Caveats in science-based news stories communicate caution without lowering interest.

Lewis Bott[1], Luke Bratton[1], Bianca Diaconu[1], Rachel C. Adams[1], Aimeé Challenger[1], Jacky Boivin[1], Andrew Williams[2], Petroc Sumner[1].

[1]School of Psychology, Cardiff University, CF10 3AT, UK.

[2]School of Journalism, Media & Cultural Studies, Cardiff University, CF10 3NB, UK.

Address for correspondence:

Lewis Bott, PhD

School of Psychology

Cardiff University

Tower Building

Park Place

Cardiff

CF10 3AT UK

BottLA@Cardiff.ac.uk

Tel: +44 (0)29 2087 4938

Abstract

Science stories in the media are strongly linked to changes in health-related behavior. Science writers (including journalists, press officers, and researchers) must therefore frame their stories to communicate scientific caution without disrupting coherence and disengaging the reader. In this study we investigate whether caveats ("Further research is needed to validate the results") satisfy this dual requirement. In four experiments participants read news reports with and without caveats. In Experiments 1 to 3, participants judged how cautious or confident researchers were, and how interesting or comprehensible they found the reports. News reports with caveats were judged as more cautious that those without, but levels of reader interest and comprehensibility were unaffected. In a fourth, we created a mock newsroom and recruited journalism students to make judgements about which press releases should be published. Here, neither caveats nor the introduction of qualifying expressions in headlines had an effect on judgements of newsworthiness, consistent with Experiments 1 to 3. The reasons participants gave for rejecting a press release rarely referred to the caveat. Our results therefore suggest that science writers should include caveats in news reporting and that they can do so without fear of disengaging their readers or losing news uptake.

Keywords: caveats; science communication; media; coherence; pragmatics

Public Significance Statement

We show that caveats in science stories communicate scientific caution, yet do not cause noticeably lower levels of reader interest in the story. Our results therefore suggest that science writers should include caveats in news reporting and that they can do so without fear of disengaging their readers or losing news uptake.

Caveats in science-based news stories communicate caution without lowering interest.

Science stories in the media are strongly linked to changes in health-related behavior and knowledge (Cram, Fendrick, Inadomi, Cowen, Carpenter, & Vijan S., 2003; Lewis, 2001; Grilli, Ramsay, Minozzi, 2002; Matthews et al., 2016; Phillips, Kanter, Bednarcyk and Tasdad, 1991; Stryker, Moriarty & Jensen, 2008). While there are positive effects of the media (e.g. increasing cancer screening rates, Cram et al.), the response of the public can be exaggerated relative to the intentions of the scientific authors or advice from practitioners. For example, following media coverage of the side effects of statins, the number of patients who stopped taking their statins medication increased (Matthews et al.), despite the advice of the medical profession and UK National Institute for Health and Care Excellence (NICE, July 2014). The reasons behind the public's response to the media are complex and multifaceted (see Lewis; Grilli, Ramsay & Minozzi) but one suggestion (e.g. Schwitzer, 2008; Sumner et al., 2014, Sumner et al., 2016; Adams et al., 2019) is that science writers (including journalists, press officers and researchers) mislead the public by oversimplifying or exaggerating scientific claims.

In this paper we investigate one common recommendation for responsible scientific writing: the inclusion of caveats, or limitations, relating to the source study. Caveats are whole sentences or blocks of text intended by the writer to express caution, tentativeness and lower certainty in the convictions of the reader (Crismore & Vande Kopple, 1999). Consider the caveat below (from press release 02-11-013, Sumner et al., 2014, describing Gilbert et al., 2011).

> "There's still a way to go before we fully understand the link between a person's vitamin D levels and their risk of cancer. There is consistent evidence that bowel cancer is less common in people with high levels of vitamin D. But we still need more

research to clarify whether vitamin D directly prevents bowel cancer or if people with higher levels are generally healthier."

The caveat emphasizes the need for further research, as is typical of many[1], and also presents an alternative explanation for the findings, "… if people with higher levels are generally healthier." The presumed intention is to inject caution into perceptions of the causal link between vitamin D and cancer.

However, how effective is the caveat? Do readers understand the associative explanation between vitamin D and bowel cancer? And even if they do, does this affect how interesting they find the story? Caveats that are ineffective are, of course, no use to the reader, but caveats that are effective but disengaging are of limited use to the writer. Here we test these two perspectives in parallel by manipulating the presence of caveats in news stories and measuring the resulting change in perceptions of caution and interest.

