulation “had no effect” simply because it did not influence how people responded to these very broad measures. I wouldn’t expect much (anything) to influence these measures, in fact. I wouldn’t really expect a change in “nutritional backlash” or “nutritional confusion”, either, though on the surface they seem more reasonable than the other 3 measures. The issue here is that you are asking participants to generalise a few nutritional headlines from an online survey into a global evaluation of “Dietary recommendations are not useful” or “Nutrition recommendations are confusing.” Well, some are and some aren’t; a single online survey is not going to influence anyone’s evaluation of this. So, in sum, I wouldn’t really draw any conclusions about your overall research question from Experiment 2. Include it in the report, please, but I would focus more on refining Experiment 1.
5. There are some problems with your analysis code which make it difficult to reproduce your analyses exactly in ‘Study 7’. Essentially, your subset arguments return a data frame with 0 observations. Not sure what the deal is there, but I spent most of my time on Experiment 1 (Study 8) anyways.

Based on Figure 1, you can see that there is an interaction such that the non-conflict conditions differ (visually) from the conflicting headlines between the generic and quantified cells. If we select those cells and compare them, we find fairly large effect sizes. For non-conflicting headlines, generic headlines versus quantified resulted in: reduced contradiction (d = .40!) Greater confusion (d = .26!) Less advancement (d = .27!)

For the conflicting headlines, generic headlines versus quantified resulted in: Similar levels of contradiction (d = .03) More confusion (d = .16) Less advancement (d = .19) I’ll append the code for this at the very end of this review, just so it doesn’t get cluttered.

These results are mostly in line with your hypotheses and a common sense explanations, but you would not detect them because you are very low on power – remember you were powered for d = .70. Most social psychological effects are much, much smaller, within the range of the effects I just reported (see: Funder & Ozer, 2019; Gignac & Szodorai, 2016). Your average effect is going to be around a d =.20,which requires more people. It might seem like a lot of people, but that’s just because most past research was terrible underpowered (and likely hacked) and underpowered studies inflates effect sizes. In general, we should have much larger sample sizes.

So, to detect a d=.20 80% of the time, I’d use a one-tailed test. You’d need 309 people in each condition. To detect the smaller effect of d = .15, I’d again use a one-tailed t-test. You’d need 550 in each condition.

library(pwr)

pwr.t.test(type= "two.sample", d = .15, sig.level = .05, power = .80, alternative = "greater")

So, I hope that the authors will address the above concerns. I’d like to see this manuscript published because I think it is interesting and it is very intuitive. But I also don’t want some important effects to be missed because the study was underpowered. I’d say the effects from experiment 1 should be replicated with a larger sample size, but I guess that decision is ultimately up to the editor.

References: Funder, D. C., & Ozer, D. J. (2019). Evaluating Effect Size in Psychological Research: Sense and Nonsense. Advances in Methods and Practices in Psychological Science, 2(2), 156–168. <https://doi.org/10.1177/2515245919847202> Gignac, G. E., & Szodorai, E. T. (2016a). Effect size guidelines for individual differences researchers. Personality and Individual Differences, 102(November 2016), 74–78. <https://doi.org/10.1016/j.paid.2016.06.069>

Code: I used your script for Study 8, and load the resulting mydata.csv file here:

NC <- dplyr::filter(mydata, Conflict == "Non-Conf.") # NC = no conflict

CF <- dplyr::filter(mydata, Conflict == "Conf.") #CF = conflict

NC <- dplyr::select(NC, Format, contradiction, confusion, advancement)

CF <- dplyr::select(CF, Format, contradiction, confusion, advancement)

psych::cohen.d(NC ~ Format)

psych::cohen.d(CF ~ Format)

**Rating scale questions**

|  | Strongly Disagree | Disagree | Neutral | Agree | Strongly Agree |
| --- | --- | --- | --- | --- | --- |
| The study/studies in this manuscript have strong construct validity (good measures and/or manipulations of the constructs the authors wish to study). (Choose “Neutral” if this is not an empirical manuscript) |  |  |  |  | ✔ |
| The study/studies in this manuscript have strong statistical validity (appropriate statistical tests, assumptions are clear and reasonable, no statistical errors, appropriate statistical inferences, etc.). (Choose “Neutral” if this is not an empirical manuscript) |  |  |  | ✔ |  |
| The study/studies in this manuscript have strong internal validity (any causal claims or implications are well-justified, alternative explanations are thoroughly considered, etc.). (Choose “Neutral” if this is not an empirical manuscript, or no causal claims are made or even vaguely implied.) |  |  | ✔ |  |  |
| The study/studies in this manuscript have strong external validity (authors appropriately constrain their conclusions based on the limits of the generalizability of their findings to other contexts (including from lab to real world), other populations, other stimuli or measures, etc.) |  |  | ✔ |  |  |

**Reviewer 2**

**Open response questions**

Please write your review here. The author(s) will see this review. Your identity will not be revealed to the authors unless you also include your name (i.e., sign your review) in this box. It is up to you whether to reveal your identity or not, either is fine.

The authors present an interesting study of the potential impact of conflict (whether participants read two headlines about nutrition research findings that conflicted with each other) and generic language (“experts say” vs. “some experts say”). They found that participants in Experiment 1 did notice the contradiction between conflicting headlines and found those headlines more confusing and less reflective of advancing science. However, conflict did not influence global beliefs about health. Generic language was not significant. I learned a lot just from the questionnaires you included!

I have a few questions/comments that might benefit from minor revision:

1. Pg. 8: “These specific headlines do not directly contradict each other.” This is true for the ones you list, but less true for others in this condition. For instance, “Everyone needs to take vitamin D supplements …” and “taking vitamin D is pointless” seem to be in direct opposition to one another (I’d say the same for your meat category). I’m not sure how important this is for highlighting conflict. Did participants differ at the item level? For instance, were the vitamin D ratings systematically higher in conflict/confusion or lower in scientific advancement than the alcohol ratings?
2. Do you have any demographic information (besides age/gender) to report about your participants?
3. Methods: It would be helpful to list the descriptors (if any) for the midpoint of each scale you sued (similar to “confidence in the scientific community”), or say explicitly that only the end points of the scale were marked. This would be especially helpful for Exp. 2, where a lot of the ratings look to be hovering around the midpoint.
4. Would ordinal regressions be more appropriate tests for your data than ANOVAs?

Other points to consider for the discussion or future studies:

Study 1: really interesting: “The mean of those exposed to non-conflicting headlines (-0.007) did not significantly differ from zero (t(146) = 0.12, p= 0.902).” From the histogram it’s clear that “same as” was the modal response (not that some people thought we knew more and some thought we knew less). I wonder what it would take for people to say that we know more than before. I’d be curious about the extent to which participants view themselves as knowledgeable about nutrition already (compared to other fields of scientific inquiry). I’m not sure how much any ratings in Exp. 2 speak to this, but at least in terms of “nutritional confusion,” participants are typically in the middle (rather than reporting high confusion, even when exposed to conflicting findings).

I would be curious how introducing a quantifier into who showed the finding (e.g., “in some people”) would contrast with your manipulation of qualifying who is making the claim (e.g., “some scientists”). Generics also gloss over variability within participants. This strikes me as especially important in nutrition research (e.g., the fasting diet could have those two effects listed on pg. 4, but for different individuals or groups of people).

Unrelated to your study, I’m curious how well the 59 participants in Exp. 1 (who did the study twice) matched in their responses from the first attempt.

