Does 'Scientists believe…' imply 'All scientists believe...'?

Individual differences in the interpretation of generic news headlines

Matthew Haigh*

Hope A. Birch

Thomas V. Pollet

*Corresponding author

Department of Psychology, Northumbria University, Newcastle upon Tyne, NE1 8ST, United Kingdom. Telephone: +44 (0)191 227 3472. E-mail: matthew.haigh@northumbria.ac.uk

Word count (exc. Abstract and References) 7858

Abstract

Media headlines reporting scientific research frequently include generic phrases such as "Scientists believe *x*" or "Experts think *y*". These phrases capture attention and succinctly communicate science to the public. However, by generically attributing beliefs to 'Scientists', 'Experts' or 'Researchers' the *degree* of scientific consensus must be inferred by the reader or listener (do all scientists believe *x*, most scientists, or just a few?). Our data revealed that decontextualized generic phrases such as "Scientists say…" imply consensus among a majority of relevant experts (53.8% in Study 1 and 60.7-61.8% in Study 2). There was little variation in the degree of consensus implied by different generic phrases, but wide variation between different participants. 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. We tested the novel hypothesis that individual differences in consensus estimates occur because generic phrases encourage an intuitive overgeneralization (e.g., *Scientists believe = All scientists believe*) that some people revise downwards on reflection (e.g., *Scientists believe = Some scientists believe*). Two pre-registered studies failed to support this hypothesis. There was no significant relationship between reflective thinking and consensus estimates (Study 1) and enforced reflection did not cause estimates to be revised downwards (Study 2). Those reporting scientific research should be aware that generically attributing beliefs to 'Scientists' or 'Researchers' is ambiguous and inappropriate when there is no clear consensus among relevant experts.

Keywords: Generics; Generalization; News Headline; Inference; Cognitive Reflection Test; Scientific Consensus

Readers, listeners, and watchers of news media frequently encounter attention-grabbing headlines reporting scientific research. Listed below are three genuine examples taken from mainstream media outlets.

1. *Scientists believe* the secret of a good night's sleep is all in our genes (The Guardian, 2017)

2. *Experts think* early humans ate grass (BBC, 2012)

3. Eating more nuts could slow weight gain, *researchers say* (Sky News, 2019)

These brief snippets capture attention and succinctly communicate science to the public, but their brevity creates ambiguity. Do *all* scientists believe the secret of a good night's sleep is in our genes, *most* scientists or just *some* scientists? Because the noun phrase is unquantified (*Scientists believe*…) rather than universally (a*ll* scientists believe…) or exactly quantified (*one* scientist believes…), readers must form their own subjective interpretation about the degree of scientific consensus.

Statements such as 1-3 are known as 'generics'. They generically attribute a claim to 'scientists' rather than specifically to one scientist or one group of scientists. Generics are not just limited to communicating science but appear frequently in everyday discourse. Phrases such 'Ducks lay eggs', 'Tigers have stripes' and 'women read magazines' are not unusual (Abelson & Kanouse, 1966; Leslie, 2008). Generics, however, do have an unusual property: they require little evidence to be judged as true (Cimpian, Brandone & Gelman, 2010). While 'All men like DIY' might be considered false (perhaps because it only takes one counterexample to falsify this claim), the generic 'Men like DIY' is more likely to be accepted as true. Such generics are relatively immune to counterexamples, with statements such as 'Ducks lay eggs' or 'Mosquitos carry the West Nile virus' seeming to be true even when these claims are clearly not true for all members of the category (i.e., male ducks do not lay eggs and 99% of mosquitos do not carry West Nile virus) (Leslie, 2008).

Generics are frequently used to report primary scientific research (DeJesus, Callanan, Solis & Gelman, 2019). Academic papers routinely make general claims about *Humans, Adults, Males, Children, Introverts* and *Extroverts* etc., that gloss over variation within each category (e.g., DeJesus et al., 2019; Simons, Shoda & Lindsay., 2017). Generic phrases are also common in secondary reporting of scientific research by the news media. One specific function of generics, identified above, is to attribute the source of scientific claims using phrases such 'Scientists say…' or 'Experts believe…' (Robbins, 2012). Just as generic claims about 'Males' gloss over variation among males (DeJesus et al., 2019), generic claims about 'Scientists' gloss over variation in scientific opinion. Sacrificing precision for simplicity creates ambiguity. For example, the generic phrase 'Experts think…' can truthfully refer to almost any proportion of relevant experts[1]. It is up to the reader or listener to infer the degree of consensus. One reader may take this phrase to mean '*All*' relevant experts (e.g., a definitive consensus statement), another may take this to mean '*Most*' relevant experts while another may perceive it as a claim relating to just one specific subset of experts (e.g., the specific authors of a Study). The purpose of the two studies presented below is to reveal 1) the degree of scientific consensus implied by commonly used generic phrases such as "Scientists say…", "Experts think…" and "Researchers believe…" 2) to reveal the extent to which estimates of

[1] Although generic phrases such as 'Experts think…' can truthfully refer to almost any proportion of relevant experts there are likely to be some pragmatic limits to their use. For example, it would be true to say 'Experts think...' when referring to a tiny minority of experts, but to do so would violate Grice's Maxim of Quantity (i.e., give the most helpful amount of information). In this case it would be more informative to say 'Some experts' or 'A few experts'.

consensus vary between people and between phrases, and 3) to examine whether this variance is associated with the tendency to engage in *cognitive reflection*.

Cognitive Reflection

Put simply, cognitive reflection is the tendency to reflect on our intuitive ‘gut feelings’. Cognitive reflection is generally understood in the context of dual process models of human thinking (e.g., Toplak, West & Stanovich, 2011). Dual process models are based on the premise that humans engage in two distinct types of thinking. The first (generally known as Type 1 or intuitive thinking) is fast, instinctive and heuristic driven (e.g., correctly verifying that 2+2=4). It is independent of cognitive ability, delivering fast but not always accurate judgements (i.e., it is particularly susceptible to cognitive biases). The second type (generally known as Type 2 or reflective thinking) is a slower, more resource demanding and conscious type of thinking that is related to individual differences in cognitive ability (e.g., correctly verifying that 17x8=136).

Several prominent dual process accounts take a default-interventionist approach to explain how the two systems work together (e.g., Evans & Stanovich, 2013). Such theories assume that intuitive Type 1 processes act first, providing a fast and intuitive *default* response. Cognitively demanding Type 2 processes then *intervene* to revise that output, if required.

Often Type 1 processing suffices to make an accurate judgment or decision (e.g., correctly categorising an object as human or non-human, correctly verifying that 2+2=4, correctly recognising a friend etc.). Other times the output of type 1 intuitive thinking can fail to produce a normative response. The following problem is a classic example of how Type 1 processes can be misled:

A bat and a ball cost $1.10 in total. The bat costs $1.00 more than the ball.

How much does the ball cost? (in cents) _________

For those seeing this question for the first time, the answer that quickly comes to mind is ‘10 cents’. This is the fast and intuitive output of default Type 1 processing. Some people are content to retain this default answer without further reflection. Other people have a tendency to reflect on and revise their gut feelings. In other words, they are more likely to allow Type 2 processes to intervene. Those who do allow this intervention soon realise that the default response (10 cents) is incorrect and revise their answer to ‘5 cents’. Those people who are more likely to spontaneously reflect on their intuitions are said to have higher levels of Cognitive Reflection (Frederick, 2005).