Caveats in the media

The use of caveats has been advised in numerous best practice guides for science writers (Academy of Medical Sciences, 2017; Science Media Centre, 2012; Straight Statistics and Sense about Science, 2010; Schwitzer, 2008) and calls for straightforward science reporting (Schwartz & Woloshin, 2004). For example, the Academy of Medical Sciences, a highly respected organization representing medical scientists, recommends that writers, "Include research caveats and clear descriptions of context." Services such as HealthNewsReview.org are also popular (receiving ~ 5000 hits per day; source: www.semrush.com), and in common with the resources cited above, encourage the declaration of limitations in press releases and news stories ("when reading about a new intervention or diagnostic tool, people should come away with a sense of how rigorous the

[1] Appendix 1 shows a list of all caveats found in health-related press releases from the Russell Group of UK Universities in 2011 (retrieved from the press release corpus collected by Sumner et al., 2014).

evidence is for the intervention"). Given the wealth of advice advocating caveats one would expect that a high proportion of press releases and media stories would contain them. The reality, however, is that very few of them do (Ionniadis, 2007; Hyland, 1996; Pulluchia, 1997; Sumner et al., 2104; Sumner et al., 2016; Wang, Boland & Grey, 2015). For example, correlational research might warrant a caveat against causal inference, yet out of all biomedical and health press releases generated by major UK universities and journals in 2011, only 14% (62 out of 428) included a suitable caveat, and of the subsequent news stories, only 18% (107/607) carried the caveat (Sumner et al., 2016).

We suggest that the low rate of caveats is due, at least in part, to two perceptions held by science writers. First, writers might perceive that readers find caveats difficult to understand. They consequently omit them on the grounds that they would be ineffective. Arguments in support of this view are that caveats in scientific articles often involve technical information (e.g. "The procedure only identifies methylated nucleotides located within the recognition sequences…"; see Hyland, 1996), which lay readers would indeed struggle with. However, caveats can be simplified (see Appendix 1) and it may not be necessary to fully understand the caveat in order to appreciate the communicative intentions of the writer. Expressions such as, "further research," "limitations," and "alternative explanations," all suggest caution even if what follows is not understood.

Second, writers might perceive that caveats disengage the reader. A caveat that makes a story uninteresting would reduce the likelihood of an editor choosing to print the story, and this concern might reduce the likelihood of professional writers (including scientists) using them (similar points have been made by Collins, 1987, Olausson, 2009, and Schäfer, 2011). There is no direct evidence to assess whether caveats do cause disengagement, but related indirect evidence is provided by Sumner et al. (2014), Sumner et al. (2016), and Schat, Bossema, Numans and Smeets (2017). These researchers assessed whether caveats were associated with lower news output. While Sumner and colleagues found no significant

relationship, suggesting that caveats do not make press releases less interesting (to editors), Schat et al. found a small negative relationship, suggesting that they might. All three studies were correlational and not experimental, however. Thus, it might be the case that caveats reduce uptake but this effect was offset by, for example, a tendency for caveats to be included when the story was more exciting, or indeed that caveats boost uptake but are only included when the study is weak.

In this study we directly test the effect of caveats on perceived researcher caution and on interest. Our goal was to gather evidence about the usefulness of caveats as a technique for conveying caution. Caveats are not the only linguistic method for conveying caution - lexical hedges, such as *may* or *could* ("Vitamin D may reduce cancer") fulfill similar pragmatic goals – but caveats are the most intuitive and are often recommended by science guides. We therefore focus on caveats and briefly consider alternative methods at the end of the paper. In the next section we discuss the cognitive factors underlying how readers understand news stories and caveats in order to provide a psychological framework for our study.

Understanding and evaluating science-based news

In order to comprehend a news story, readers must construct a representation or mental model of the story's content (e.g. Johnson-Laird, 1980; Kintsch, 1988). This would include the main protagonists (e.g. the scientists), the themes (what the story is about, e.g. vitamin D and cancer), the actions (what happened, e.g. a correlational study was conducted) and the causality relations (e.g. the causal link between vitamin D and cancer), amongst other elements. Readers would also have to make inferences about relevant, unstated information (e.g. that the protagonists were scientists; Graesser, Singer & Trabasso, 1994). Finally, they would have to decide whether the resulting representation is credible, and if so, update their prior beliefs (e.g. add propositions that Vitamin D is important to their health; see e.g. Oaksford & Chater, 2007; Corner & Hahn, 2009; Hahn, Harris & Corner, 2009). The inclusion of caveats could alter any or all of these processes.

Caveats must alter the representation of the news story if they are to be useful. They might reduce the number of causal links in the representation (e.g. the Vitamin D example above warns against a causal link between Vitamin D and cancer) or lower the belief in some of the propositions. If the individual elements of a story do not connect sensibly and causally, or belief in the propositions is low, however, the resulting mental model becomes less coherent ("why are they telling me about vitamin D and cancer if there is not a causal relationship between the two?"). Readers would consequently find the text more difficult to understand and, presumably, less appealing as a result. Work in other areas also suggests that less coherent texts are less persuasive, in that readers are less likely to alter their beliefs or to act on the content of the text (Corner & Hahn, 2009; Pennington & Hastie, 1988, 1992, 1993; Lagnado, 2011; Thagard, 2000). For example, Pennington and Hastie (1988) showed that when legal evidence was presented in an order that made a witness story more difficult to construct (less coherent), there were fewer judicial verdicts in favour of the less coherent story. Similarly, Corner and Hahn found that arguments about science were rated less strongly when they contained "mixed evidence" (i.e. they lacked evidential coherence) than when they contained either entirely confirmatory or entirely disconfirmatory evidence. Caveats might therefore prevent readers from altering their beliefs, exactly as the writer intended. But if they achieve this by making the news story less coherent, caveats would come at a cost of making the story less appealing. For the same reason, less coherent stories would be judged less newsworthy by journalists.