**Rating scale questions**

|  | Strongly Disagree | Disagree | Neutral | Agree | Strongly Agree |
| --- | --- | --- | --- | --- | --- |
| The study/studies in this manuscript have strong construct validity (good measures and/or manipulations of the constructs the authors wish to study). (Choose “Neutral” if this is not an empirical manuscript) |  |  |  | ✔ |  |
| The study/studies in this manuscript have strong statistical validity (appropriate statistical tests, assumptions are clear and reasonable, no statistical errors, appropriate statistical inferences, etc.). (Choose “Neutral” if this is not an empirical manuscript) |  |  | ✔ |  |  |
| The study/studies in this manuscript have strong internal validity (any causal claims or implications are well-justified, alternative explanations are thoroughly considered, etc.). (Choose “Neutral” if this is not an empirical manuscript, or no causal claims are made or even vaguely implied.) |  |  |  | ✔ |  |
| The study/studies in this manuscript have strong external validity (authors appropriately constrain their conclusions based on the limits of the generalizability of their findings to other contexts (including from lab to real world), other populations, other stimuli or measures, etc.) |  |  |  | ✔ |  |

**Author Response**

Jan 18, 2021

Dear Professor Vazire,

Thank you for the opportunity to revise and resubmit our manuscript (*When 'Scientists say' coffee is good for you one day and bad for you the next….*). We found the comments very helpful. We have addressed all comments point-by-point and have made corresponding revisions to the manuscript as required. Thank you for considering our manuscript and we hope it is now appropriate for publication in Collabra:Psychology.

Best wishes

Matthew Haigh & Hope Birch

**Editor and Reviewer comments are reproduced in full below. Our replies are highlighted in grey.**

**Editor Comments**

I have now received two reviews of your manuscript, “When ‘Scientists say’ coffee is good for you one day and bad for you the next: Do generic attributions to ‘Scientists’ and ‘Experts’ amplify perceived conflict?” from qualified researchers. I also independently read the manuscript before consulting these reviews. I agree that your manuscript has important strengths and also that there are some issues that need to be addressed. I therefore encourage you to submit a revised version for further consideration at Collabra: Psychology.

Thank you for the opportunity to revise and resubmit. We have addressed each of the comments and made corresponding revisions to the paper where necessary.

The reviewers did an outstanding job in their reviews, I won’t discuss each of their points here. Below I add a few other points or elaborate on some of the points raised in the reviews. In your resubmission, please include a document with a point-by-point response to both the points I list here and the reviewers’ comments, outlining each change made in your manuscript or providing a suitable rebuttal.

1. My biggest concern is about how to interpret the null results across your studies. Like Reviewer 1, I am concerned that the non-significant differences are difficult to interpret because of low power/precision. In the case of the medium-sized (but not significant) effects that Reviewer 1 calculated for Experiment 1, and the two interaction effects in Table 1 that are not-quite-significant, I am not sure that that these are necessarily meaningful results that you were underpowered to detect – part of the problem of low power/precision is that we can’t easily distinguish between those and medium/large effects that are not quite significant but are flukes (i.e., the true effect sizes is close to zero, but because of the lack of precision you happened to get a large estimate). My own priors are consistent with Reviewer 1’s (i.e., that these are likely real, meaningful effects), but I don’t think we can assume that. Thus, I worry that the informational value of Experiment 1 is currently quite limited – I am not sure how we can conclude much one way or the other about the effects of the “some” qualifier. To be honest, I don’t see a way to address this without collecting more data. I am open to other ideas you might have, but I suspect the best approach would be to run Experiment 1 again with a larger sample. If you do this, I would like you to also pre-register a plan for how to test for evidence of absence in the case of non-significant results (see my next point about Experiment 2 – to anticipate this possibility in a new study, I think it is a good idea to pre-register an equivalence test or Bayesian analysis to quantify evidence of absence).

Below, we have replied in detail to Reviewer 1’s comments about power of Experiment 1 and what this means for our non-significant results. To avoid repetition, we will summarise our argument here and full details can be found in response to Reviewer 1. In short, we believe that our level of power is not as low as suggested by Reviewer 1 and that the not-quite-significant interactions (p=.066) are not actually driven by the effects we predicted. Because these effects are non-significant and inconsistent with our predictions, we believe they should only be interpreted with the extreme caution, if at all.

Reviewer 1 was concerned that Experiment 1 was underpowered and may have therefore missed important effects (i.e., some of the null results may have actually been false negatives). The effects of particular concern to Reviewer 1 were the interaction between Format and Conflict in Experiment 1 (both the confusion and advancement DVs had interaction terms of *p*=.066, which might be considered ‘marginal’). We had a high level of power to detect an interaction effect. This was sufficient to detect a ‘medium’ sized interaction effect >99% of the time (or a smaller effect of *d*=0.32 80% of the time). However, despite having a high level of power the interaction *p* values were non-significant (*p*>.05). Additional analysis reported in the revised manuscript further supports our conclusion that there was no meaningful effect. First, this analysis shows that the effect size confidence intervals for the interaction effects included zero (i.e., no effect). Second, Bayes factors for the interaction terms provide substantial evidence in favour of H0 (this additional analysis was carried out in response your point 2). For these three reasons, we are confident that we did not miss any meaningful interaction effects.

Because the interactions were non-significant and had small effect sizes (with confidence intervals that included zero) we did not conduct any post-hoc tests. In the pre-registration we stated that post-hoc tests would only be carried out if an interaction was significant. However, Reviewer 1 correctly observed that our level power for the post-hoc tests (simple effects) was lower than the power we had to detect main effects or interactions. This is arguably a moot point as our omnibus test was non-significant and therefore post-hocs were not justified. Regardless of whether or not we should have conducted the post-hocs, the issue with power to detect simple effects was not as severe as it may have seemed. Reviewer 1 calculated the largest simple effects we would be able to detect in Experiment 1 and concluded that the largest effect we could have detected was d=0.7 (i.e., larger than most effects in psychology). They were therefore concerned that we may have missed effects smaller than this. Importantly, this calculation was based on the assumption that our desired level of power was 0.99. In other words, we had power to detect an effect of d=0.7 (a ‘large’ effect) 99% of the time. While power of 0.99 would be ideal, this is not often achieved and not what we aimed to achieve. If we change the desired level of power to the widely used minimum of 0.8 (Cohen, 1988) it reveals that we had power to detect simple effects of d=0.46 (a ‘medium’ effect) 80% of the time. Reviewer 1 pointed out that the effect sizes of the simple effects were smaller than this (in the region of d=0.2) but importantly, these effects were not in the direction we predicted (this is discussed in reply to R1). For these reasons we have chosen not to interpret ‘marginal’ interactions as they are driven by small effects that we did not predict. We have opted not to re-run Experiment 1 because the results based on 300 participants give us little evidence for the interaction effect we predicted.

1. For Experiment 2, the non-significant results seem clearer (i.e., the effects seem closer to zero), but it would still be important to quantify/test the amount of evidence you have for a zero/trivial effect (i.e., evidence of absence). How much precision do you have for these results? If you use non-informative priors, does a Bayesian analysis suggest these results are strong evidence for the null? There are a variety of ways to address this, but I think you need to provide readers with more information about what these results do or don’t mean.

Thank you, we agree that it is important to provide evidence of absence in this situation. In addition to our pre-registered ANOVAs we have now also conducted Bayesian ANOVAs (with non-informative priors) and reported the associated Bayes Factors (BF10) in Tables 1 and 2. Because this analysis was not pre-registered, we have labelled it as Exploratory. This exploratory analysis has helped us to quantify evidence for the null hypothesis (H0). In Experiment 2 there is ‘substantial’ evidence on favour of H0 for all main effects and interactions (the greatest BF10 was 0.3). In Experiment 1, there were clear effects of our conflict manipulation with decisive evidence in favour of H1 (BF10 values in excess of 100). For our Format manipulation (generic vs qualified) the evidence was in favour of H0. The same is true for the interaction terms. As discussed above, two of the interaction terms were marginally non-significant when analysed using traditional ANOVA (both interactions had a *p* value of .066) but the Bayes factors indicate that we should not attempt to interpret these interactions as the data are ‘substantially’ more likely under H0 (the BF10 values of these marginal interaction terms were .16 and .37). This additional analysis has allowed to us to strengthen our conclusions about the absence of effects.