Cognitive Reflection and scientific consensus

We propose that generic news media headlines lure readers towards an intuitive Type 1 conclusion in a similar way to the bat and ball problem. Specifically, we propose that generic headlines encourage the ‘hasty generalisation’ that ‘*Experts = All experts*’. This is consistent with work demonstrating that people are susceptible to overgeneralising from generic statements; a phenomenon known as the ‘Generic Overgeneralization Effect’ (Leslie, Khemlani & Glucksberg, 2011). If a generic statement is believed to be true (e.g., Ducks lay eggs) then some people erroneously overgeneralize that the equivalent universal statement is also true (e.g., *All* ducks lay eggs). Likewise, when participants have no knowledge about the truth of a generic (e.g., “Lorches have purple feathers”) some perceive it as referring to nearly all members of that category (i.e., nearly all Lorches have purple feathers) (Cimpian et al., 2010).

In the context of generic news headlines (e.g., Experts believe…) a hasty (over)generalisation would be that the opinion of ‘experts’ is the opinion of *all* or *nearly all* experts. This would be consistent with the work of Aklin and Urpelainen (2014) who concluded that in the absence of any dissenting information "*...the general public's default assumption is a very high degree of scientific consensus*". However, a little reflection (Type 2 thinking) reveals this generalisation is not necessarily true. If an individual reflects, they may identify counterexamples to their intuition. For example, they might question the plausibility of their initial judgement (complete consensus on any topic is rare), they might question the semantics of the phrase ('experts' could actually refer to just a few experts) or they might question the source of the message (the news media often sensationalise and gloss over the details). The application of Type 2 reflective thinking may therefore lead to the hasty generalisation being revised downwards (e.g., from 'most' or 'all' experts to 'some experts').

Our hypothesis is that people initially interpret generic phrases reporting scientific findings by making a hasty (over)generalisation (e.g., Experts think = All experts think). Those who reflect on this appealing inference are more likely to identify reasons (counterexamples) to revise the initial estimate of consensus downwards. Variation in estimates of consensus should therefore be negatively associated with variance in cognitive reflection (i.e., more reflective people perceive a lower degree of consensus).

This hypothesis was tested by two pre-registered studies. The first examined whether an individual's degree of trait Cognitive Reflection (measured both objectively and subjectively) is associated with the degree of consensus they perceive when reading common news media phrases such as 'Scientists believe…' and 'Experts say…'. We predicted that more reflective people would produce lower estimates of consensus, as they are more likely to override and revise their hasty (over)generalizations. The second Study then used time pressure and cognitive load to experimentally induce intuitive (Type 1) processing as participants rated the degree of consensus. We then encouraged them to actively reflect on their initial judgement with the option to revise it. By manipulating state reflection using this 'two-response paradigm' (Thompson, Prowse Turner & Pennycook, 2011) we aimed to determine whether there is a causal link between processing style and the revision of hasty (over)generalisations. We predicted that participants would produce relatively high estimates of scientific consensus (hasty generalisations) when engaged in fast Type 1 processing but would revise these downwards when encouraged to reflect on their initial gut feeling.

Study 1

Method

Participants were asked to estimate the degree of scientific consensus implied by common generic phrases such as “Scientists say…” and “Experts believe…”. We predicted that variance in these estimates would negatively correlate with variance in the ability to reflect on and override appealing but incorrect intuitions.

We chose to approach this question in the most controlled way possible, by eliminating the influence of context and prior beliefs on the interpretation of our generic phrases. To do this we used decontextualized phrases (e.g., *Scientists believe*…) rather than fully contextualised phrases. Consensus ratings of a contextualised headline such as “*Scientists believe climate change is due to human activity*” are more likely to be based on an individual’s prior beliefs about consensus on climate change than on the phrase “Scientists believe”. By presenting the generic phrases in isolation, we can be sure that estimates of consensus relate to the specific generic phrase and not to prior beliefs about specific issues. The aim is to generate relatively ‘pure’ estimates of consensus implied by different generic phrases. We acknowledge that these consensus estimates will inevitably change with the addition of context (e.g., “*Scientists believe climate change is due to human activity*” is likely to imply greater scientific consensus than the decontextualized “*Scientists believe…*”. Likewise, we would expect that “*Scientists believe the earth is flat*” to implies less scientific consensus. Our goal is to examine the degree of consensus implied by generic phrases in isolation. Consensus estimates will inevitably vary between participants and we predict that this variability will be associated with variability in Cognitive Refection.

The following protocol was pre-registered on the Open Science Framework prior to data collection https://osf.io/n7pj8/

Design

A cross-sectional design was used with the measures being: (i) an objective measure of cognitive reflection (CRT-7) (ii) a subjective measure of rational ability (REI Rational Ability subscale), (iii) a subjective measure of rational engagement (REI Rational Engagement subscale) (iv) scientific consensus estimates for nine decontextualized target phrases (e.g., “Scientists say…”, “Experts believe…”).

Participants

Because we planned to analyse the data using structural equation modelling, a power simulation was performed using the simsem package in R (code for this simulation can found on the Study OSF page). This indicated that a minimum sample of 200 participants would be required to achieve power of .8 (α=.05, two-tailed). We pre-registered a target sample size of 350 to account for possible data exclusions and to achieve a level of power in excess of .8.

Participants were recruited online via the www.prolific.co participant pool. This pool consists of over 70,000 registered users. A recent comparison of participant pools showed that Prolific users are naiver and more diverse than MTurk users, while providing a comparable quality of data (Peer, Brandimarte, Samat & Acquisti, 2017). Pre-screening ensured that the Study was only advertised to those aged 18 years or older and who spoke English as their first language. IP addresses were not collected as Prolific take a number of steps to avoid duplicate responses and automated responses by bots (Bradley, 2018).

In total, 355 participants consented to take part. Pre-registered exclusion criteria dictated that participants would be excluded if they did not complete the survey (these were considered to have withdrawn, as per our ethical approval conditions) or declared that they did not complete the survey seriously. Four participants did not complete the survey and were excluded. All remaining participants declared that they responded seriously, leaving a final sample of 351 participants aged 18-72 (M_{age} = 35.03, *SD* = 11.33). Of these 113 identified as male, 237 identified as female and one selected 'Other'. Participants were paid £1.30.

Materials

The survey was constructed using the Qualtrics online survey platform.

Measures of thinking style

Cognitive reflection was measured both objectively by the Cognitive Reflection Test and subjectively by the REI Rational Ability and Rational Engagement subscales. Higher scores on the CRT imply objectively greater ability reflect on and override Type 1 intuitions. Likewise, higher scores on the REI rationality subscale indicate a greater engagement with effortful, reflective thinking.

Objective measure of Cognitive Reflection. The Cognitive Reflection Test was used as an objective measure of reflective thinking. The validity of the original three-item CRT (Frederick, 2005) has been threatened by the widespread publication of the test materials, such that individuals with prior exposure to materials score significantly higher than those with no prior exposure (Haigh, 2016; Steiger & Reips, 2016). This raw score increase does not affect the Test's predictive ability (Bialek & Pennycook, 2018) but to minimise the impact of prior exposure we opted to use the longer CRT-7 (Toplak, West & Stanovich, 2014) which contains additional, less familiar questions. Designed to assess the ability to engage in analytic thinking, it assesses the tendency to override an appealing but incorrect intuitive response and engage in further reflection leading to the correct response. The CRT-7 comprises of seven mathematically worded problems (including the bat and ball problem) that cue an initial incorrect intuitive response that must be overridden to arrive at the correct conclusion. The overall CRT-7 score was calculated as the sum of all correct responses, with higher scores indicating a more reflective thinking style. The lowest possible score was 0 and the highest 7. In this Study the internal reliability measured using Cronbach's alpha was 0.77.