Just because caveats are initially encoded, however, there is no guarantee that they will remain in the text representation and subsequently influence belief revision. Caveats typically occur at the end of the text and qualify or correct an inference that might already have been derived (e.g. that there is a causal relationship between two variables). However, evidence from multiple sources demonstrates that correcting pre-existing information is difficult (see Lewandowsky, Ecker, Seifert, Schwartz, and Cook, 2012, for a review). For example, in

studies investigating the retraction of information, participants are presented with a written story in which some key information is included early in the text (e.g. "the cause of the fire was found to be gas cylinders in a nearby closet"), much like a news story, and then retracted either immediately or within a few sentences (e.g. "the closet was found to be empty"). The typical finding is that participants continue to reference the retracted information when asked about the cause of the fire (e.g. Ecker, Lewandowsky, Swire & Chang, 2011; Wilkes & Leatherbarrow, 1988; Johnson & Seifert, 1994). Work on narrative contexts show similar effects (Guéraud, Harmon & Perrachi, 2005; O'Brien, Rizella, Albrecht, and Halleran, 1998; Rapp & Kendeou, 2007). For example, O'Brian et al. showed that readers' expectations of events was influenced by earlier character descriptions, even when those descriptions were later qualified. Finally, there are many real-world examples of persistent beliefs in the face of correcting evidence, such as the continued belief in the link between autism and the MMR vaccine long after the initial claims were discredited (Hargreaves, Lewis & Spears, 2003), or the belief that Barack Obama was born in Kenya (Travis, 2010) and therefore not eligible to become President of the United States, despite the presentation of his birth certificate. If caveats behave like retractions, they would have a limited effect on the lasting representation of the story.

In summary, caveats present a possible solution to the problem of under-qualified news stories. However, the psychological literature predicts that either they would be ineffective at changing representations – much like retractions – or that they would lower coherence and thus appeal/newsworthiness. These predictions are aligned with the likely reasons caveats are rarely included by writers (regardless of whether they have read the literature reviewed above): caveats are either perceived to be ineffective at communicating caution, or perceived to reduce the appeal of news stories. To test these predictions we conducted four experiments that directly investigate perceptions of caveats in science-based news stories.

Experiment 1

The aim of Experiment 1 was to test what sorts of caveat might be effective and whether they alter how interesting news stories appear. Participants read science-based news stories with and without caveats. There were three conditions. In the general caveat condition, the stories contained a caveat referring to the need for further research (see Table 1). In the specific caveat condition, the stories also included a more detailed sentence about why the caveat was needed, and in the no-caveat condition, the stories contained the same basic material but no caveat. Immediately after reading the stories, participants were asked questions about their perceptions of the researchers and their understanding of the story.

In Experiment 1a, participants were asked either "How confident do you think the researchers are in their findings?" or "How interesting did you find the story?" Researcher confidence measured the uncertainty conveyed by the caveat, and so captured whether the caveat was effectively communicating the concerns of the writer, and interest measured the general appeal, testing concerns about whether caveats made the story less interesting. In Experiment 1b, we asked different questions using the same materials, as we discuss later.

Table 1. Example materials for Experiments 1 and 2.

Headline	Extra testosterone reduces empathy
Story	The researchers found that administration of testosterone led to a significant reduction in mind reading. Given that people with autism have difficulties in mind reading, and that autism affects males more often than females, the study provides further support for the 'extreme male brain' theory of autism. A new study from Utrecht and Cambridge Universities has for the first time found that an administration of testosterone under the tongue in volunteers reduces a person's ability to 'mind read' which is an indication of empathy. The researchers used the 'Reading the Mind in the Eyes' task as the test of mind reading, which tests how well someone can infer what a person is thinking or feeling from photographs of facial expressions from around the eyes. Mind reading is one aspect of empathy, a skill that shows significant sex differences in favour of females.
General caveat (Experiment 1)	However, the scientists emphasize the need for more research before generalizing their results.
Specific caveat (Experiment 1)	However, the scientists emphasize the need for more research before generalizing their results. They note that the task was quite simple and that there are other components of mind reading that were not captured by their study.
Nonsense caveat (Experiment 2)	One caveat to the findings, however, is that testosterone also led to increased activation in the fusiform face area, part of the fusiform gyrus, an area of increasing debate about its function.
Exciting statement (Experiment 2)	In an exciting development, researchers also found that testosterone led to increased activation in the fusiform face area, part of the fusiform gyrus, an area of increasing debate about its function.

Note. All participants read the headline and story. In Experiment 1, participants read either the story with no caveat, with a general caveat, or with a specific caveat. In Experiment 2, they read the story with no caveat, with a nonsense caveat or with an exciting statement.