1. Reviewer 1 mentions that the direction of the response scale is switched for one of the scales in Experiment 2 (Nutritional Backlash) compared to the other scales. I tried to find this information but could not find it (I assume it is in the Qualtrics file on the OSF page, but I don’t know how to open a Qualtrics file, which I know is a little bit ridiculous. However, for other readers like me, I think it would be good if you could include files with the materials that are easy to read within the browser).

In addition to the .qsf files (which allow anyone with a Qualtrics account to see the survey exactly as participants saw it and to quickly conduct an exact replication) we have also uploaded to the OSF Microsoft Word versions of the Experiments to give an approximation of what participants saw. In response to your comments immediately below (relating to the direction of the scales) we also include some selected screenshots in this letter, so that you can see exactly what participants saw.

If this is in fact the case, that presents a major threat to the interpretability of the Experiment 2 results, in my opinion. If the scales were presented in the order that they are reported in Figure 2, that would mean that the scale was in one direction for the first measure (confusion), then switched directions for the second measure (backlash), then switched directions again for the third measure (mistrust). To my mind, this makes me doubt the interpretability of the results for at least the second and third measures, but possibly all measures after that as well. If this is what happened, I am not sure what to do about it. I would not be in favor of file-drawering this study, but I think it would be very difficult to know what to make of its results.

We anticipated that the change in scale direction might be missed by some participants and we will explain below how we robustly mitigated this risk when designing the study. For this reason, we are very confident in the validity/interpretability of our data.

Your comment about switching the scales is based on a comment from Reviewer 1 (R1, point 6). Their concern was that some participants might not notice when the scales changed direction (for example, a participant might erroneously select ‘Strongly Agree’ when intending to select ‘Strongly Disagree’). Because we wanted to present each scale in a way that was faithful to the original, it meant that some scales were labelled differently to others. When setting up the study we anticipated that this change in scale direction might be missed by some participants, so we made sure to address this in two ways. First, each scale was presented on a separate page with every option labelled using words rather than numbers (see screenshot below).



Second, when a participant moved on to a new page (to complete a new scale) we included a clear message in bold if the scale changed direction (e.g., “***Note that the order of the scale has changed:*** *Strongly agree is on the left and strongly disagree on the right”*). An example screenshot is below.



For these two reasons, we are confident that participants will have noticed the change in labelling between the scales. We are also confident that participants in the final sample paid attention to these instructions. This is because participants were only included in the final sample if they passed an attention check and declared that they answered seriously.

I do agree with Reviewer 1 that most of us would probably have relatively low priors for the predicted effects for Experiment 2, because changing people’s views on such broad issues should be harder than just presenting them with 19 headlines read in 8-9 minutes (especially if we assume most people have a lot more than 8-9 minutes of prior experience with nutrition/health science news). So that may make the null results more plausible and less likely to be an artifact of the scale-switching aspect of the design. However, if the scales were switched I think this needs to be very clear to readers and needs to be taken into account in the interpretation of the Experiment 2 results.

We have now made it clear in the method section of Experiment 2 that the scales were ordered differently and that we took pre-emptive action to mitigate the risk of participants not noticing the switch.

1. More broadly, I think it is an interesting question how the results of Experiment 2 (assuming the negative results are interpretable and can be taken as evidence of absence) should impact your overall conclusions. Reviewer 1 argues that you should not put much weight on these results because they are not a particularly good test of your research question, because the outcome measures are too broad. As I mentioned above, I am sympathetic to the view that we should not expect it to be this easy to push around people’s broad, stable views about nutrition research in general. However, you presumably designed and ran the study because you did believe it would speak to your broad research question, and you thought the manipulation might impact the outcome variables.

That is correct – we explain below how our manipulation and outcome measures were grounded in previous research (i.e., studies using similar manipulations have demonstrated these manipulations can affect the type of outcome measure we used).

Personally, I felt that your conclusions (the abstract, and the discussion section) did not give much (if any) weight to the results of Experiment 2, and I disagree with Reviewer 1 that you should give these results even less weight in your overall conclusion. For example, in the first paragraph of the discussion, you write that “Across two experiments, we found that relative to those exposed to non-conflicting headlines, those exposed to conflicting headlines felt they were more contradictory, created more confusion and resulted in us knowing less about how to be healthy.” But, unless I am misunderstanding the results of Experiment 2, I don’t think you found this across two studies.

Supported by the new Bayesian analysis we have now given more weight to the results of Experiment 2 in the Abstract and Discussion, but made it clear that the absence of a Conflict effect meant we could test our primary hypothesis in this experiment (i.e., because there was no effect of our Conflict manipulation we could examine whether Format was able to moderate Conflict). We have also clarified exactly what was found in each study. In the example quote you use above we created ambiguity by trying to summarise the results of both studies in a single sentence, so we have clarified this in the manuscript by always discussing the Experiments separately.

How you weight the results of Experiment 2 will depend a lot on how you deal with the issues raised in points 2 and 3 above, but putting those aside (if it turns out the negative results in Experiment 2 are interpretable and provide evidence of absence), I don’t think it is fair to discount them in your conclusion. Even though I think Reviewer 1 makes a good point that Experiment 2 might not be a great test of your broad research question, deciding that post hoc, after seeing the results, puts us at very high risk of bias. If you designed and ran the study thinking it would be informative, and then change your mind after seeing the results, I think you should provide quite an extensive discussion of why you changed your mind about the information value of this study, and clearly flag for readers that this interpretation goes against the perspective you had when you designed and ran the study, and so readers should be warned that this new interpretation could be motivated/biased. At a minimum, I would want you to change your description of the results in the discussion, but I hope you would go beyond this and “steel man” the interpretation of the Experiment 2 results.

In response to Reviewer 1 (point 8) we explained how our hypothesis was grounded in the findings of published research. Previous studies have shown that similar manipulations have shifted ratings on some of the same scales used in Experiment 2. To be clear, we did not expect that brief exposure to conflicting headlines would have any lasting effect on global beliefs, but we did think it was plausible that priming participants with conflicting information immediately before completing questionnaires about nutrition/science beliefs might be enough to temporarily shift their responses. As an example, it seemed plausible to us that participants recently exposed to conflicting headlines might be less inclined to agree with scale items such as “*It is not always clear to me what foods are best for me to eat*”.

On seeing the Experiment 2 data we have updated our beliefs. There was clearly no evidence that our manipulation affected any of the global belief scales. We have revised the paper to make it clearer that our predictions were grounded in previous research and were justified based on what we knew prior to data collection. We have also updated our conclusions to be clear that that the Experiment 2 data provide no evidence to support our predictions. In hindsight (after seeing the data) it is easy to say that our conflict manipulation was never going to affect global beliefs, but we feel we were justified in making that prediction at the time.

1. Related to point 4 above, I think Reviewer 1’s question about whether you’ve conducted any other studies that test this same broad research question is a very important one. Please make sure to be completely transparent about this.

We have not conducted any other studies directly relevant to this research question. We were awarded funding by the Leverhulme Trust to conduct eight experiments focused on the interpretation of generic headlines and only two of those experiments are relevant to the current submission. Prior to this, we had never conducted any research on generics. Below is a brief summary of the funded experiments.

* Two studies focused on individual differences in the interpretation of generics (these have been published in Collabra and that paper is cited in the manuscript).
* Two studies looking at the effect of generics on perceived conflict (in the domain of nutrition). These are the experiments reported in our current submission.
* Two studies looking at the degree of consensus implied by generic headlines (these are currently being written up as a single paper)
* Two studies will examine the pragmatic inferences made from generic headlines (data collection due to start very soon).
* There is also a PhD student in our lab (funded from a different source) who is currently conducting an experiment on the interpretation of generics used in a health communication context.