Subjective measure of Cognitive Reflection. The Rational Ability and Rational Engagement subscales of the Rational-Experiential Inventory (REI; Pacini & Epstein, 1999) were used as a measure of subjective preference for Type 2, reflective thinking. Participants were asked to read each statement and rate the extent that the statements referred or did not refer to them: e.g. "*I have a logical mind*" (1 = *definitely not true of myself*, 5 = *definitely true of myself*). The presentation order of items was randomised (see Keaton, 2017). Seventeen items were reverse scored and subscale scores were calculated as the mean of the relevant items. In this Study the internal reliability measured using Cronbach's alpha was 0.83 for the Rational Ability subscale and 0.86 for the Rational Engagement subscale.

Scientific Consensus measure. Participants were presented with nine decontextualized generic phrases (e.g., "Scientists say…", "Experts believe…", "Researchers think…") made up of a subject and a verb followed by ellipsis. See Table 1 for the list of phrases used.

Table 1: Target phrases rated by participants in Study 1

Scientists believe...	Researchers believe...	Experts believe...
Scientists say...	Researchers say...	Experts say...
Scientists think...	Researchers think...	Experts think...

Participants were asked to estimate how many relevant scientists [experts/researchers] they thought each phrase applied to on a sliding scale ranging from zero '*no [scientists/experts/researchers]* ' to 100 '*all [scientists/experts/researchers]'*.

"Please estimate how many relevant [scientists/experts/researchers] you think this statement applies to. We are simply interested in your personal opinion. There is no right or wrong answer."

The scale was not numbered, however a number between 0-100 was visible to participants as they moved the slider.

The subjects of our nine target phrases were "Scientists" [three items], "Experts" [three items] and "Researchers" [three items] followed by a base verb that implied some degree of consensus (e.g., 'believe', 'say', 'think'). The three subjects were chosen as a corpus search showed they are frequently used as general terms to describe the authors of scientific work. Other subjects considered after searching for synonyms of the three nouns above were Academics, Scholars, Doctors, Lecturers and Professors but a corpus search revealed that these were infrequently used to convey scientific consensus. More specific job titles such as 'psychologists', 'neuroscientists' or 'physicists' were considered beyond the scope of this investigation.

The subjects were paired with verbs that implied some degree of scientific consensus. To select the verbs, we searched the British National Corpus using an online interface (https://www.english-corpora.org/bnc/) to identify the 30 base verbs that most frequently collocate with "Scientists", "Experts" and "Researchers". From these lists we selected three base verbs that frequently accompanied each of our subjects. These were:

1) '*believe*' (ranked as the most frequent collocate of 'Scientists' and 'Researchers', ranked second most frequent collocate of 'Experts)

2) '*say*' (ranked as the most frequent collocate of 'Experts' and second most frequent collocate of 'Scientists' and 'Researchers')

3) '*think*' (ranked as the third most frequent collocate of 'Scientists' and 'Researchers' and the seventh most frequent collocate of 'Experts')

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. These were paired with three verbs that frequently accompany the subjects (these were 'suggest', 'agree', 'argue'). These different verbs were chosen to make the task less repetitive for participants. The filler phrases are presented in Table 2.

Table 2: Filler phrases rated by participants in Study 1

Some scientists argue…	Some researchers suggest...	Some experts agree...
Many scientists suggest...	Many researchers agree...	Many experts argue...
Few scientists agree...	Few researchers argue...	Few experts suggest...

The internal reliability measured using Cronbach's alpha was 0.97.

Seriousness check. To exclude non-serious responses, participants were asked at the end of the Study whether they took part seriously. Seriousness checks have been suggested to substantially improve the quality of data collected (Aust, Diedenhofen, Ullrich & Musch, 2013). Participants were advised they would still be paid even if they did declare non-serious responding. Our pre-registered exclusion criteria dictated that non-serious participants would be excluded from the analysis.

Procedure

Ethical approval for all studies in this paper was granted through the faculty ethics committee of the authors' institution. An information sheet at the start of the Qualtrics survey informed participants they would be asked about their understanding of some phrases commonly used by the media, be asked to solve some simple problems and to answer questions regarding their thinking style.

Participants first saw the 18 phrases (9 target items plus 9 fillers) and provided consensus estimates for each. These were presented one per page in a different random order to each participant. They then completed the CRT-7, the REI (with statements presented in a different random order for each participant), and then the seriousness check. All items required a response. Participants could not progress beyond a page until all questions had been answered. Mean completion time was 13.15 minutes (*SD* = 7.71).

Results

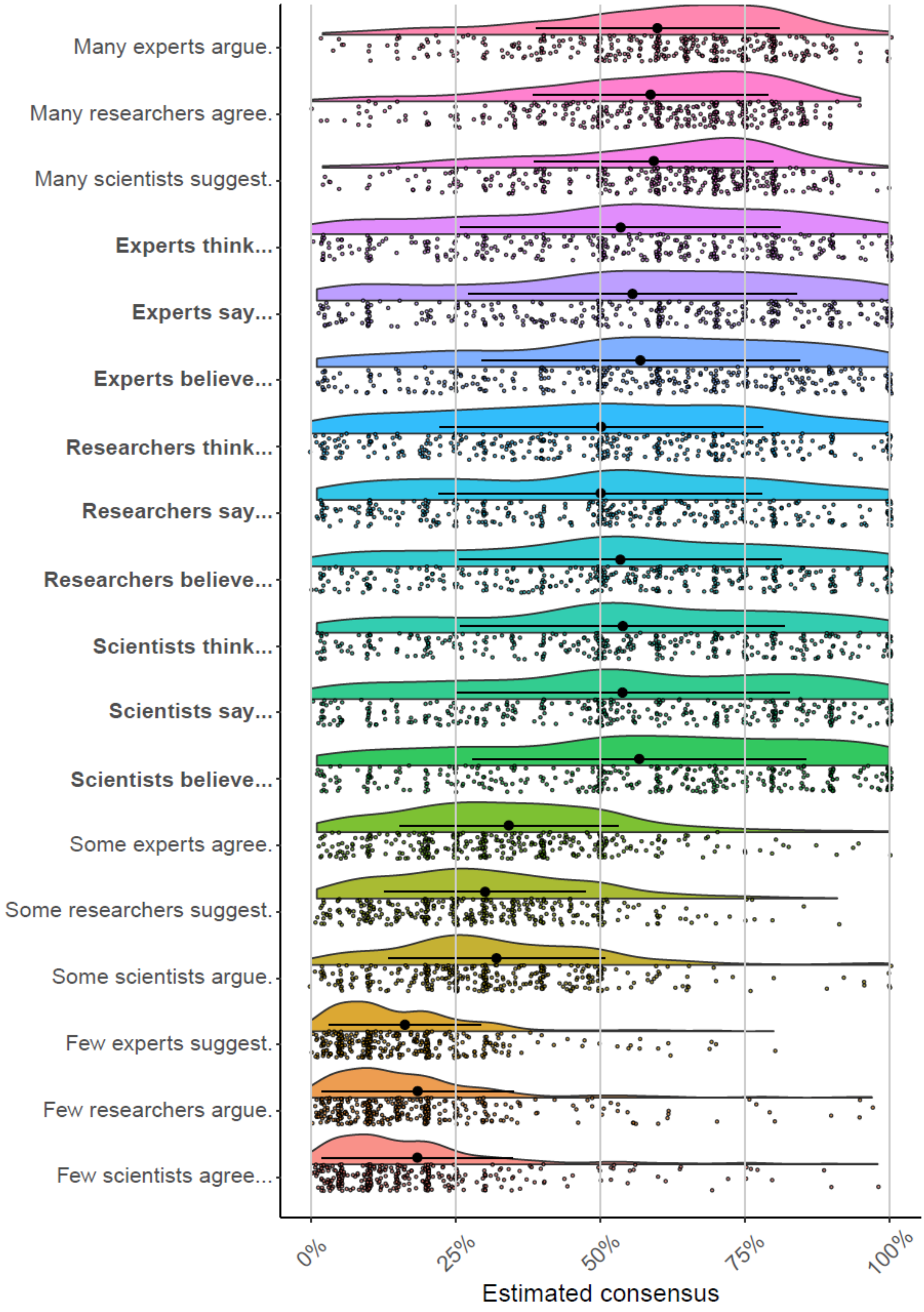
Pre-registered analysis

The following analyses were pre-registered on the Open Science Framework prior to data collection https://osf.io/n7pj8/registrations. The raw data and R analysis script are publicly available via the OSF https://osf.io/n7pj8/ . There were no missing data as all questions required a response.