Method

Participants. One hundred and sixty-four participants were recruited via the online platform Prolific. Participants were randomly allocated to the confidence question, N= 83 (48 females, 35 males; aged, 18-66, *M* = 38) or the interest question, N = 81 (54 females, 27 males; aged, 18-71, *M* = 38), and to one of six counterbalancing lists, N = 28, 28, 27, 29, 27, 26 per list. The average completion time was 9 minutes and participants were paid £1.30 for participation. All experiments were approved by the School of Psychology Research Ethics Committee, Cardiff University.

Design and materials. The items were shortened versions of press releases taken from Sumner et al. (2014) (see Appendix 2). We described them as "news stories" to participants however, and use this label throughout Experiments 1 to 3. We used press releases rather genuine news stories because press releases had fewer irrelevant details, such as news outlet specifics ("as our Sun correspondence says…"). Press releases are different to news stories, in that the former are written by press officers at universities, journals or in industry, and the latter by journalists, but there is considerable overlap between the two (Lewis, Williams, & Franklin, 2008; Schwartz, Woloshin, and Andrews, 2012; Sumner et al. 2014; 2016; Yavchitz A et al., 2012) and many press releases are written in such way that they can be inserted directly into a news publication without alteration. Participants would therefore have been unlikely to notice a discrepancy between the label "news story" and the press release content. Similarly, we expect the results from our studies to generalise from press releases to news stories.

There were nine stories, each with a bespoke caveat. All general caveats emphasized the need for more research. All specific caveats added a further sentence referring to alternative explanations for the results (see Table 1 for an example). The caveat always appeared as the final text block in the trial. The no-caveat condition contained the basic story without any caveat. Note that length of story (e.g. number of words) was confounded with

caveat condition (stories in the general caveat condition had more words than those in the no-caveat condition). We return to this point in the General Discussion but we note that this situation mirrors the environment in that stories with caveats will be longer than stories without caveats, all else being equal.

Participants saw all nine stories, three for each caveat condition. Stories were counterbalanced using a pseudo-latin square design across three counterbalancing lists. Lists were constructed in such a way that each participant saw three stories from each caveat condition and no participant saw the same story twice. Across lists, all stories appeared in each caveat condition. Stories were presented in a random sequence.

Participants answered either the confidence or the interest question. A between subject design was adopted to avoid carry over effects from one question to the other (e.g. participants copying their response from the caveat to the interest question). The question was presented immediately after each story.

Participants were also asked a multiple-choice memory question. This was included so that trials in which the participant had been inattentive could be removed.

Procedure. Each story was divided into five blocks of text and each block presented on a different screen. The title was one block and the remaining text was divided into another four. There were the same number of blocks for each condition, which meant that the final block for the caveat conditions was slightly longer than for the no-caveat condition. Participants read the title and then advanced onto the next and subsequent blocks with a mouse click. After each story was completed, the confidence or interest question was presented on a different page to the story text. Responses were made using a mouse to select a point on a visual analogue scale from 0 to 100. The mouse started at the zero point for each trial. Zero was labelled as “Not confident at all” / “Not interesting at all” and 100 as “Extremely confident” / “Extremely interesting.” Participants then answered a multiple-choice memory question also with the mouse.

The experiment was conducted online.

Results

Data screening. One-hundred-and-thirty-seven trials (out of 1476; 9%) were removed from the analysis because the memory question was answered incorrectly (confidence: 28, 19, 25; interest: 18, 25, 22; for no, general, and specific caveats respectively). These were evenly distributed across caveat type for each question χ^2's (2) < 1.75, p's > 0.42.

Analysis overview. In all experiments the data were analysed as mixed models with the lme4 package in R (Bates, Maechler, Bolker, & Walker, 2015). *p*-values were computed with the Kenward-Roger and Satterthwaite approximations to degrees of freedom, implemented in the lmerTest package (Kuznetsova, Brockhoff, Christensen, 2016).

We used Bayes Factors (BF) to interpret the evidential value of relevant nonsignificant findings (Dienes, 2011, 2014; Rouder, Speckman, Sun, Morey, & Iverson, 2009). The BF is the likelihood of the data under the alternative hypothesis relative to the data under the null hypothesis. BF > 3 is interpreted as "substantial" evidence for the alternative hypothesis and BF < 0.33 "substantial" evidence for the null hypothesis (Dienes, 2011, 2014; Schonbrodt, Wagenmakers, Zehetkeitner, & Perugini, 2017). We assumed the default JZS prior (r = 0.707; Rouder et al., 2009), which is a non-informative objective prior that minimises assumptions regarding expected effect size. Bayes Factors were computed using the BayesFactors package in R (Morey & Rouder, 2015).

All data and analysis is available online at https://osf.io/3kqtp/.

Model specification. Responses to the confidence question and the interest question were analysed using separate models. For both, the model structure was as follows (expressed in R pseudo code):

```
score~caveat*group+(1+caveat|subject)+(1+caveat+group|item)
```

with `caveat` (general, specific or no-caveat) and `group` (counterbalancing group; 1, 2 or 3) as fixed factors, and `subject` (participant) and `item` (news story) as random factors.

We initially included age and its interactions as a continuous predictor in the model but found it did not significantly reduce variability for either dependent measure, χ^2's (9) < 9.67, p's > 0.38. We therefore omitted age from the model.