All of the completed studies were preregistered prior to data collection. All new studies will also be pre-registered and data shared publicly on the OSF. None of the other experiments are directly relevant to the current submission. All of the research we have conducted is in the public domain, so nothing has been or will be left in the file drawer.

1. Please also make it very clear to readers which analytic (and any design) decisions were not pre-registered or deviated from the pre-registration. You should assume most readers will not check your pre-registration, and it is important to make sure that readers do not misinterpret a design or analysis decision as planned if it was not (or if the plan was ambiguous).

We have now split the Results into ‘pre-registered’ and ‘exploratory’ analysis.

1. Related to some of the points raised above, it seems that you sometimes interpret p-values in the range that is sometimes called “marginal” (i.e., between .05 and .10) as evidence for an effect, and sometimes not. It would be better to be consistent about this, or, if you are going to selectively interpret some as evidence for a positive result and others not (or even as evidence of a negative result), that should explicitly acknowledge the reasons for this inconsistency (e.g., maybe it depends on your subjective priors?) should be articulated, so that readers can decide whether your reasons are compelling.

In the revised paper we no longer refer to ‘marginal’ effects. Any effect with a *p* value >.05 is treated as non-significant, consistent with our pre-registered analysis plan. We also report Bayes Factors which show that all of our non-significant effects provide evidence in support of H0.

1. In the framing of your paper (introduction, discussion), there is an underlying issue that I think could be tackled more directly. Specifically, some of your language implies that it is necessarily a good thing to reduce people’s perception of conflict. Similarly, you seem to assume that people should ideally come to see that conflict is the result of real variation (e.g., in context, etc.) rather than the result of error or bias (things that should make them trust the process of science less). I am curious why you believe that this is obviously the case (or if you do not, then I think some of the language in your manuscript would need to be revised).

Specifically, it seems possible to me that it is completely reasonable for non-scientists to be frustrated with inconsistency/conflict, it could be completely warranted to interpret that conflict as a sign of error or bias in the scientific process, and therefore for that conflict to make them trust the process of science less. When that is vs. is not the case (i.e., when contradictions/conflict do vs. don’t indicate a problem with the scientific process in a particular field) is a complex question, but it seems very bold to assume that it is not an appropriate/fair reaction to lose trust or be frustrated by conflicting headlines. If you do take this position, I think you should at least acknowledge that reasonable people could disagree, and that you are making some strong assumptions about what is and is not reasonable for a non-scientist to conclude on the basis of variable headlines/results.

We have edited the manuscript to make it clearer that we do not aim to reduce the perception of conflict per se. The idea behind these studies was that generic headlines, which are commonly used by the media, may artificially inflate the perception of conflict. Our aim was to find out if the use of qualifiers would reduce the perception of conflict to a more realistic level. We shall explain with an example.

Previous research has indicated that generic attributions to experts (e.g., *Researchers believe*….) imply on average that a majority of experts hold that opinion (I.e., *Researchers believe* = *Most researchers believe*…). The risk of generic headlines such as 1 and 2 below is that they are interpreted in the same way as headlines 3 and 4.

1. *Fasting diet could regenerate pancreas and reverse diabetes, researchers say*
2. *Fasting diets may raise risk of diabetes, researchers warn*
3. *Fasting diet could regenerate pancreas and reverse diabetes,* ***MOST*** *researchers say*
4. *Fasting diets may raise risk of diabetes,* ***MOST*** *researchers warn*

Interpreting headlines 1 and 2 in the same way as 3 and 4 implies a dramatic U-Turn or lack of direction within the field (which may impact on outcomes such as trust and confusion). This degree of conflict is clearly inflated. It gives the impression that ‘experts say one thing one day and another the next…”. This is not the reality in most situations. Generally, when there is conflict in science it relates to the findings of one research group conflicting the with findings of another. Our Format manipulation is therefore intended to give a more accurate perception of conflict by implying that the conflict is *between* different researchers (some researchers believe one thing and some believe another). We did this by inserting the qualifier ‘some’ – see example 5 and 6.

1. *Fasting diet could regenerate pancreas and reverse diabetes,* ***some*** *researchers say*
2. *Fasting diets may raise risk of diabetes,* ***some*** *researchers warn*

Our aim was not to reduce or remove the perception of conflict per se, but to make the perception of conflict more accurate (i.e., incremental disagreement between specific experts rather a wholesale U turn by the field of experts).

1. Some of the intro and methods sections were quite repetitive (e.g., I think there are three almost identical descriptions of the design/summary of methods).

We have now streamlined the introduction and Method to reduce repetition.

1. Please report effect sizes and confidence intervals throughout.

We now report effect sizes for all effects with confidence intervals around these effect sizes. We also plot confidence intervals around the condition means in Figures 1 and 3.

In summary, I think this is a promising manuscript and, I hope you will revise it for further consideration at Collabra: Psychology. I look forward to receiving your revision.

**Reviewer 1**

This is a really interesting study, and I’m glad to be invited to review it. I think the idea is great and very intuitive and this is a pretty interesting finding.

Thank you - we are delighted that you like the idea. We appreciate the time and effort you put in to review the OSF pages and analysis scripts in addition to the reviewing paper itself.

My main concern is that the authors are missing some effects because they are very underpowered.

We hope that our replies below will allay your concerns about power.

I’ll go through some points below. I’ve also included some snippets of code here and there and some code at the end after the references which I used in my reanalysis of the data.

I’m very happy that the authors made their code and data public and that the studies were preregistered. This made evaluation pretty easy, but I do also have some comments regarding this.

Abstract

1. In the abstract, you call the formats Generic and Quantified format, but do you perhaps mean qualified format? It’s nit-picky, but quantified brings to mind something like “5 researchers say…” whereas you are “qualifying” the statement by indicating that “not all researchers say…”.

We agree that ‘qualified’ is more suitable term than ‘quantified’ and have changed this throughout the manuscript.

1. One quick note about the OSF page. In the future, I suggest creating a single project page and then creating multiple registrations on that page. This keeps everything organised in one place. Creating new pages for each ‘study’ can be confusing. This leads to my next comment…

We completely agree. This was our first experience of using the OSF. When we started this project, it seemed sensible to have a separate page for each of our eight funded experiments but in retrospect it makes much more sense to have one OSF page for each multi-experiment paper. We will be sure to do that in future.

1. Based on the data files on the OSF page, this also is called Study 7 and Study 8, but in the manuscript, this is referred to as Experiment 1 and 2. I’m wondering why this discrepancy occurs. Further, experiment 1 links to the prereg for ‘Study 8’ and Experiment 2 links to the prereg for ‘Study 7’. Please include a statement describing (and ideally linking to) other experiments that have been done in this line of work so that any potential overlap and file drawer effects and such can be evaluated.

There is a simple explanation here (see also our reply to the Editor). We were funded to run 8 experiments on generics. In our funding application we numbered the experiments 1-8. In an attempt to be organised we created eight OSF pages as soon as the funding was awarded (labelled study 1 through to Study 8). In retrospect, it would have made more sense to create one OSF page for each multi-experiment paper.

The two experiments reported in the current submission were labelled 7 and 8 in the funding application. We have now updated the OSF pages to make the labelling of these Experiments consistent with the labelling used in the paper.

These are the only two studies that we conducted on Conflict/Nutrition and these were always intended to make up a single two-experiment paper. Our other work on generics is not directly relevant to this paper and will be submitted as separate papers (for example, two of the studies were recently published as a paper in Collabra). Please rest assured that no studies are being left in the file drawer.

While we are discussing transparency, it is worth noting one other factor that might help clear up some of the confusion over study labels. Experiment 2 in the paper (Experiment 8 on OSF) was actually conducted before Experiment 1 (Experiment 7 on OSF). We switched the presentation order in the paper for narrative purposes. We felt it more logical to present results about the effect of conflict on specific beliefs before presenting the effect on global beliefs.