The mean CRT-7 score was 2.65 correct answers (*SD*= 2.16). The mean REI rational ability and rational engagement subscale scores were 3.48 (*SD*= 0.62) and 3.44 (*SD*=0.67) respectively. The mean rating of consensus implied across our nine generic phrases was 53.82 (*SD*= 25.60) on a 0-100 scale (see Figure 1 for mean ratings of each individual phrase).

As we aimed to measure the extent to which estimates of consensus vary between people and between phrases, we calculated descriptive statistics averaged over our 351 participants (*M*=53.82, *SD*=25.60) and averaged over our nine items (*M*=53.82, *SD*=2.48) separately. The variability in means between participants was approximately ten times greater than the variability between items.

Figure 1: Mean degree of consensus implied by the phrases used in Study 1 (Averaged over 351 participants, Error bars represent one standard deviation). Generic phrases in bold. The scale was anchored at 0 (no scientists/experts/researchers) and 100 (all scientists/experts/researchers).



Consistent with previous research, the CRT-7 was positively correlated with the REI Rational Ability (*r*(349) =.285 *p*<.001) and REI Rational Engagement (*r*(349) =.276 *p*<.001) subscales. Ratings of all nine generic consensus ratings were strongly and significantly correlated with each other (correlations ranged from *r*(349) = .74, *p*<.001 to *r*(349) = .86, *p*<.001).

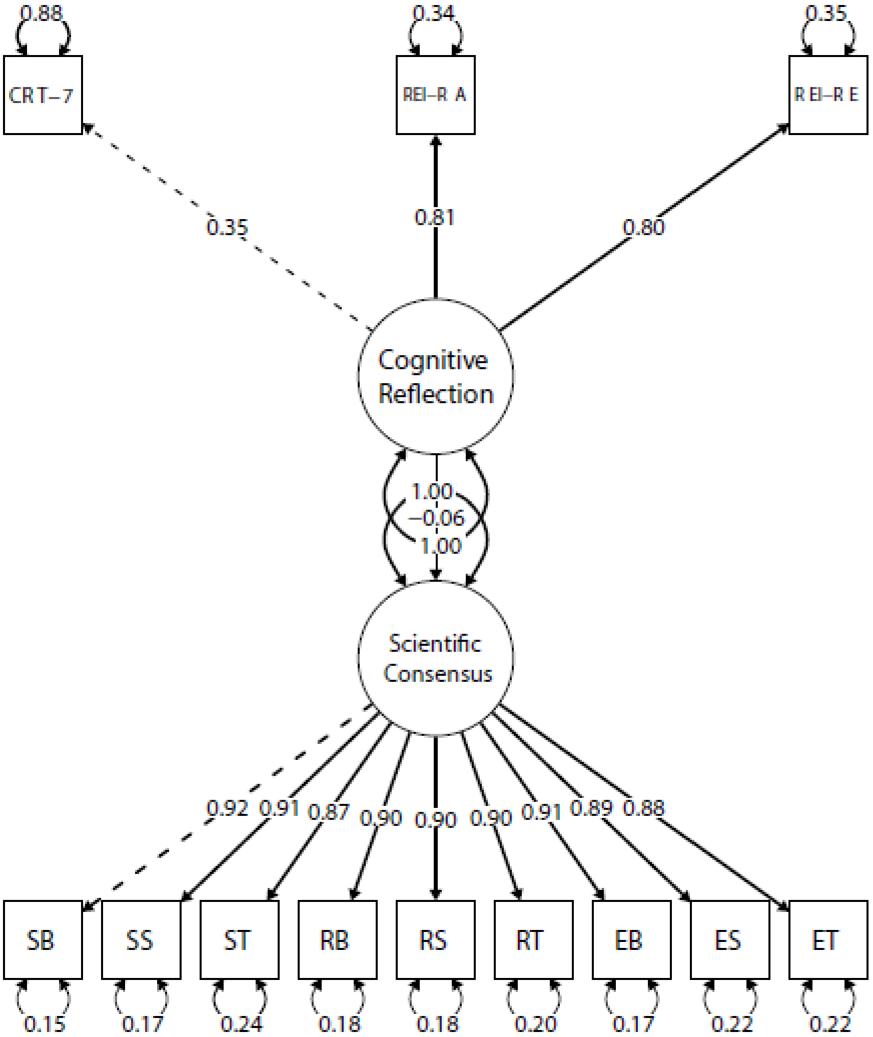
Following our pre-registered analysis plan we used Structural Equation Modelling (SEM) to examine the direct path between a latent 'Cognitive Reflection' variable and a latent 'Scientific Consensus' variable (see Figure 2). We predicted there would be a significant negative relationship between these latent variables (i.e., more reflective people make lower estimates of consensus). The latent Cognitive Reflection variable was measured using CRT-7 total score, REI Rational Ability subscale and REI Rational Engagement subscale. The latent 'Scientific Consensus' variable was measured using consensus ratings to the nine generic phrases described above. The filler phrases qualified with ‘some’, ‘many’ and ‘few’ were not part of the latent 'Scientific Consensus' variable.

Analysis was conducted using the lavaan package (Rosseel, 2012) in R version 3.6.0 (R Core Team, 2019). Prior to conducting the analysis, scientific consensus ratings were transformed to *z* scores to reduce the difference in variance magnitude between these ratings and the other measures, this was unforeseen but required as our models did not converge. This transformation was not pre-registered but transformations such as these are commonly employed to achieve convergence and do not affect the fundamental results (Little, 2013:17)

Model fit was assessed using $\mathcal{X}^2$, CFI, TLI and RMSEA. The model $\mathcal{X}^2$ was significant ($\mathcal{X}^2$ (53) =141.848, *p*<.001), which is typical for large samples (Bollen, 1989; Lance & Vandenberg, 2001). CFI (0.979) and TLI (0.974) values were greater than 0.95 and RMSEA was 0.069 (90% CI: 0.055 - 0.083) indicating an adequate model fit.

Figure 2 shows that there was a negative relationship between the latent 'Cognitive Reflection' variable and a latent 'Scientific Consensus' variable, but this was weak and not statistically significant (β = -0.063, *SE*=0.075, *p*=.304).

Figure 2: Structural equation model of the relationship between Cognitive Reflection and estimates of Scientific Consensus implied by nine generic phrases. Cognitive Reflection and Scientific Consensus are latent variables (Standardised solution, N=351). The model $\mathcal{X}^2$ was significant ($\mathcal{X}^2$ (53) =141.848, *p*<.001), CFI (0.979) and TLI (0.974) values were greater than 0.95 and RMSEA was 0.069 (90% CI: 0.055 - 0.083). Figure labels: CRT-7 = Cognitive Reflection Test (7 item version), REI-RA = REI Rational Ability subscale, REI-RE = REI Rational Engagement subscale, SB= Scientists believe…, SS= Scientists say…, ST= Scientists think…, RB= Researchers believe…, RS= Researchers say…, RT= Researchers think…, EB= Experts believe…, ES= Experts say…, ET= Experts think...