Analysis. Perceptions of researcher confidence were influenced by caveats (see Figure 1) such that both general and specific caveat conditions were significantly lower than the no-caveat condition: β = -13.59, SE = 2.19, t = 6.21, p < .001, d = 0.90; β = -18.93, SE = 2.29, t = 8.27, p < .001, d = 1.18, respectively. Thus even simple, generic caveats are effective at communicating caution. The specific caveat condition was also lower than the general caveat condition, β = -5.35, SE = 2.59, t = 2.058, p = .022, d = 0.36.

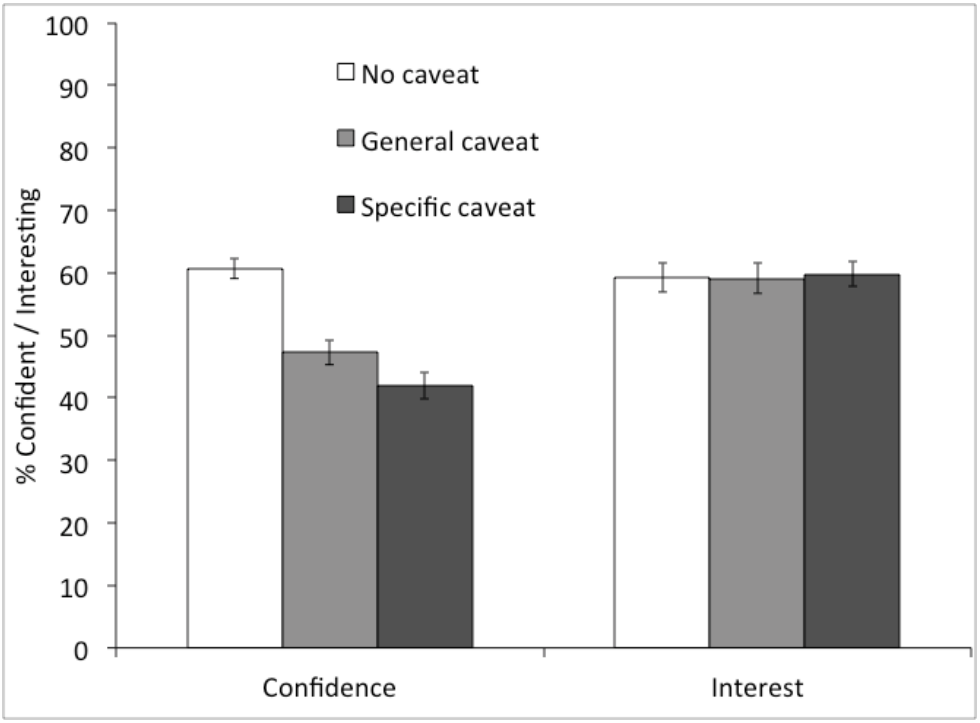

Figure 1. Mean confidence and interest scores for Experiment 1. Error bars are the standard error of the mean.

In contrast, levels of interest were unaffected. Neither general nor specific caveat conditions were significantly lower than the no-caveat condition, β = -0.064, SE = 2.02, t = 0.033, p = 0.98, BF = 0.11, d = 0.0078; β = 0.38, SE = 1.90, t = 0.20, p = 0.81, d = 0.027, BF = 0.12, and nor did the specific caveat condition differ from the general caveat condition, β =

0.45, SE = 1.95, t = 0.23, p = 0.82, d = 0.043, BF = 0.11. Note that the results were not due to general insensitivity since the Bayes Factor analyses showed that the probability of the data under the null hypothesis (that there is no effect of the caveat) was over eight times that under the alternative hypothesis (that there is an effect of the caveat).

Discussion

We observed large effects of caveats on perceptions of researcher confidence. Participants rated stories with specific or general caveats as being less confident than those with caveats, and those with specific caveats as being less confident than those with general caveats. Conversely, there were no effects of caveats on levels of interest. Together, the data illustrate that caveats can lower confidence without lowering levels of interest.

Experiment 1b

In Experiments 1b and 1c we used the same materials and basic design as Experiment 1a but asked different questions. In Experiment 1b, we asked readers about researcher *caution* ("Please rate how cautious the researchers were in their conclusions") rather than researcher *confidence* (Experiment 1a). Caution and confidence can sometimes diverge (e.g., researchers could be confident about their associative data but cautious about inferring causality) and we wanted to make sure that our conclusions would generalise over these two concepts. Caution also adds methodological generality to our claims from Experiment 1a because caution implies a reversed scale relative to confidence. Readers who are sensitive to the caveat should provide low caution scores for stories with specific caveats and high caution scores for stories with no caveats, whereas the reverse is true for confidence.

We also collected education levels of participants at the end of the study. Although previous research has found no effects of education on understanding of science news stories (Adams et al. 2016; Bleske-Rechek et al., 2015; Jensen, Pokharel, Scherr, King, Brown & Jones, 2017; Norris et al., 2003; Norris & Phillips, 1994), it remains possible that caveats are influenced by education in a way that general science news comprehension is not.

