**Peer review comments**

Article title: Does ‘Scientists believe…’ imply ‘All scientists believe...’? Individual differences in the interpretation of generic news headlines

Author names: Matthew Haigh, Hope A. Birch, Thomas V. Pollet

Handling editor: Kevin King

**Editor decision Revise & Resubmit**

June 22, 2020

Dear Dr. Haigh,

I have now received all reviews of your manuscript, “Does ‘Scientists believe…’ imply ‘All scientists believe…’? Individual differences in the interpretation of generic news headlines” from qualified researchers. Thank you for your patience in awaiting this decision. 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 I therefore encourage you to submit a revised version for further consideration at Collabra: Psychology.

I was very fortunately to recieve outstanding reviews from two experts in the field. I believe their comments are quite clear, and I will not re-capitulate their thoughts here. I believe all of their comments should be quite easy addressable in a revision. 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.

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,

Kevin King

**Reviewer 1:**

Your name will not be revealed unless you wish to sign your review. If you chose to show your name, please type it in the text field below. If you prefer your comments to be anonymous, leave the field blank.

This research starts with an interesting question about perceived expert consensus and tests a specific hypothesis about the interplay of intuition and deliberation in the construction of consensus estimates. The broad conclusion is that people, on average, interpret a statement such as “scientists believe…” as indicating that about 60% of those scientists hold the belief in question. But there is also substantial individual variability in the consensus estimates, which the authors frame as consistent with the logic of such generic statements: It is logically true to say “scientists believe…” whether the actual proportion of scientists who believe is 2% or 98%. Gricean principles of conversational pragmatics, of course, might suggest otherwise. If only a tiny minority of scientists believe something, it seems deceptive to describe that fact with an unqualified statement that “scientists believe…”. There are two surprising results here. First, because the authors make a good case that cognitive reflection should play some role in adjusting estimates downward from an initial intuitive interpretation of “scientists believe…” as applying to 100% of scientists, it comes as something of a surprise that they find no evidence for this account despite using a couple different methodologies for testing it. Second and, to me, more surprising, is that the mean estimated level of consensus is only around 60%. It seems highly plausible that people would, at least initially, interpret “scientists believe…” as referring to all scientists unless that statement was qualified in some way. Partly, as mentioned above, this seems plausible from the perspective of conversational norms. Additionally, it seems plausible that members of the public would expect scientists – or experts in any domain for that matter – to typically exhibit a high level of agreement on matters within their areas of expertise. (This is separate from the question of whether that is in fact true of scientists generally.) A study worth considering in this regard is by Aklin and Urpelainen (2014), who argue that the public generally assume such a high level of consensus among scientists that drawing attention to even a small dissenting minority can raise public doubts.
This makes me wonder if something about the methodology of the current study led to lower consensus estimates than would otherwise be the case. For instance, the authors identify as one limit of their work the decontextualized phrases used as targets of judgment. More generally, if the statement was embedded in a newspaper-like article summarizing a scientific conclusion, and participants were later asked what proportion of scientists believed the conclusion that was attributed to them, it is possible the estimated consensus level would be higher. More importantly, in my view, is the use of a within-subjects design in which participants were exposed to a set of statements and asked to estimate the level of consensus associated with each. Long-understood principles of scale use (e.g., Parducci’s 1965 range-frequency theory) imply that participants will use the response scale in a way that distributes items across its full range. So, if some of the “filler” statements (e.g., “many scientists believe…”) encountered by participants would seem to call (in the comparative context established by the other items) for a higher estimate than the key set of target statements, and if other fillers (e.g., “few scientists believe…”) seem to call for lower estimates, then this would force the ratings of the target statements toward the middle of the scale. On this account, the observed mean consensus estimate for the target statements becomes difficult to interpret in any absolute way as its value may be, to a large extent, an artifact of the context in which the target statements were evaluated.

Aklin, M., & Urpelainen, J. (2014). Perceptions of scientific dissent undermine public support for environmental policy. Environmental Science & Policy, 38, 173-177.

Parducci, A. (1965). Category judgment: a range-frequency model. Psychological review, 72(6), 407.

**Reviewer 2:**

The authors present an interesting study of readers’ interpretation of generic headlines (such as “scientists say…”). I have previously reviewed this paper and I still find the paper to be clearly written and on a topic of growing interest in psychology and other fields. The figure design and the authors’ commitment to open science practices is commendable.