Power analysis

1. In Study 1, you explain how you determined the sample size, but I believe there is an error here. It seems you calculated power for the interaction term rather than for the simple comparisons (ie using the “ANOVA: special effects, main effects, and interactions options”). The interaction term itself always requires less power, but the simple comparisons are where the power is needed.

Thank you for pointing this out – we did indeed calculate power for the interaction term using this method (“ANOVA: special effects, main effects, and interactions options”). We agree that it was an error to not explicitly consider the power of the simple comparisons when deciding our sample size. Each simple comparison is based on two cells (roughly half the sample) so will inevitably have lower power than main effects or interactions, which are based on the full sample. While not ideal, we hope to demonstrate below that this does not detract from our conclusions.

For example, if you have a 2x2 with 300 people, that’s 75 per cell; you would need to determine the smallest effect you could detect in a comparison of two cells with n = 75 (e.g. N = 150). This is a d = .70/f = .35, 99% of the time with a two-tailed test, which is much larger than any social psychological effect on average.

To double check our understanding we first reproduced your calculation:

pwr.t.test(n=75, power=0.99, sig.level = .05) #min.detectable effect = 0.7

This calculation is of course correct, but importantly it assumes minimum power of 0.99, rather the more commonly used level of 0.8. We can completely see why you used the figure of 0.99 because we stated in our pre-registration and in the manuscript that we had >0.99 power (to detect an interaction effect). I think we caused confusion here due to the way we determined our sample size. Rather than determining sample based on desired power and anticipated effect size, we used a more pragmatic method. We knew how many participants we could afford to recruit (n=75 /N=300) and intended to recruit all of these to maximise power. We therefore calculated the power that a sample of 300 would give us to detect a ‘medium’ sized effect (in this case >0.99). The power level of >0.99 was therefore a description of the power we achieved, rather than a minimum desired level of power.

If we change the desired level of power to 0.8 it tells us we had power to detect a ‘medium’ sized (d=0.46) simple effects 80% of the time.

pwr.t.test(n=75, power=0.8, sig.level = .05) #min.detectable effect = 0.46

We acknowledge that many effects in psychology are smaller than d=0.46, but we hope to have demonstrated that the power to detect simple effects was not as low as it initially appeared.

To detect an f = .25/d=.50 99% of the time (still much larger than most effects), you would need 148 in each condition, so 148\*4 = 592 total, as per the below code:

library(pwr)

pwr.t.test(d = .50, sig.level = .05, power = .99)

As with our point above, if we change the desired level of power to 0.8 our sample size looks more reasonable.

pwr.t.test(d = .50, sig.level = .05, power = .8) #n=63.8

This calculation indicates that we would need a minimum of 64 participants per group. We actually recruited more than this (75 per group). However, we acknowledge your point that effects in psychology are often smaller than d=0.5.

This is for your future reference, but also has important bearing on the results you report here.

Because, for example, in Table 1, two of your interactions are around p = .06, suggesting that you might be a little low on power.

Here we are discussing the power of the interaction term (above we discussed the power of the simple effects). To retrospectively check whether we were low on power we used GPower (see Figure 1 screenshot below) to determine the smallest interaction effect we could have reliably detected.

Figure 1: GPower sensitivity analysis for Experiment 1



This tells us that with 300 participants we had sensitivity to detect an interaction of at least f=0.162/d=0.324 80% of time. An effect of d=0.32 is midway between a ‘small’ (d=0.2) and ‘medium’ (d=0.5) effect size. Prior to running the study this would have seemed a reasonable level of power. However, we acknowledge that the observed *p* values of .066 do call this into question – with a larger sample it is possible that these *p* values would have fallen below .05. It is also possible that the *p* value could go the other way with a larger sample, as the confidence intervals for the interaction effect sizes included zero. In support of this interpretation, exploratory Bayesian analysis reported in the revised manuscript (conducted in response to the Editor’s comments) suggests that our interaction terms were more likely under the null hypothesis.

Some of your main effects are significant, because these comparisons are pooled across two of the cells, yielding increased power. However, you don’t report effect sizes here. More importantly, the lack of power resulted in you not comparing the simple effects, which I show below have some fairly large effects but you didn’t pick them up because of low power. See point 9.

Because the interactions were non-significant and the effect size was small (with confidence intervals that include zero) we do not feel there is justification to report post-hoc simple effects analysis in the paper. However, we do we discuss the simple effects in more detail below (in response to point 9), as we believe that the simple effects driving this apparent interaction were actually not predicted and so should be interpreted with caution.

1. Please report effect sizes throughout the manuscript. R and/or cohen’s d or standardized mean differences would be preferred over eta-squared.

We now report Cohen’s f with 95% confidence intervals for all of our main effects and interactions (see Tables 1 and 2).

1. The nutritional backlash items (experiment 2) seem to be scaled in the opposite direction (eg 1 is strongly agree and 5 is strongly disagree). In the past, I’ve encountered issues where subjects don’t notice the scale labels when they are switched. How can you address this possibility?

We anticipated that the change in scale direction might be missed by some participants, so we made sure to robustly address this in two ways when setting up the survey. Details of exactly how we anticipated and mitigated this issue can be found in response to the Editor (point 3).

1. The analysis plan in the preregistration does not completely outline the analyses to be conducted. It simply says “2x2 Independent groups ANOVA. If there is a significant interaction this will be decomposed (simple main effects). ” and specifies and alpha level for the interaction and “simple main effects”. So, it’s not clear what actually counts as a preregistered analysis- for example, the one-sample t-tests on the ‘scientific advancement’ measure were not preregistered. I’m not sure I agree with that analysis, anyways: I want to know whether they differ from each other, not whether they are different from 0 on their own.

We have now split the results section into ‘pre-registered’ and ‘exploratory’ analysis. The 2x2 ANOVAs are reported as pre-registered analysis. We did not report any post-hoc simple effects analysis in the paper as the interaction terms were all non-significant (our pre-registration stated that we would only run post hoc tests if the interaction was significant). However, we acknowledge that we should have pre-registered full details of the proposed post-hoc tests, rather than just broadly stating the type of analysis (we address this in detail in response to point 9). All other analysis is now labelled as exploratory (i.e., the one sample t-tests and the new Bayesian analysis suggested by the Editor).

1. There were virtually no significant effects in Experiment 2 and the graphs clearly show that the means across the four conditions are basically identical. However, I am concerned that these measures go a step beyond what would normally be influenced by these kinds of materials and manipulations. The questions (mistrust of expertise, confidence in science, epistemic beliefs) seem to be really general and broad evaluations of science and knowledge. I have serious doubts that a one-shot survey manipulation would change people’s evalautions of knowledge. I would not conclude that the headline manipulation “had no effect” simply because it did not influence how people responded to these very broad measures. I wouldn’t expect much (anything) to influence these measures, in fact. I wouldn’t really expect a change in “nutritional backlash” or “nutritional confusion”, either, though on the surface they seem more reasonable than the other 3 measures. The issue here is that you are asking participants to generalise a few nutritional headlines from an online survey into a global evaluation of “Dietary recommendations are not useful” or “Nutrition recommendations are confusing.” Well, some are and some aren’t; a single online survey is not going to influence anyone’s evaluation of this. So, in sum, I wouldn’t really draw any conclusions about your overall research question from Experiment 2. Include it in the report, please, but I would focus more on refining Experiment 1.

Prior to running the study, we had good reason to predict that our manipulations would affect the measures used. The design was inspired by published studies showing that brief online experimental manipulations of conflict can affect ratings of Nutritional Backlash and Confusion. For example, Clark, Nagler & Niederdeppe (2019) asked participants in an online study to simply read two short news articles about nutrition that were either conflicting or non-conflicting, before completing the same nutritional confusion and nutritional backlash scales (they used an 11-item version of backlash scale, we used a 6-item version). Clark et al. found that their experimental manipulation led to significantly increased nutritional confusion in the Conflict condition (relative to Non-Conflicting condition). There was also numerically increased nutritional backlash (although this comparison was not significant).