Method

Participants. Ninety-nine participants were recruited via the online platform Prolific (73 females, 25 males, 1 non-binary; aged 18-65, *M* = 33). Participants were randomly allocated to each counterbalancing list within question type, N = 33 in each. Potential participants were ineligible if they had completed previous experiments in the study. The average completion time was 9 minutes and participants were paid £1.30 for participation.

Participants had the following education profile: GCSE/O-level (UK national qualifications taken at around 15 years old) or other pre-16 year old qualifications, N = 10; A-level (UK national qualifications taken at around 17 years old, used for university entry) or other 16-18 year old qualifications, N = 28; Undergraduate degree, N = 38; Postgraduate degree, N = 20; Other, N =4.

Design, materials and procedure. The design, materials and procedure were identical to Experiment 1a, except for the different question asked and the education level requested at the end of the experiment. The education question asked participants to indicate their maximum level of education from a list of common UK qualifications (as seen in the Participants section).

Results

Data screening. Ninety-three incorrect memory responses were removed (34, 27, 32 from no caveat, general caveat and specific caveat respectively). There were no significant differences across conditions, χ^2 (2) = 0.84, p = .66. For analyses regarding levels of education, we removed 3 participants who answered "Other" to the education question. This is because we could not determine what education level they were.

Model specification. We started with the same model as Experiment 1a and tested whether to include age and education using a χ^2. Education significantly reduced variability and so was included in the model, χ^2 (9) = 18.05, p = 0.035, but age did not, χ^2 (1) = 0.45, p = .50. The model was therefore:

```
score~caveat*education+(1+caveat|subject)+(1+caveat+education|item)
```

with `caveat` (general, specific or no-caveat) and `education` (GCSE, A-level, undergraduate, or postgraduate) as fixed factors, and `subject` (participant) and `item` (news story) as random factors. We removed counterbalancing group from the model to aid convergence.

Analysis. There were large and robust effects of the caveat on caution (see Figure 2). Ratings were in the opposite direction to those of confidence, as expected, with significantly lower scores for the no caveat condition $M = 41$ (SD = 25) compared to the general $M = 57$ (SD = 21) and the specific conditions $M = 62$ (SD = 20), $\beta = 12.97$, SE = 2.89, $t = 4.45$, $p < 0.001$, $d = 0.72$; $\beta = 17.44$, SE = 2.93, $t = 5.95$, $p < .001$, $d = 0.83$, and significantly lower scores for the general than the specific condition, $\beta = 4.47$, S.E = 1.82, $t = 2.46$, $p = 0.027$, $d = 0.30$.

Including education and its interaction with caveat significantly reduced variability compared to a model with neither effect ($\chi^2(9) = 18.05$, $p = 0.035$; see Model Specification). However, the main effect of education alone was not significantly different to a model without education, $\chi^2(1) = 2.19$, $p = 0.14$, and the addition of the interaction term only approached significance, $\chi^2(6) = 11.60$, $p = 0.072$. There is thus no evidence that the effect of caveats differs across levels of education.

Discussion

Experiment 1b demonstrates that the effects of caveats are robust across different questions. When we asked participants about how cautious the researchers were, stories with general or specific caveats were rated as more cautious than those with no caveats, and stories with specific caveats were rated as more cautious than those with general caveats. These findings mirror those we observed in Experiment 1a when we asked about the confidence of researchers. Furthermore, levels of education did not interact with the effects of the caveat.

Experiment 1c

In Experiment 1c, we used the same materials and design as Experiments 1b and 1c but asked about comprehension ("How easy was it to understand the story?"). One way in which caveats might be effective is that they remove important links between variables in the representation of the story. For example, by denying evidence of a causal relationship between two variables, a caveat might leave the reader with an unsatisfactory knowledge gap about why an association arises at all. This would lead to a less coherent representation and consequently a drop in comprehensibility (less coherent stories are more difficult to understand, see Gernsbacher, 1990). One might expect this to be reflected in interest scores, as measured in Experiment 1a, since texts that are difficult to understand are presumably uninteresting. However, it may be that readers focus on general content to judge interest and ignore comprehension, whereas when directly asked about understanding, they judge the coherence of the text. As in Experiment 1b, we also asked about education levels.

Method

Participants. Ninety-nine participants were recruited via the online platform Prolific. (64 females, 35 males; aged 18-70, *M* = 36). Participants were randomly allocated to each counterbalancing list within question type, N = 33 in each. Potential participants were ineligible if they had completed previous experiments in the study. The average completion time was 9 minutes and participants were paid £1.30 for participation.

Participants had the following education profile: GCSE/O-level or other pre-16 year old qualifications, N = 11; A-level or other 16-18 year old qualifications, N = 33; Undergraduate degree, N = 38; Postgraduate degree, N = 16; Other, N =1.

Design, materials and procedure. The design, materials and procedure were identical to Experiment 1a, except for the different question asked and the collection of educational levels at the end of the experiment.

Results

Data screening. Seventy-one responses were removed (17, 24, 30 from the caveat conditions respectively). There were no significant differences across conditions, χ^2 (2) = 3.58, p = .17. For analyses regarding levels of education, we removed 2 participants who answered "Other" to the education question.