Exploratory analysis

In addition to the nine generic phrases used in Study 1, which were the focus of this Study, participants also saw nine 'filler' phrases which were included to make the task less repetitive. Three of these phrases were quantified with 'some', three with 'many' and three with 'few' (see Table 2). During the peer review process, it became apparent that our hypothesis that people overestimate consensus and then revise downwards on reflection, may also be relevant to the interpretation of verbal quantifiers more generally (i.e., phrases such as 'many scientists…' and 'few scientists…'). To explore this possibility, we calculated bivariate correlations between each of our nine fillers and our three measures of cognitive reflection (see Table 3). Interpretations of the 'many' fillers were largely unrelated to our cognitive reflection measures. Interpretations of the three 'Few' statements had weak negative relationships with our three measures of cognitive reflection. There was also some evidence of a similar pattern with our 'some' statements.

Table 3: Exploratory bivariate correlations (Pearson's *r*) between each of our filler phrases and our three measures of cognitive reflection.

	CRT-7	**Rational Ability**	**Rational Engagement**
Some scientists argue…	-0.12*	-0.06	-0.06
Some researchers suggest...	-0.1†	-0.04	-0.09
Some experts agree…	-0.13*	-0.1†	-0.12*
Many scientists suggest...	-0.03	-0.03	-0.12*
Many researchers agree...	0.04	-0.02	-0.07
Many experts argue…	0.0001	-0.03	-0.05
Few scientists agree...	-0.21**	-0.16**	-0.15**
Few researchers argue...	-0.16**	-0.13*	-0.11*
Few experts suggest...	-0.14**	-0.15**	-0.1†

Note: †*p*<0.1, **p*<0.05, ***p*<0.01

Discussion

The first aim of this paper was to reveal the degree of scientific consensus implied by commonly used generic phrases such as "Scientists say…", "Experts think…" and "Researchers believe". Study 1 shows that such decontextualized phrases imply a slim majority of scientists (Mean estimate of consensus was 53.8% of relevant scientists/experts/researchers). This mean consensus estimate was greater the mean estimates for filler phrases quantified with 'Few' (17.7%) or 'Some' (32.1%) and lower than fillers quantified with 'Many' (59.3%) (see Figure 1).

The second aim was to reveal the extent to which estimates of consensus vary between people and between phrases. Mean estimates varied very little between our nine commonly used phrases (*SD*=2.5) but varied widely between our 351 participants (*SD*=25.6). On average, the nine phrases we selected each implied a similar level of consensus (ranging from 50.1% to 56.9%) suggesting that generics imply a similar degree of consensus, regardless of the subject or verb used (e.g., 'Researchers say...' tends to be interpreted in much the same way as 'Scientists think...'). In contrast, our participants behaved very differently from each other, with 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. Both extremes are plausible interpretations of a generic.

The third aim was to examine whether this variance in estimates between participants is associated with variance in cognitive reflection. We predicted that a latent Cognitive Reflection variable (in which cognitive reflection was measured both objectively and subjectively) would be negatively associated with estimates of consensus implied by generic phrases (e.g., Experts think…). This prediction was based on the hypothesis that generic phrases encourage people to make an intuitive overgeneralization (e.g., Experts think = All experts think) that is then revised downwards by those who spontaneously engage in cognitive reflection. A weak negative relationship was observed between cognitive reflection and estimates of consensus, but this was not statistically significant. Therefore, the data from this highly powered Study do not support our hypothesis.

Exploratory analysis was also conducted on the filler items (i.e., phrases quantified with 'some', 'many' and 'few') which was not pre-registered. This allowed us to explore whether hypothesis that people overestimate consensus and then revise downwards on reflection may be relevant to the interpretation of verbal quantifiers more generally (i.e., phrases such as 'many scientists…' and 'few scientists…'). This analysis revealed weak negative relationships between filler phrases quantified with 'few' (e.g., Few scientists agree...) and our measures of cognitive reflection. A similar but less consistent pattern was observed for fillers quantified with 'some'. This suggests that those who are more reflective tend to give lower consensus estimates to phrases quantified with 'few' and 'some', consistent with the idea that people initially overestimate consensus and revise their estimate downwards on reflection. Future confirmatory research is required to confirm this negative relationship between the interpretation of verbal quantity phrases (such as 'few' and 'some') and cognitive reflection.

One explanation for the null findings in Study 1 is that estimates of scientific consensus implied by generic phrases are unrelated to trait cognitive reflection. However, before accepting this conclusion, an alternative possibility should be explored. This possibility is that the Study protocol actively encouraged participants to engage in analytical thought, making estimates of consensus insensitive to trait variance in cognitive reflection. Explicitly asking participants to assign a quantity to the nine phrases may have encouraged them to reflect on their meaning in a much deeper way than they typically would (i.e., engaging Type 2 thinking). Therefore, all participants may have revised their initial estimates before giving a final response, rather than just those who more readily engage in reflective thinking.

In Study 2 we sought to overcome this limitation and determine whether there is a cause and effect relationship between processing style and estimates of consensus. To do this, we used a two-response paradigm (Thompson et al., 2011) that was designed to force an intuitive estimate (response 1), before giving participants the opportunity to reflect on and revise that estimate (response 2). If estimates of consensus are related to thinking style, then this direct experimental manipulation will result in relatively higher estimates of consensuses when the initial intuitive estimate is given (due to participants over generalising) and relatively lower estimates when they are encouraged to reflect on their intuition.

Study 2

Method

In this second online Study, participants were asked to estimate the degree of scientific consensus implied by common generic phrases such as "Scientists say…" at two time points. At Time 1 (T1) participants made an intuitive consensus judgement under cognitive load and time pressure to encourage Type 1 processing by suppressing Type 2 processing. At Time 2 (T2) the cognitive load was removed, and participants were given unlimited time to consider their first response with the option to revise their estimate, if desired. We predicted that intuitive consensus estimates at T1 would be relatively high and would be revised downwards on reflection at T2. This is an adaption of the two-response paradigm developed by Thompson et al. (2011).

The following protocol was pre-registered on the Open Science Framework prior to data collection https://osf.io/2vh9w/

Design

A repeated measures experimental design was used with the response time point (T1 and T2) being the repeated measures independent variable and the degree of scientific consensus participants assigned to the target phrase the dependent variable.

Participants

Power analysis was conducted assuming one fixed factor (which is repeated measures with two levels) and two random factors (participants and items). Power analysis was conducted using the two random factor Power calculator developed by Westfall, Kenny & Judd (2014). With an anticipated effect size of *d*=0.5, we found that the most efficient and practical design to achieve power of 0.8 would require a minimum of 16 target items and 118 participants.

As per Study 1, participants were recruited online via the www.prolific.co participant pool. Pre-screening ensured that the Study was only advertised to those aged 18 years or older, who spoke English as their first language and who did not take part in either Study 1 or the pre-test for Study 2.

In total, 194 participants consented to take part. Participants were excluded if they did not complete the survey (N = 12). All participants declared that they completed the survey seriously. This left a final sample of 182 participants aged 18-73 (M_{age} = 36.92, SD = 11.68). Of these, 93 identified as male, 88 identified as female and one selected 'Other'. Participants were paid £1.50.

Materials

A detailed description of the Study 2 materials can be found in Appendix 1.

The survey was created using the Qualtrics online platform. All nine of the generic decontextualized phrases from Study 1 were used along with an additional seven generic phrases (16 target items in total). Four of the additional phrases used "Academics" as the subject; it being the fourth most frequently used subject to describe authors of scientific work. This was paired with the same base verbs as Study 1 ("Academics Believe…", "Academics Say…", "Academics Think…"). An additional base verb "agree" was then selected as a corpus search revealed it to be the fourth most popular verb that frequently collocated with our four subjects and implied some degree of scientific consensus (ranked as the third most frequent collocate for "Researchers" and "Academics", ranked seventh most frequent collocate of Researchers, and ranked 29th most frequent collocate of

“Scientists”). “Agree” was paired with the four subjects to create the additional decontextualized phrases: “Scientists Agree…”, “Experts Agree…”, “Researchers Agree…”, “Academics Agree…”.