I also have a few of the same concerns. Namely, Figure 1 suggests an additional conclusion that participants are treating “some” and “few” statements differently than “many” and the target statements (e.g., “Experts think”), both in terms of the means and the distribution of the data. This suggests that there may be something missing by treating all these statements as one latent “scientific consensus” variable (although it is possible that there is too much individual-level variability for this to be a viable hypothesis). The authors do a nice job unpacking these high levels of variability across participants (as opposed to across statements) in the discussions; I found the point about “some estimating that, on average, the nine statements referred to as few as 2% of relevant scientists and others estimating that they refer to 100% of relevant scientists” (pg. 13) especially compelling. But I think there is room to examine this difference in item type more explicitly. I appreciate that this may not be part of the authors’ pre-registered analysis, but since they divide statements into “target” and “filler,” that suggest some a priori thinking about potentially different interpretations of these phrases. This may or may not be related to cognitive reflection, but would be interesting to understand nonetheless.

The authors might be interested in this recent paper: Van Der Bles, A. M., van der Linden, S., Freeman, A. L., & Spiegelhalter, D. J. (2020). The effects of communicating uncertainty on public trust in facts and numbers. Proceedings of the National Academy of Sciences, 117(14), 7672-7683.

I wish the authors the best in their future work!

**Author response**July 6, 2020

Dear Dr King

Thank you for the opportunity to revise and resubmit our manuscript (*Does ‘Scientists believe…’ imply ‘All scientists believe...’? Individual differences in the interpretation of generic news headlines*). We found the reviewer comments very helpful. We have addressed their comments point-by-point and have made corresponding revisions where 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 & Thomas Pollet

**The reviewer comments are reproduced in full below. Our replies are highlighted in grey.**

**Reviewer 1**

This research starts with an interesting question about perceived expert consensus and tests a specific hypothesis about the interplay of intuition and deliberation in the construction of consensus estimates. The broad conclusion is that people, on average, interpret a statement such as “scientists believe…” as indicating that about 60% of those scientists hold the belief in question. But there is also substantial individual variability in the consensus estimates, which the authors frame as consistent with the logic of such generic statements: It is logically true to say “scientists believe…” whether the actual proportion of scientists who believe is 2% or 98%. Gricean principles of conversational pragmatics, of course, might suggest otherwise. If only a tiny minority of scientists believe something, it seems deceptive to describe that fact with an unqualified statement that “scientists believe…”.

This is a good point, thank you for raising it. We now acknowledge that there are likely to be some pragmatic limits to the use of generic phrases (see page 3 footnote). Although it would be true to say ‘scientists believe...’ if only a tiny minority believe something, it would violate Grice’s Maxim of Quantity (i.e., give the most helpful amount of information). A more helpful amount of information would be to say ‘A few scientists believe…’. Despite being factually accurate, from a pragmatic perspective the use of a generic phrase in this situation would certainly be ‘economical with the truth’.

There are two surprising results here. First, because the authors make a good case that cognitive reflection should play some role in adjusting estimates downward from an initial intuitive interpretation of “scientists believe…” as applying to 100% of scientists, it comes as something of a surprise that they find no evidence for this account despite using a couple different methodologies for testing it. Second and, to me, more surprising, is that the mean estimated level of consensus is only around 60%. It seems highly plausible that people would, at least initially, interpret “scientists believe…” as referring to all scientists unless that statement was qualified in some way. Partly, as mentioned above, this seems plausible from the perspective of conversational norms. Additionally, it seems plausible that members of the public would expect scientists – or experts in any domain for that matter – to typically exhibit a high level of agreement on matters within their areas of expertise. (This is separate from the question of whether that is in fact true of scientists generally.) A study worth considering in this regard is by Aklin and Urpelainen (2014), who argue that the public generally assume such a high level of consensus among scientists that drawing attention to even a small dissenting minority can raise public doubts.

Thank you for bringing this paper to our attention. It is very relevant, so we have now cited it in the introduction. Aklin and Urpelainen’s conclusion that “..*.the general public's default**assumption is a very high degree of scientific consensus*” (in the absence of any other dissenting information) strengthens our argument that participants would be expected to initially infer a high degree of scientific consensus (e.g., *Scientists believe = All scientists believe*).

This makes me wonder if something about the methodology of the current study led to lower consensus estimates than would otherwise be the case. For instance, the authors identify as one limit of their work the decontextualized phrases used as targets of judgment. More generally, if the statement was embedded in a newspaper-like article summarizing a scientific conclusion, and participants were later asked what proportion of scientists believed the conclusion that was attributed to them, it is possible the estimated consensus level would be higher.