We also believe there was justification in including the other measures relating to beliefs about scientific development and trust in experts. To be clear, we did not expect brief exposure to conflicting or non-conflicting headlines to have any long-lasting or profound effect on beliefs. Instead, we expected that having examples of conflicting headlines in mind might be enough to temporarily affect responses at that time. For example, we felt it plausible that recent exposure to conflicting headlines would affect responses to items such as “*I find nutrition recommendations to be confusing*” at least temporarily. In this case, a participant who has just memorised a series of conflicting headlines may be more inclined to agree with the statement.

In sum, we believe the manipulation we used and the measures we selected were not unreasonable prior to running the study (although the results suggest different). We have strengthened our justification for these measures based on our prior beliefs but re-worded our conclusions (based on the data) to acknowledge that the measures may be a step too far too far removed from our manipulation. We hope this addresses your concerns while also satisfying the Editor’s request for us to place *more* focus on the Experiment 2 results (Editor’s letter point 4).

Reference

Clark, D., Nagler, R. H., & Niederdeppe, J. (2019). Confusion and nutritional backlash from news media exposure to contradictory information about carbohydrates and dietary fats. *Public Health Nutrition, 22*(18), 3336-3348.

1. There are some problems with your analysis code which make it difficult to reproduce your analyses exactly in ‘Study 7’. Essentially, your subset arguments return a data frame with 0 observations. Not sure what the deal is there, but I spent most of my time on Experiment 1 (Study 8) anyways.

You may have guessed that we are relative newcomers to R! However, we have checked the code carefully and can only conclude that the issue is due to differences in hardware or software. On the day of writing this (15/01/2020) the script published on OSF ran without any problem using R Studio (version 1.1.383) / R (version 4.0.3) on a Windows 10 PC. Using this set-up, the subset arguments returned a data frame (mydata) with 400 rows, as we expected.

As a workaround, we have exported this data frame as a .csv file and uploaded it to the OSF so that the subsequent analysis can be reproduced from that stage onwards. We have also uploaded the complete analysis output as a PDF document.

In general, we can see reproducibility issues relating to R scripts becoming more and more common (for example, as packages are updated over time scripts may stop working or produce different output) so we now use the Checkpoint packaged in the revised scripts. This will ensure that the same package versions that we used are also used by those trying to reproduce the analysis.

Based on Figure 1, you can see that there is an interaction such that the non-conflict conditions differ (visually) from the conflicting headlines between the generic and quantified cells. If we select those cells and compare them, we find fairly large effect sizes. For non-conflicting headlines, generic headlines versus quantified resulted in: reduced contradiction (d = .40!) Greater confusion (d = .26!) Less advancement (d = .27!)

For the conflicting headlines, generic headlines versus quantified resulted in: Similar levels of contradiction (d = .03) More confusion (d = .16) Less advancement (d = .19) I’ll append the code for this at the very end of this review, just so it doesn’t get cluttered.

These results are mostly in line with your hypotheses and a common sense explanations, but you would not detect them because you are very low on power – remember you were powered for d = .70.

Below we explain why we believe these effects are actually not in line with our predictions. To reiterate our point relating to power, we had power to detect effects of d=0.7 99% of the time and d=0.46 80% of time.

Most social psychological effects are much, much smaller, within the range of the effects I just reported (see: Funder & Ozer, 2019; Gignac & Szodorai, 2016). Your average effect is going to be around a d =.20,which requires more people. It might seem like a lot of people, but that’s just because most past research was terrible underpowered (and likely hacked) and underpowered studies inflates effect sizes. In general, we should have much larger sample sizes.

So, to detect a d=.20 80% of the time, I’d use a one-tailed test. You’d need 309 people in each condition. To detect the smaller effect of d = .15, I’d again use a one-tailed t-test. You’d need 550 in each condition.

library(pwr)

pwr.t.test(type= "two.sample", d = .15, sig.level = .05, power = .80, alternative = "greater")

So, I hope that the authors will address the above concerns. I’d like to see this manuscript published because I think it is interesting and it is very intuitive. But I also don’t want some important effects to be missed because the study was underpowered. I’d say the effects from experiment 1 should be replicated with a larger sample size, but I guess that decision is ultimately up to the editor.

As we outlined above, we do not feel the marginal interaction effects give us justification to report the post hoc simple effects analysis. This is partly because the interaction terms were non-significant (see also the Bayesian analysis) but also because we believe that the simple effects that you describe above are actually inconsistent with our predictions.

The best way to begin addressing your comment is to clarify the type of interaction we predicted, because we did not make this clear in the pre-registration (which is completely our fault). In the pre-registration we stated that we would follow up a significant interaction by looking at the simple main effects but did not specify exactly how, which is an important omission. We were also vague in the ‘hypothesis’ section of the pre-registration when it came to describing the type of interaction we expected. This was an error on our part due to lack of experience with pre-registration (but hopefully better than no pre-registration). In the initial submission we spelled out more clearly how we predicted that the interaction effect would turn out.

*“…we predicted an interaction effect, whereby the size of these conflict effects would be smaller or non-existent when headlines were qualified with ‘some’”*

In Figure 2 below we’ve illustrated the type of effect we expected below, with the ‘conflict effect’ (i.e., the difference between those who saw conflicting headlines and those who saw non-conflicting headlines) being relatively smaller when headlines were qualified.

Figure 2: Illustration of what was meant by *“…we predicted an interaction effect, whereby the size of these conflict effects would be smaller or non-existent when headlines were qualified with ‘some’”*



Qualified

Another (better) way of framing this predicted effect would be to say we expected no effect of our headline Format manipulation when the headlines were non-conflicting. This is because there was no conflict in this condition therefore there was no reason to predict that headline format would have any effect on perceptions of conflict or confusion - this was simply a control condition. In contrast we did expect an effect of headline Format when the headlines were conflicting (with qualified headlines causing less confusion and perceived contradiction). We’ve illustrated this in Figure 3 below.

Figure 3: Alternative way of framing our predictions



Qualified

On initial inspection of the means, there appeared to be some evidence for our prediction on the Contradiction and Confusion measures. For the contradiction measure the conflict vs no conflict effect size reduced from d=3.66 to d=2.74 (see Figure 4) and on the confusion measure the effect size reduced from d=1.24 to d=0.76 (see Figure 5).

Figure 4: Experiment 1: Measure = Contradiction (note: lowest possible contradiction score = 6)



Qualified

Figure 5: Experiment 1: Measure = Confusion



Qualified

However, when we look at Figure 4 and 5 in more detail, the interactions do not pattern as we expected. We found little evidence for an effect of Format when the headlines were conflicting (blue line). The effect sizes were d=0.03 (CI95 -0.29, 0.36) for the contradiction (Figure 4) measure and d=0.16 (CI95 -0.17, 0.48) for the conflict measure (Figure 5). The lower confidence interval for both of these effects crosses zero.

We expected no effect of Format when participants had seen non-conflicting headlines (orange line), but unexpectedly the effect sizes were relatively large in this condition. Adding a qualifier to non-conflicting headlines made them appear more contradictory (d=0.4, CI95 .07, .72) and created more confusion (d=0.26, CI95 -0.06, 0.59). It is these unexpected effects relating to the non-conflicting headlines that are driving the marginally non-significant interaction terms in Experiment 1.

In sum, we are very cautious about interpreting the marginally significant interaction effects as they are driven by simple effects that are not consistent with our hypothesis. The simple effects we did predict were very small, with effect size confidence intervals that include zero. We have opted not to re-run Experiment 1 because the results based on 300 participants give us little evidence for the interaction effect we predicted.