Using the same slider layout as Study 1, participants were asked to estimate how many relevant experts they thought each phrase applied to on a sliding scale ranging from zero ‘*No scientists/experts/researchers/academics’* to 100 ‘*All scientists/experts/researchers/academics’*.

The 16 target items were presented amongst four filler items which used the same subject and base verbs as the targets but were prefaced with quantifiers (“Some Scientist Believe...”, “Most Experts Say...”, “Many Researchers Think...”, “Few Academics Agree”). These were simply to make the task less repetitive for participants.

Prior to starting the experimental task participants completed four practice items to accustom them to making intuitive judgements under time pressure. These were also structured using the same subject and base verbs as the target items but with a quantifier prefacing them (worded differently to the filler items) to accustom subjects to using the slider. Given the time pressure element, a pre-screening requirement was that participants could only complete the Study on a desktop computer with a click-based mouse (i.e. not a touchpad).

Procedure

In this experiment the rating of scientific consensus was given twice for each phrase. The two-response format was similar to that introduced by Thompson et al., (2011) which was developed to gain insight into the time-course of intuitive and deliberate responses. Participants provided two estimates for each generic phrase. For the first response, participants were asked to give the very first estimate that came to mind (i.e. an intuitive estimate). To ensure that participants gave an intuitive response they gave their initial estimate under time pressure and under cognitive load (this application and time pressure and cognitive load in an online Study has been previously been used by Bago & De Neys, 2019a, 2019b; Raoelison & De Neys. 2019). Cognitive load was applied by asking participants to memorise a pattern of four crosses presented in a 3x3 matrix for later recall (see Appendix 1 for details).

Time pressure was applied by instructing participants to select a point on the sliding scale within three seconds (this time limit was determined by a reading time pre-test conducted with 41 participants, see Appendix 2 for details of the pretest). This initial fast response is thought to reflect the output of System 1 processes based on evidence that that rapid responses rely more on automatic heuristic processing and therefore likely reflect System 1 output (Evans & Stanovich, 2013; Pennycook, De Neys, Evans, Stanovich & Thompson, 2018).

Following the initial intuitive response, participants were asked to identify the pattern they saw from four options (thus removing the cognitive load). Feedback was immediately given to indicate whether the answer was correct or incorrect. The generic phrase was then presented again. Participants were reminded of their initial intuitive estimate (e.g., “Your intuitive estimate under time pressure was 85 on a scale of 0-100”) and shown a slider that was automatically set to this initial estimate. They were instructed that they could take as long as they liked to reflect on their initial estimate and had the option to revise their answer (if desired) by moving the slider. If participants did not wish to revise their intuitive response, they were advised to press the ‘continue’ button to start the next trial. This second response (T2) is thought to reflect the output of System 2 processes which are believed to be intentional, conscious and time consuming (Evans & Stanovich, 2013; Pennycook et al., 2018).

Prior to completing the experimental task, participants completed a practice phase (described in detail in Appendix 1) where the memory and ratings tasks were introduced in stages to build familiarity. After the practice phase, each participant saw the target and filler phrases presented in a randomised order and completed the two-response protocol for each phrase (see Appendix 1 for detailed procedure). All phrases at T1 required a response and participants could not progress until a response was given. Mean completion time was 14.10 minutes (*SD* = 7.70).

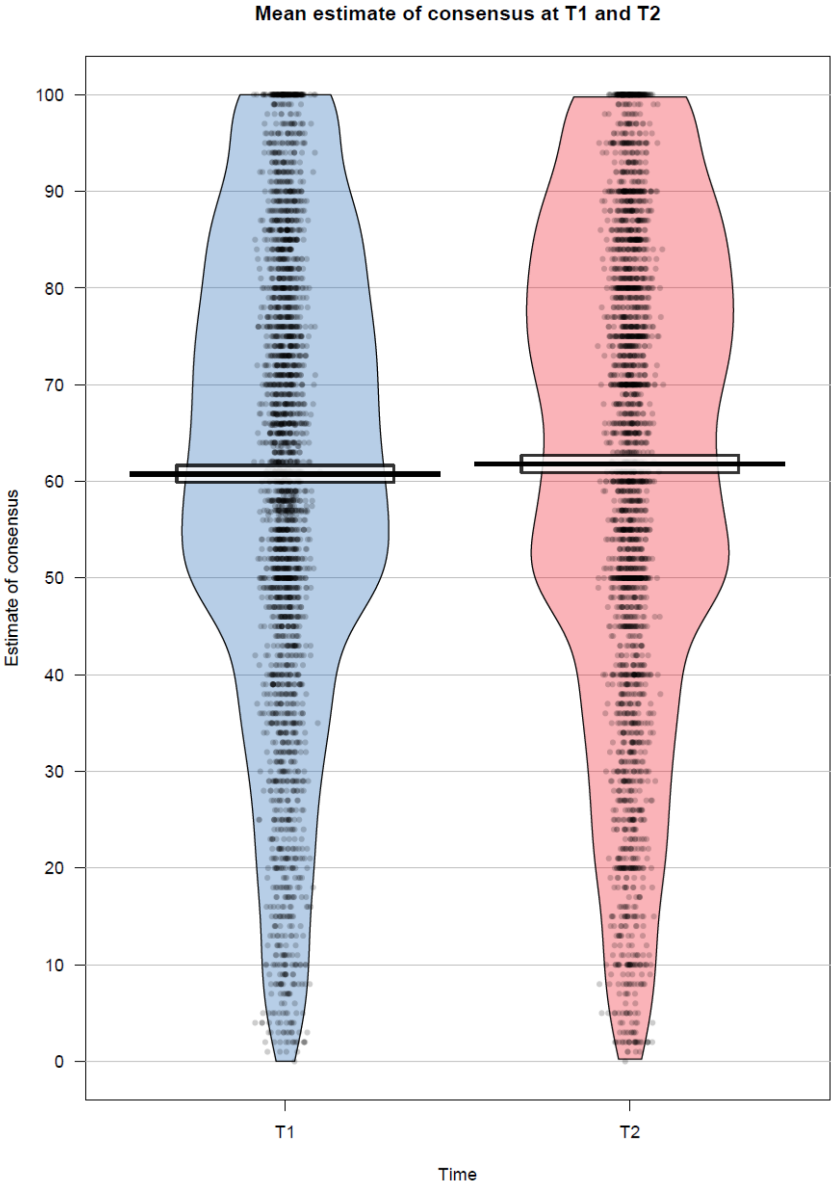
Results

The following analyses were pre-registered on the Open Science Framework prior to data collection https://osf.io/2vh9w/registrations. The raw data and analysis script are publicly available via the OSF https://osf.io/2vh9w/.

There were no missing data as all T1 questions required a response. Descriptive statistics suggest that participants engaged with the task as instructed. First, the average participant correctly recalled 92.2% of the matrices (*SD*=7.8) suggesting that cognitive load was successfully applied at T1 (in other words, participants engaged with the instruction to memorise each matrix). Second, participants revised their intuitive (T1) rating at T2 on 47.8% of trials, suggesting that they frequently revised their initial (T1) estimate, rather than simply retaining it.

Following our pre-registered protocol, we excluded 146 T1 estimates where a response was not given within 3.25 seconds. These excluded observations represented 5% of the T1 data. Excluded values were replaced using mean imputation. Figure 3 shows that the mean T1 estimate was 60.74 (*SD*= 20.15) and the mean T2 estimate was 61.82 (*SD*=21.96).