We agree that ratings could be different if the generic phrases were fully contextualised and we acknowledge this in the manuscript (Abstract: “*These ratings of decontextualized phrases will inevitably be labile and prone to change with the addition of context, but under controlled conditions people interpret generic consensus statements in very different ways*”). When designing these experiments, we spent a lot of time thinking about the trade-off between experimental control and ecological validity. Presenting the generic phrases in isolation ensured the maximum degree of experimental control by eliminating the effects of prior beliefs relating to the topic of the headline (with the limitation that the task is much more abstract than evaluating contextualised scientific conclusions). For example ‘*Scientists believe climate change is due to human activity*’ and ‘*Scientists believe the earth is flat*’ would lead to very different estimates of how many ‘scientists believe’, based on prior beliefs about climate change and the shape of the earth. This is obviously an extreme example, but hopefully it illustrates our rationale. For this reason, we decided to begin our programme of research using isolated abstract phrases. The next stage of our research programme is to investigate contextualised scientific conclusions. You may be interested to know that we are currently working on studies (which test a different hypothesis to the one in this paper) where we ask participants to evaluate generic headlines in their original context (e.g., [*Scientists think* *Pluto once had rivers and lakes of liquid nitrogen*](https://www.theverge.com/2016/3/22/11282648/pluto-lake-river-liquid-nitrogen-new-horizons)). In these experiments we have balanced experimental control and ecological validity using a large sample of genuine headlines on topics that participants are expected to have few prior beliefs about.

More importantly, in my view, is the use of a within-subjects design in which participants were exposed to a set of statements and asked to estimate the level of consensus associated with each. Long-understood principles of scale use (e.g., Parducci’s 1965 range-frequency theory) imply that participants will use the response scale in a way that distributes items across its full range. So, if some of the “filler” statements (e.g., “many scientists believe…”) encountered by participants would seem to call (in the comparative context established by the other items) for a higher estimate than the key set of target statements, and if other fillers (e.g., “few scientists believe…”) seem to call for lower estimates, then this would force the ratings of the target statements toward the middle of the scale. On this account, the observed mean consensus estimate for the target statements becomes difficult to interpret in any absolute way as its value may be, to a large extent, an artifact of the context in which the target statements were evaluated.

This is a good point that we have explored with some additional analyses. Before addressing your comment, we will outline our rationale for including filler items. We chose to include filler items for three reasons:

1. To avoid too much repetition for participants and keep them engaged (so they did not get the impression they were being asked to rate exactly same type of statement again and again)
2. To mask our specific interest in generic phrases by mixing them up with other phrases
3. To ensure that participants had to think about each phrase on its merits and discourage them from giving exactly the same rating to every phrase (similar to reason 1).

In the manuscript we phrased our rationale for including fillers in the following way “*The nine target items were presented amongst nine filler items, three of which referred to ‘Some [scientists/experts/researchers] …”, three to ‘Many [scientists/experts/researchers] ...’ and three to ‘Few [scientists/experts/researchers] …’. This was to ensure that participants remained engaged and used the full range of the scale*.”

It is a good point that the ‘many scientists...' and ‘few scientists...’ fillers (which quite obviously call for high and low ratings respectively) might force ratings of the more ambiguous ‘scientists believe…’ towards the middle of the scale. We were able to test this with some additional analysis of our Study 1 data (We have shared this on OSF <https://osf.io/n7pj8/> Supplementary analysis folder). The 18 phrases were presented in different random order to each participant (each phrase presented on a separate page). We identified all those participants whose **first** trial (i.e., the first item they rated) was one of our generic target phrases (n=169). The first trial for these participants could not have been affected by context, as they had not seen any of the fillers at this stage. If the fillers did force the ratings of the target statements toward the middle of the scale, then we would expect the mean of these first trial ratings (which are unaffected by context) to be higher than the overall mean across *all* generic trials (which could have been affected by context). The mean rating on these specific trials was 52.7, which is actually slightly lower than the overall mean across all generic trials (53.8). We therefore feel confident that ratings of generic phrases were not constrained (forced towards the middle of the scale) by ratings given to filler phrases.

Aklin, M., & Urpelainen, J. (2014). Perceptions of scientific dissent undermine public support for environmental policy. Environmental Science & Policy, 38, 173-177.

Parducci, A. (1965). Category judgment: a range-frequency model. Psychological review, 72(6), 407.