Model specification. As in Experiment 1b, we started with the model from Experiment 1a and tested whether age or education significantly reduced variability. Neither did, however, $\chi^2(1) = 0.17$, $p = 0.68$, χ^2 (9) = 3.46, $p = 0.94$. We therefore used the same mixed model as Experiment 1a.

Analysis. Comprehensibility was unaffected by the caveat. Neither general caveat, M = 81 (SD = 20), nor specific caveat M = 81 (SD = 19), were significantly different from no caveat, M = 82 (SD = 19), β = -1.53, S.E = 1.38, t =1.11, p = 0.29, d = 0.15, BF = 0.17; β = -1.94, S.E = 1.32, t =1.47, p = 0.17, d = 0.20, BF = 0.17; and nor was general caveat different to specific caveat, β = 0.41, S.E = 1.13, t =0.37, p = 0.76, d = 0.063, BF = 0.096.

Discussion

In Experiment 1c we asked about comprehension. Here, as with levels of interest in Experiment 1a, we found no effects of the caveat. Caveats do not make news stories more difficult to understand and, consequently, they are unlikely to make the text less coherent.

One potential concern is that we measured perceived comprehension of the text rather than actual comprehension. There is evidence from the metacognition literature that the two can diverge (e.g. Bower & Winchester, 1970; Vesonder & Ross, 1985) so that, for example, participants in our task might have perceived an equal level of comprehension across conditions but would nonetheless differ in what they could recall. However, we found no evidence of recall differences across conditions when we analysed the data screening questions, either here or in any of the experiments, so the discrepancy between the two cannot be large for our stimuli. Differences in coherence (the focus of the experiment) would also not necessarily manifest themselves in objective measures of comprehension such as

recall, since coherence is not underpinned with factual information but instead with the relationship between facts.

Experiment 2

The caveats in Experiment 1 were all simple and easy to understand, by design. However, writers may not always be able to construct such simple caveats, or they may feel that the simplicity is misleading and does not reflect the complexity of their concerns. In these circumstances would it be better not to include a caveat, or to include one knowing it might be too complex for many people to understand?

In Experiment 2 we tested this by using nonsense caveats. These were statements that did not make sense as a caveat, in that they did not place limitations on the research and were not relevant to the conclusions of the study (see Table 1). What defined them as a caveat was the introductory clause. In the caveat condition, this explicitly stated that subsequent material described a limitation, such as, "One limitation of the research was that…" or "However, one caveat to our findings…" We reasoned that this would mimic the situation in which a reader would fail to understand a caveat used in a news story. There were two control conditions. The first was the basic story with no caveat. The second was the same main clause material as the caveat but with an introductory clause designed to make the subsequent material an exciting consequence of the findings, such as, "In an exciting new development…." (see Table 1). The *exciting* condition was included so that we could test whether the mere presence of the nonsense material influenced judgements, or whether an explicit statement of caveat was necessary.

In Experiment 2a we asked participants how confident they perceived the researchers to be. This tests the main goal of Experiment 2, which was to establish the effectiveness of caveats when readers find them difficult to understand. In a separate Experiment, 2b, we asked how interesting readers found the stories. We wanted confirm and extent the conclusions from Experiment 1 about how much caveats alter levels of interest. We present

the methods and results of Experiments 2a and 2b together because they used the same materials, experimental design and analysis.

Method

Participants. All participants were recruited via Prolific. We completed Experiment 2a (confidence) first, N = 99 (65 females, 34 males; aged, 17-73, *M* = 37) and then Experiment 2b (interest), N = 99 (50 females, 49 males; aged, 16-75, *M* = 31). Participants were randomly allocated to one of three counterbalancing lists within each experiment, N = 33 per list, subject to the constraint that there was an equal number of participants in each list. No participants had completed Experiment 1 and participants were prevented from completing both Experiments 2a and 2b. The average completion time was 10 minutes and participants were paid £1.50 for participation.

Design and materials. The same nine stories were used as in Experiment 1. These were combined with three sorts of caveat introductory clause (see Appendix 3): "One caveat to the findings…"; "However, one limitation of the…", and "The scientists warn that…," and three sorts of exciting introductory clause: "Interestingly, …"; "The scientists were intrigued to find…"; and "In an exciting development…" Participants saw all nine stories, three for each caveat condition. Stories were counterbalanced across three counterbalancing lists in such a way that each participant saw three stories from each caveat condition but no participant saw the same story twice, and across lists, all stories appeared in each caveat condition. Furthermore, all participants read each introductory clause and no introductory clause more than once. Stories were presented in a random sequence for each participant.

The confidence and interest questions were the same as Experiment 1.

Procedure. The procedure was the same as Experiment 1.

Results and discussion

Data screening. Two-hundred-and-eight trials (out of 1782; 12%) were removed because the memory question was incorrect (confidence: 28, 30, 31; interest: 41, 37, 41; in

the no, nonsense, and exciting conditions respectively). These were distributed evenly across caveat conditions in each question, χ^2's (2) < 0.27, *p*'s = 0.87.