Figure 3: Mean estimate of consensus at T1 and T2 (*N*=182). The mean is represented by a horizontal line. The band represents 95% Bayesian Highest Density Interval.



A linear mixed effects model was fitted using the lme4 package (Bates, Maechler, Bolker & Walker, 2015) in R version 3.6.0 (R Core Team, 2019). The fixed factor was Time (T1 vs T2). Participants and Items were treated as random factors. We began with a maximal random effects structure (Barr, Levy, Scheepers & Tily, 2013) with random slopes and intercepts for Participants and Items. This resulted in a singular fit. We therefore systematically removed random effects until we achieved a non-singular fit. We first removed random slopes by items. This reduced model also resulted in a singular fit. We further reduced the model by removing random slopes by participants. This 'intercepts only' model achieved a non-singular fit.

The final non-singular model included random intercepts for both Participants and Items. We compared this intercepts-only model to a null model (without the fixed effect) using a likelihood ratio test. There was a significant difference between the likelihood of these two models (χ^2 (5) =10.533, *p*=.00116). This indicates that the fixed effect of Time affected perceived scientific consensus, with T2 estimates of consensus an average of 1.07 points (*SE*=0.33) greater than T1 estimates (on our 101 point scale). This effect was in the opposite direction to our pre-registered prediction. A standardised effect size (*d*) was calculated by dividing the mean difference between conditions by the square root of the pooled variance components (Westfall, Kenny & Judd, 2014). The size of this effect was small (*d*=0.043).

Discussion

In Study 2, participants estimated the degree of consensus implied by 16 generic phrases (e.g., Scientists say…) on a 0-100 scale that ranged from 'No scientists' to 'All scientists'. They did this on two occasions. The first (T1) was under time pressure and under cognitive load. The second (T2) was an opportunity to revise the initial estimate without any time pressure or cognitive load. We predicted that participants would provide relatively high estimates of consensus at T1, with fast and intuitive System 1 processing leading to a hasty overgeneralization. At T2 we predicted that this initial estimate would be revised downwards by more deliberative System 2 processing. We found the opposite effect. The average consensus estimate at T1 was 60.74 and at T2 the average estimate was revised upwards to 61.82. The size of this upwards revision was small (*d*= 0.043).

These data do not support our hypothesis that participants would revise their intuitive T1 estimates of consensus downwards at T2. The most common behaviour was for participants was to make no revision to their intuitive estimate (T1 and T2 estimates were identical on 52.2% of trials). When a revision was made, this was most frequently an upwards revision (T2 estimates were higher than T1 on 30.8% of trials). Revisions were only made in the predicted (downwards) direction on 17% of trials.

General Discussion

Generic news headlines (e.g., "*Scientists believe the secret of a good night's sleep is all in our genes"*) are inherently ambiguous. They require the reader or listener to infer the degree of consensus among 'Scientists'. Do most scientists believe there is an important link between sleep and genes or just a select few? Those who do not read beyond the headline must rely on a subjective inference. These inferences can be consequential, as perceived scientific consensus on a given issue is an important factor in determining our own beliefs on that issue (e.g., van der Linden et al., 2015). In this paper we set out to 1) reveal the degree of scientific consensus implied by commonly used generic phrases such as "Scientists say…", "Experts think…" and "Researchers believe…" 2) to reveal the extent to which estimates of consensus vary between people and between phrases , and 3) to examine whether this variance is associated with a reflective thinking style. We did this by running two highly powered, pre-registered studies.

In terms of the first aim, we revealed that commonly used generic phrases such as "Scientists say…", "Experts think…" and "Researchers believe…" imply consensus among over half of all relevant experts (53.8% in Study 1, 60.74% in Study 2 T1 and 61.82% in Study 2 T2) . These ratings were based on decontextualized phrases and are therefore labile; prone to change with the addition of context. These baseline estimates are nonetheless important, as they show that in the absence of any other context generic phrases about experts imply consensus among a *majority* of experts. This starting point may be revised upwards or downwards with additional context but in cases where the context is new or unfamiliar to the reader (e.g., *Scientists think it rains diamonds on Neptune*) the generic consensus statement (e.g., Scientists think…) may be all a they have to go on.

While in some cases it may be true that a generic such as 'scientists believe…' refers to a majority of scientists, much of the novel research hitting the headlines is the work of just one group of authors. Using the generic 'scientists' or 'experts' risks implying that the opinion of a small subset of experts is the opinion of the majority of experts. This risks being misleading and confusing to the public. For example, when research by one group of authors is attributed generically to experts (e.g., Scientists believe one glass of wine per day is *beneficial*) it may imply that this is the opinion of a majority of experts; when contradictory findings by a different group of authors are reported (e.g., Scientists believe one glass of wine per day is *harmful*) it implies that the majority of experts have changed their minds. When this happens several times, there is a risk that it will damage trust in science and expert advice (e.g., *Scientists say one thing one day and another the next!*) (see Koehler & Pennycook, 2019 for related work). For these reasons, responsible reporting should avoid implying consensus through generic statements and instead use more specific quantity terms (e.g., *A group of scientists*…; '*Some scientists…*').

In terms of our second aim, the data revealed that mean estimates varied little between phrases (Study 1 *SD*=2.5) but varied widely between participants (Study 1 *SD*=25.6). In other words, estimates of consensus were very similar for all of our generic phrases (mean estimates of consensus ranged from 50.1 to 56.9). Generics, therefore, appear to imply the same degree of consensus regardless of whether the subject is 'scientists', 'experts' or 'researchers' or whether they 'believe', 'think' or 'say'. In contrast, participants varied widely in their mean estimates of our generics (mean estimates of consensus ranged from 2% to 100%). The variance between participants was approximately ten times greater than between items. People thus draw very different inferences about the degree of consensus implied by generics, with some inferring they refer to very few experts and others to *all* experts. Those who make the strong inference that that 'scientists' refers to all or nearly all scientists are particularly at risk of changing their beliefs and behaviours based on the opinion of what could be a small minority of experts and of perceiving major scientific U-turns when different groups of experts report contradictory findings. In contrast, those at the lower end of the spectrum may be less inclined to change their beliefs or behaviours even when there is widespread consensus.

The wide variability between participants in Study 1 may in part be an artefact of our decision to use decontextualized phrases. The variability in consensus estimates of "scientists believe…" is inevitably going to greater than the variability corresponding to the same phrase used in context. What this does show is that in the absence of any other context, different people have different starting points for interpreting generics. The purpose of this Study was to examine whether variation in this decontextualized baseline estimate was related to variance in cognitive reflection. We tested the novel prediction that decontextualized generic phrases such as "Scientists believe…" encourage an intuitive overgeneralization about the degree of scientific consensus (e.g., *Scientists believe = All scientists believe*) that is then revised downwards to varying extents, on reflection (e.g., *Scientists believe = Some scientists believe*).

In Study 1 we found no evidence to suggest that trait cognitive reflection is associated with estimates of scientific consensus. Likewise, in Study 2, we found no evidence that enforced reflection causes people to revise their intuitive estimates downwards. This suggests that the interpretation of a generic does not involve a process of intuitive overgeneralization and reflective belief revision. This leaves us without a satisfactory explanation for why estimates of consensus vary. Possible avenues for future research may be to explore the roles of experiential factors such as engagement with science, knowledge of the scientific process, level of education or cognitive factors such critical thinking, intelligence and need for cognition. Research should also focus on the implication of this variation in perceived consensus to determine the effects it has on subjective beliefs and behaviours.