**Reviewer 2**

The authors present an interesting study of readers’ interpretation of generic headlines (such as “scientists say…”). I have previously reviewed this paper and I still find the paper to be clearly written and on a topic of growing interest in psychology and other fields. The figure design and the authors’ commitment to open science practices is commendable.

I also have a few of the same concerns.

Thank you for these kind words – we hope this version further addresses your concerns.

Namely, Figure 1 suggests an additional conclusion that participants are treating “some” and “few” statements differently than “many” and the target statements (e.g., “Experts think”), both in terms of the means and the distribution of the data. This suggests that there may be something missing by treating all these statements as one latent “scientific consensus” variable (although it is possible that there is too much individual-level variability for this to be a viable hypothesis).

We believe this comment may stem from us not being clear enough that the latent ‘scientific consensus’ variable was measured using only the nine generic phrases (all of which have very similar means and distributions). Ratings to the ‘some’, ‘many’ and ‘few’ statements were not used to create the latent variable. These were simply filler items to make the task less repetitive for participants (we further explain the purpose of our filler items above, in response to Reviewer 1). We have now clarified this in the paper.

The authors do a nice job unpacking these high levels of variability across participants (as opposed to across statements) in the discussions; I found the point about “some estimating that, on average, the nine statements referred to as few as 2% of relevant scientists and others estimating that they refer to 100% of relevant scientists” (pg. 13) especially compelling. But I think there is room to examine this difference in item type more explicitly. I appreciate that this may not be part of the authors’ pre-registered analysis, but since they divide statements into “target” and “filler,” that suggest some a priori thinking about potentially different interpretations of these phrases. This may or may not be related to cognitive reflection, but would be interesting to understand nonetheless.

This is something we have now further investigated with some new exploratory analyses. The fillers were included simply to make the task less repetitive for participants, so we did not have any apriori hypotheses about their interpretation or their relationship with cognitive reflection (informally we assumed there would be no relationship). As a result, we did not pre-register any confirmatory analysis relating to the filler items. Following on from your comment, we decided to run and report some exploratory analyses on the Study 1 data. Bivariate correlations between each of our nine fillers and our three measures of cognitive reflection revealed an interesting trend whereby interpretations of the three ‘Few’ statements had weak negative relationships (*p*<0.05) with our three measures of cognitive reflection (see Table 3 in revised manuscript). There was also some evidence of a similar pattern with our ‘some’ statements. We have added an Exploratory Analysis section to the Study 1 Results section where we now report these findings but emphasise in the Discussion that the analysis was exploratory, the relationships were weak (the strongest correlation was -0.21) and the relationships were not hypothesised. These findings do open the door for future confirmatory studies to examine whether numerical estimates of verbal quantity phrases (such as ‘few’ and ‘some’) are related to cognitive reflection. A negative relationship like those we found would be expected if people initially overestimate quantifiers and then revise downwards on reflection (with more reflective people making larger downward revisions).

In addition to reporting bivariate correlations, we also plotted the means for each filler item in Figure 1 and reported the average estimates of ‘some’, ‘many’ and ‘few’ in the Study 1 Discussion. We refrained from any detailed discussion about the mean difference in consensus ratings between the filler and the target generic phrases, as these comparisons are not like with like. This is because the filler phrases used different verbs to the generic phrases (the generic phrases included the verbs ‘believe’, ‘say’ and ‘think’ whereas the filler phrases included the verbs ‘argue’, ‘suggest’ and ‘ agree’).

The authors might be interested in this recent paper: Van Der Bles, A. M., van der Linden, S., Freeman, A. L., & Spiegelhalter, D. J. (2020). The effects of communicating uncertainty on public trust in facts and numbers. Proceedings of the National Academy of Sciences, 117(14), 7672-7683.

Thank you for bringing this paper to our attention. It is very relevant to experiments we are currently working on. These experiments look at the psychological effects of taking generic headlines, which imply a high degree of certainty among experts (e.g., “*Doctors recommend Coca-Cola to treat stomach blockages*”), and making them less certain by inserting the word ‘some’ (“*Some doctors recommend Coca-Cola to treat stomach blockages*”).

I wish the authors the best in their future work!

Thank you!

Editor second decision - Accept

July 24, 2020

Dear Dr. Haigh,

I have now had a chance to read over your manuscript “Does ‘Scientists believe…’ imply ‘All scientists believe…’? Individual differences in the interpretation of generic news headlines”, along with the letter describing the changes you made. Thank you for your responsiveness to the concerns that the reviewers and I raised. 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, Kevin King