References: Funder, D. C., & Ozer, D. J. (2019). Evaluating Effect Size in Psychological Research: Sense and Nonsense. Advances in Methods and Practices in Psychological Science, 2(2), 156–168. <https://doi.org/10.1177/2515245919847202> Gignac, G. E., & Szodorai, E. T. (2016a). Effect size guidelines for individual differences researchers. Personality and Individual Differences, 102(November 2016), 74–78. <https://doi.org/10.1016/j.paid.2016.06.069>

Code: I used your script for Study 8, and load the resulting mydata.csv file here:

NC <- dplyr::filter(mydata, Conflict == "Non-Conf.") # NC = no conflict

CF <- dplyr::filter(mydata, Conflict == "Conf.") #CF = conflict

NC <- dplyr::select(NC, Format, contradiction, confusion, advancement)

CF <- dplyr::select(CF, Format, contradiction, confusion, advancement)

psych::cohen.d(NC ~ Format)

psych::cohen.d(CF ~ Format)

**Reviewer 2**

The authors present an interesting study of the potential impact of conflict (whether participants read two headlines about nutrition research findings that conflicted with each other) and generic language (“experts say” vs. “some experts say”). They found that participants in Experiment 1 did notice the contradiction between conflicting headlines and found those headlines more confusing and less reflective of advancing science. However, conflict did not influence global beliefs about health. Generic language was not significant. I learned a lot just from the questionnaires you included!

Thank you – we are very pleased you found it interesting!

I have a few questions/comments that might benefit from minor revision:

1. Pg. 8: “These specific headlines do not directly contradict each other.” This is true for the ones you list, but less true for others in this condition. For instance, “Everyone needs to take vitamin D supplements …” and “taking vitamin D is pointless” seem to be in direct opposition to one another (I’d say the same for your meat category). I’m not sure how important this is for highlighting conflict. Did participants differ at the item level? For instance, were the vitamin D ratings systematically higher in conflict/confusion or lower in scientific advancement than the alcohol ratings?

To be consistent with the type of conflict that people experience in real life we focused on what Carpenter et al. (2016) describe as ‘decisional conflict’. Sometimes the headline pairs were direct contradictions (such as the examples 1 and 2 in the paper which relate to fasting diets) and sometimes they were not direct contradictions (such as examples 5 and 6 in the paper which relate to alcohol). Importantly, regardless of whether the headlines were exact opposites or not, all headline pairs were conflicting in the ‘decisional’ sense. For example, the conflicting pair of headlines related to fasting implied opposite courses of action (headline 1 in the paper implies that fasting is a good decision but headline 2 implies it is a bad decision) as did the pair of conflicting headlines relating to alcohol (headline 5 in the paper implies that drinking a moderate amount of alcohol is a good decision but headline 6 implies it is a bad decision).

To help address your comment, we conducted item level analysis on the contradiction ratings collected in Experiment 1 where participants were directly asked the extent to which they agreed with statements such as:

* “*The two headlines about* ***dietary fat*** *contradicted one another*”
* “*The two headlines about* ***Vitamin D*** *contradicted one another*”
* *Etc.*

Responses were measured on a Likert scale where 1=Strongly Disagree, 3= neither agree nor disagree, 5 = Strongly Agree.

For each of the six headline topics we calculated the mean contradiction rating in the conflict condition and in the no-conflict condition. We also calculated the effect size of this difference separately for each headline topic. These figures are reported in the Table below

|  |
| --- |
| **Perceived Contradiction (Exp. 1)**“*The two headlines about* ***[…….]*** *contradicted one another*”**1**=Strongly Disagree, **3**= Neither agree nor disagree, **5** = Strongly Agree. |
|  | **NON-CONFLICTING PAIR OF HEADLINES** | **CONFLICTING PAIR OF HEADLINES**  | **Cohen’s d**  |
| Vitamin D   | 1.87  | 4.33  | 2.36  |
| Meat   | 2.11  | 4.46  | 2.28  |
| Coffee   | 2.73  | 4.14  | 1.04  |
| Fat   | 3.03  | 3.82  | 0.70  |
| Alcohol   | 1.96  | 4.20  | 1.96  |
| Fasting   | 1.75  | 4.31  | 2.52  |

For every topic, the conflicting pair of headlines was seen as more contradictory than the non-conflicting pair. The effect size was ‘large’ for every item (if d=0.8 is considered a large effect) and some effect sizes were larger than others. Consistent with your observation, the pairs of headlines that were directly conflicting (e.g., Vitamin D, Meat, Fasting) produced the largest effects (d>2). The pairs of headlines that where the conflict was less direct (Coffee, Fat, Alcohol) still had large effect sizes but not quite as large as those above (ranging from d=0.7 to d=1.96).

The purpose of showing participants six pairs of conflicting or non-conflicting headlines was to prime a general sense of conflict or consistency and to find out whether this affected ratings of how confusing the headlines were (as a whole set) and how much they advanced our knowledge (as a whole set). The questions we asked are below:

CONFUSION: “*The headlines I was asked to remember create confusion about how to be healthy*”

ADVANCEMENT: “*When we take the results reported in these headlines together, do we now know more, less or the same as we did before about how to be healthy?*”

The fact that the conflict effect was large for all six topics suggests that our manipulation worked. It was inevitable that some topics would have stronger effects than others, but we don’t see that causing any issues as all that was important was building up a general sense of conflict during the memory encoding task.

1. Do you have any demographic information (besides age/gender) to report about your participants?

We only requested age and gender from participants as we try to minimise the collection of personal identifying information. In addition to the demographic data we collected directly, we also know that all participants spoke English as their first language and had not taken part in related experiments, as these were pre-screening requirements (the study was only advertised to those aged 18+, who spoke English as their first language and had not taken part in related experiments).

We do not have further details of the specific participants in our study, but below we provide demographic details about the wider pool of participants from which our sample was drawn (source <https://www.prolific.co/demographics/> ). We have revised our Participants section to better describe the population. A summary of the population demographics is provided below for your convenience.

• Most participants in the pool were born in the UK or USA (33% and 28% respectively).

• 45% are male and 55% female

• Most (70%) report their ethnicity as white/Caucasian

• Most of the participants are aged between 20-30 years (50%) followed by 26% aged 30-40 years

• 61% are in either full-time or part-time employment

• 39% are university students

• 33% have an undergraduate degree, 29% have A Levels or high school qualifications, and 17% a postgraduate degree

1. Methods: It would be helpful to list the descriptors (if any) for the midpoint of each scale you sued (similar to “confidence in the scientific community”), or say explicitly that only the end points of the scale were marked. This would be especially helpful for Exp. 2, where a lot of the ratings look to be hovering around the midpoint.

Further information about scale descriptors has now been added to the Method section. This midpoint of the 5-point Likert scales used was always labelled ‘Neither agree nor disagree’.

1. Would ordinal regressions be more appropriate tests for your data than ANOVAs?

There are lots of conflicting opinions on this one, so we will try and cover all bases to show our analysis was robust. We answer this in two parts, as there is a different answer depending on whether a scale was based on multiple items or just a single item.

*Scales based on multiple items:* Most of our measures involved summing or averaging over a number of items. Each item was rated on an ordinal Likert scale, then we calculated the sum or the average of these responses. For multiple item scales like these, it is justifiable to treat scores on the overall scale (i.e., the collection of items) as interval level data, rather than ordinal (Carifio & Perla, 2008). It is analogous to summing the number of correct answers on a multiple-choice exam (i.e., each response scale is ordinal, but the sum of correct answers can be treated as interval level data) (Norman, 2010). For this reason, we chose to use ANOVA over ordinal regression for these multi-item scales.