Generic attributions to 'scientists', 'researchers' and 'experts' are common in the media and social media, but they are inherently ambiguous. When generic attributions are presented out of context estimates of consensus vary widely, but on average they imply that a majority of the category (e.g., researchers, scientists, experts) are in agreement. For this reason, those reporting scientific research should avoid using generics when there is not a majority consensus among relevant experts and consumers of news should be aware that generics can intentionally or unintentionally misrepresent the true degree of consensus.

References

Abelson, R. P., & Kanouse, D. E. (1966). Subjective acceptance of verbal generalizations. In S. Feldman (Ed.), *Cognitive consistency* (pp. 171-197). Academic Press.

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

Aust, F., Diedenhofen, B., Ullrich, S., & Musch, J. (2013). Seriousness checks are useful to improve data validity in online research. *Behavior Research Methods*, *45*(2), 527-535. https://doi.org/10.3758/s13428-012-0265-2

Bago, B., & De Neys, W. (2019a). The intuitive greater good: Testing the corrective dual process model of moral cognition. *Journal of Experimental Psychology: General, 148*(10), 1782-1801. https://doi.org/10.1037/xge0000533

Bago, B., & De Neys, W. (2019b). The smart System 1: Evidence for the intuitive nature of correct responding on the bat-and-ball problem. *Thinking & Reasoning*, *25*(3), 257-299. https://doi.org/10.1080/13546783.2018.1507949

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255-278. https://doi.org/10.1016/j.jml.2012.11.001

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*(1), 1-48. https://doi.org/10.18637/jss.v067.i01

BBC Newsround. (2012, November 12). *Experts think early humans ate grass.* https://www.bbc.co.uk/newsround/20302443

Bialek, M., & Pennycook, G. (2018). The cognitive reflection test is robust to multiple exposures. *Behavior Research Methods*, 50(5), 1953-1959. https://doi.org/10.3758/s13428-017-0963-x

Bollen, K. A. (1989). *Structural equations with latent variables*. John Wiley.

Bradley, P. (2018, August 10). Bots and data quality on crowdsourcing platforms. *Prolific Blog*. https://blog.prolific.co/bots-and-data-quality-on-crowdsourcing-platforms/

Cimpian, A., Brandone, A. C., & Gelman, S. A. (2010). Generic statements require little evidence for acceptance but have powerful implications. *Cognitive Science*, *34*(8), 1452-1482. https://doi.org/10.1111/j.1551-6709.2010.01126.x

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. https://doi.org/10.1073/pnas.1817706116

Evans, J. S. B., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science*, *8*(3), 223-241. https://doi.org/10.1177/1745691612460685

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic perspectives, 19*(4), 25-42. https://doi.org/10.1257/089533005775196732

Haigh, M. (2016). Has the standard cognitive reflection test become a victim of its own success? *Advances in Cognitive Psychology*, *12*(3), 145-149. https://doi.org/10.5709/acp-0193-5

Keaton, S. A. (2017). Profile 53. Rational-Experiential Inventory–40 (REI-40). In Worthington, D. L., & Bodie, G. D. *The sourcebook of listening research: Methodology and measures*. John Wiley & Sons.

Koehler, D. J., & Pennycook, G. (2019). How the public, and scientists, perceive advancement of knowledge from conflicting Study results. *Judgment & Decision Making*, *16*(6), 671-682.

Lance, C. E., & Vandenberg, R. J. (2001). Confirmatory factor analysis. In F. Drasgow & N. Schmitt (Eds.), *Measuring and analyzing behavior in organizations: Advances in measurement and data analysis* (pp. 221-256). Jossey-Bass.

Little, T. D. (2013). *Longitudinal structural equation modeling*. Guilford press.

Leslie, S. J. (2008). Generics: Cognition and acquisition. *Philosophical Review*, *117*(1), 1-47. https://doi.org/10.1215/00318108-2007-023

Leslie, S. J., Khemlani, S., & Glucksberg, S. (2011). Do all ducks lay eggs? The generic overgeneralization effect. *Journal of Memory and Language*, *65*(1), 15-31. https://doi.org/10.1016/j.jml.2010.12.005

McKie, R. (2017, April 9). *Scientists believe the secret of a good night's sleep is all in our genes*. The Guardian. https://www.theguardian.com/science/2017/apr/09/science-of-sleep-genetics

Mee, E. (2019, September 24). *Eating more nuts could slow weight gain, researchers say*. Sky News. https://news.sky.com/story/eating-more-nuts-could-slow-weight-gain-researchers-say-11817911

Pacini, R., & Epstein, S. (1999). The relation of rational and experiential information processing styles to personality, basic beliefs, and the ratio-bias phenomenon. *Journal of Personality and Social Psychology*, *76*(6), 972-987. https://doi.org/10.1037/0022-3514.76.6.972

Peer, E., Brandimarte, L., Samat, S., & Acquisti, A. (2017). Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, *70*, 153-163. https://doi.org/10.1016/j.jesp.2017.01.006

Pennycook, G., De Neys, W., Evans, J. S. B., Stanovich, K. E., & Thompson, V. A. (2018). The mythical dual-process typology. *Trends in Cognitive Sciences*, *22*(8), 667-668. https://doi.org/10.1016/j.tics.2018.04.008

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Raoelison, M., & De Neys, W. (2019). Do we de-bias ourselves?: The impact of repeated presentation on the bat-and-ball problem. *Judgment and Decision Making*, 14 (2), 170 - 178.

Robbins, M. (2012, March 6). *"Scientists say..."*. The Guardian. https://www.theguardian.com/science/the-lay-scientist/2012/mar/06/2

Rosseel Y (2012). "lavaan: An R Package for Structural Equation Modeling." *Journal of Statistical Software*, 48(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on generality (COG): A proposed addition to all empirical papers. *Perspectives on Psychological Science*, *12*(6), 1123-1128. https://doi.org/10.1177/1745691617708630

Stieger, S., & Reips, U. D. (2016). A limitation of the Cognitive Reflection Test: familiarity. *PeerJ*, *4*, e2395. https://doi.org/10.7717/peerj.2395

Thompson, V. A., Prowse Turner, J. A., & Pennycook, G. (2011). Intuition, reason, and metacognition. *Cognitive Psychology*, *63*(3), 107-140. https://doi.org/10.1016/j.cogpsych.2011.06.001

Toplak, M. E., West, R. F., & Stanovich, K. E. (2011). The Cognitive Reflection Test as a predictor of performance on heuristics-and-biases tasks. *Memory & Cognition*, *39*(7), 1275-1289. https://doi.org/10.3758/s13421-011-0104-1

Toplak, M. E., West, R. F., & Stanovich, K. E. (2014). Assessing miserly information processing: An expansion of the Cognitive Reflection Test. *Thinking & Reasoning*, *20*(2), 147-168. https://doi.org/10.1080/13546783.2013.844729

van der Linden, S. L., Leiserowitz, A. A., Feinberg, G. D., & Maibach, E. W. (2015). The scientific consensus on climate change as a gateway belief: Experimental evidence. *PloS one*, *10*(2), e0118489. https://doi.org/10.1371/journal.pone.0118489

Westfall, J., Kenny, D. A., & Judd, C. M. (2014). Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *Journal of Experimental Psychology: General*, *143*(5), 2020-2045. https://doi.org/10.1037/xge0000014

Contributions

Contributed to conception and design: MH, HB

Contributed to acquisition of data: MH, HB

Contributed to analysis and interpretation of data: MH, HB, TP

Drafted and/or revised the article: MH, HB, TP

Approved the submitted version for publication: MH, HB, TP

Funding

This work was supported by a Leverhulme Trust Research Project Grant (RPG-2019-158) awarded to the first author.

Competing Interests

The authors have no competing interests.

Data accessibility

All the stimuli, presentation materials, participant data, and analysis scripts can be found on the Open Science Framework.

Study 1 https://osf.io/n7pj8/

Study 2 https://osf.io/2vh9w/