*Scales based on a single item:* Three of our measures were based on single item Likert scales. These measures were ‘Confusion’ and ‘Scientific Advancement’ in Experiment 1 and ‘Confidence in the Scientific Community’ in Experiment 2. The data from these measures are clearly ordinal and ordinal regression is therefore an appropriate method. It has also been argued that ANOVA is an equally suitable measure, as in practice it is robust when ordinal data are used (i.e., produces unbiased results) with the advantage of having greater power (Norman, 2010). To cover all bases and demonstrate the robustness of our ANOVAs we also conducted ordinal regressions (see revised analysis scripts on OSF) which produced exactly the same pattern of results. A comparison of *p* values produced by these two methods is summarised in the Table below. This shows that regardless of analysis method, we would draw exactly the same conclusions. We opted to report ANOVA in the paper as the method has greater power.

|  |  |  |  |
| --- | --- | --- | --- |
|  | ANOVA *p* values  | Ordinal Regression *p* values  | Same conclusion? |
| **Exp 1 Confusion**  |  |  |  |
| Conflict  | <.001\* | <.001\* | Yes |
| Format | .402 | .325 | Yes |
| Conflict \* Format  | .066 | .109 | Yes |
|  |  |  |  |
| **Exp 1 Advancement**  |  |  |  |
| Conflict  | .001\* | .001\* | Yes |
| Format | .053 | .062 | Yes |
| Conflict \* Format  | .707 | .652 | Yes |
|  |  |  |  |
| **Exp 2 Confidence** |  |  |  |
| Conflict  | .516 | .617 | Yes |
| Format | .390 | .343 | Yes |
| Conflict \* Format  | .865 | .992 | Yes |

References

Carifio, J., & Perla, R. (2008). Resolving the 50‐year debate around using and misusing Likert scales. *Medical education*, *42*(12), 1150-1152.

Norman, G. (2010). Likert scales, levels of measurement and the “laws” of statistics. *Advances in health sciences education*, *15*(5), 625-632.

Other points to consider for the discussion or future studies:

Study 1: really interesting: “The mean of those exposed to non-conflicting headlines (-0.007) did not significantly differ from zero (t(146) = 0.12, p= 0.902).” From the histogram it’s clear that “same as” was the modal response (not that some people thought we knew more and some thought we knew less).

We’ve made this point clearer in the manuscript. Regardless of condition, the modal response was ‘we know the same’.

I wonder what it would take for people to say that we know more than before. I’d be curious about the extent to which participants view themselves as knowledgeable about nutrition already (compared to other fields of scientific inquiry). I’m not sure how much any ratings in Exp. 2 speak to this, but at least in terms of “nutritional confusion,” participants are typically in the middle (rather than reporting high confusion, even when exposed to conflicting findings).

This is an interesting point. The modal response for those who saw consistent (non-conflicting headlines) was to choose ‘we know the same’. We expected that seeing details of research studies that reach consistent conclusions would result in participants selecting ‘we know more’. Indeed, the only other paper to have used this Scientific Advancement scale reported that the modal response, in the absence of any conflict, was to select ‘we know more’ (Koehler & Pennycook, 2019). There are a couple of differences between our Experiment 1 and the work of Koehler and Pennycook, which might explain the findings. First, Koehler & Pennycook presented participants with factual summaries of research projects (these were several sentences in length with details about the design and specific outcomes). Second, the research summaries used by Koehler and Pennycook were not about nutrition (instead the three research topics they used were on 1. the effects of a specific method of teaching mathematics 2. an exercise programme for arthritis 3. nonverbal aggression in videogame players). In short, previous research suggests that participants will select ‘we know more’ when they can see the details and the topics are relatively uncontroversial.

We suspect that both of the differences described above are relevant to how participants responded in our study. Instead of using detailed summaries carefully written in neutral language, the materials in our study were genuine news headlines. These headlines inherently lack detail and are written as much to attract attention as they are to communicate facts. The use of headlines may therefore cause participants to be sceptical of the claims, even when they are consistent. It is rational to withhold judgement on scientific progress in the absence of any significant detail, in which case ‘we know the same’ would be the most suitable response.

The subject of the materials may also have played a part. Koehler and Pennycook focused on topics such as Education and Psychology rather than nutrition. The research studies they described were relatively uncontroversial and on topics where participants are likely to have few prior beliefs. In contrast, there is a substantial body of research showing that most of us are regularly exposed to (seemingly) conflicting nutritional information in the media. This prior experience may well increase the weight of evidence required to perceive a scientific advance in this area. Even if two studies report consistent findings, it may feel inevitable that conflicting research will ultimately cancel it out. This type of apathy/desensitisation may again cause participants to withhold judgement on scientific progress.

Finally, it could simply be a matter of trust in the media. If readers place little trust in tabloid headlines (because they are known to sensationalist / non-neutral) they may again abstain from making any strong judgement on scientific progress. Our participants in Experiment 2 had low ‘mistrust of expertise’ (implying high trust in experts), but they have had mistrust in how the media reports the work of experts.

I would be curious how introducing a quantifier into who showed the finding (e.g., “in some people”) would contrast with your manipulation of qualifying who is making the claim (e.g., “some scientists”). Generics also gloss over variability within participants. This strikes me as especially important in nutrition research (e.g., the fasting diet could have those two effects listed on pg. 4, but for different individuals or groups of people).

We completely agree that this is an important question for future research. A PhD student in our lab is planning to do work along these lines, following on from work by DeJesus et al. (2019). He’ll be looking at the interpretation of generic health claims made about groups such as ‘Men’, ‘Women’ and ‘Children’ (where it would be more accurate to say, ‘Some men’, ‘Some women’ or ‘Some children’).

Reference

DeJesus, J. M., Callanan, M. A., Solis, G., & Gelman, S. A. (2019). Generic language in scientific communication. *Proceedings of the National Academy of Sciences, 116*(37), 18370-18377.

Unrelated to your study, I’m curious how well the 59 participants in Exp. 1 (who did the study twice) matched in their responses from the first attempt.

We looked at the data from these 59 participants as we thought it might provide a useful insight into how consistent responses were across the two attempts. However, we soon realised that comparisons between the two attempts are unlikely to be meaningful. Of the 59 participants who started the study a second time only 38 completed the entire study again and in most cases these participants were randomly assigned to a different condition on their second attempt. For this reason, we cannot make any meaningful comparison of responses. Participants saw different stimuli on each occasion so we would expect different responses. In addition, seeing the same headlines worded differently on the second occasion will have made the experimental manipulation obvious, which might have affected responses. We also need to consider the motivations of the participants on their second attempt – they were likely frustrated at having to complete the study twice (even though this belief was mistaken) so more likely to click through at speed simply to ensure they were paid. As we might have expected, those who completed the study twice took less time on their second attempt (possibly because they did not need to read the instructions a second time and possibly because they rushed through).

**Editor Final Decision—Accept**

Feb 16, 2021

Dear Dr. Haigh,

I have now had a chance to read over your manuscript “When ‘Scientists say’ coffee is good for you one day and bad for you the next: Do generic attributions to ‘Scientists’ and ‘Experts’ amplify perceived conflict?”, along with the letter describing the changes you made. Thank you for your responsiveness to the concerns that the reviewers and I raised - your letter was especially thorough and helpful to me in evaluating your revision. I am happy to say that your paper is now officially accepted for publication in Collabra: Psychology. Congratulations on this excellent work, I think it will make an important contribution to the literature and I look forward to seeing it published! I hope your experiences with Collabra: Psychology have been positive and that you will continue to consider it as an outlet for your work.

As there are no further reviewer revisions to make, you do not have to complete any tasks at this point. Our managing editor will contact you in case there are any pre-prodution file related questions. You will have an opportunity to check the page proofs before we publish your article. Thank you again for publishing in Collabra: Psychology.

Sincerely, Simine Vazire Editor in Chief Collabra: Psychology