Model Specification. The data were analysed using mixed models and Bayes Factors, as in Experiment 1. The model structure was as follows:

```
score ~ caveat*group+(1+caveat|subject)+(1+caveat+group|item)
```

where `caveat` (no-caveat, caveat, or exciting) and `group` (counterbalancing group; 1, 2 or 3) were fixed factors, and `subject` (participant) and `item` (news story) were random factors. `score` was either confidence (Experiment 2a) or interest (Experiment 2b). We also tested whether age and its interactions were significant predictors of score and found that they were not, χ^2's (9) < 7.54, *p*'s > 0.58. We therefore omitted age from the model.

Analysis. For confidence, the nonsense caveat condition was significantly lower than the no-caveat condition (see Figure 3), β = -6.26, SE = 1.73, *t* = 3.62, *p* = 0.033, *d* = 0.21, even though the caveat was nonsense. The nonsense caveat condition was also significantly lower than the exciting statement condition, β = -7.50, SE = 1.82, *t* = 4.12, *d* = 0.46, *p* = 0.007, suggesting that the effect of the caveat was not due to the nonsense material *per se* but the introductory clause identifying it as a caveat. Further support for this is shown by the absence of a difference between exciting statement and no-caveat conditions, β = 1.24, SE = 2.49, *t* = 0.50, *p* = 0.65, *d* = 0.077, BF = 0.14.

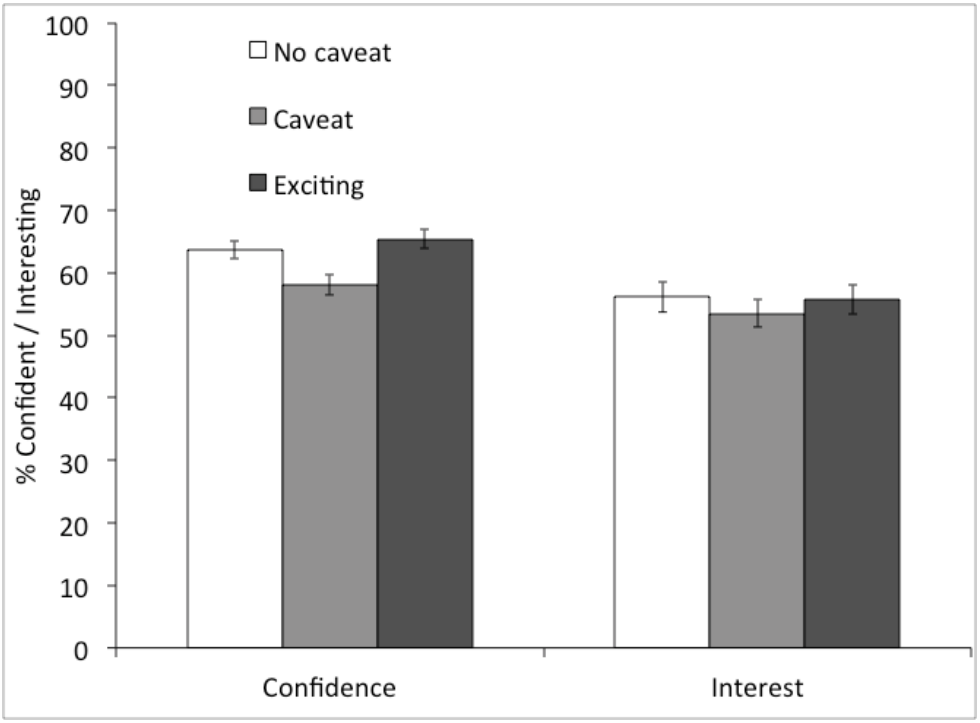

Figure 3. Mean confidence and interest scores for Experiment 2. Error bars refer to the standard error of the mean.

In contrast to the confidence data, there were no effects of the nonsense caveat on interest. The nonsense caveat condition was not lower than the no-caveat condition, β = -2.49, SE = 2.60, t = 1.0, p = 0.38, d = 0.15, BF = 0.15, nor the exciting statement condition, β = -2.27, SE = 2.59, t = 0.88, p = 0.48, d = 0.19, BF = 0.18. There was also no difference between the exciting statement and the no-caveat conditions, β = -0.22, SE = 2.74, t = 0.081, p = 0.94, d = 0.039, BF = 0.10.

Discussion

The results of Experiment 2 mirrored those of Experiment 1. We found robust effects of caveats on judgements of researcher confidence (Experiment 2a) but no effects on levels of interest (Experiment 2b), just as in Experiment 1. The difference was that here, the caveats were nonsense. Experiment 2 therefore suggests that writers should include the caveat even if they fear it would be incomprehensible to most readers. Providing that the caveat is signalled appropriately, it should still be effective. Of course, when the caveats are comprehensible, as in Experiment 1, the effect of the caveat will be stronger than when they are not. This can be

seen by comparing the reduced effect size of the caveat manipulation in this experiment, *d* =0.21, against that of the previous experiment: general *vs* no-caveat *d* = 0.90, specific *vs* no-caveat, *d* = 1.18.

Experiment 3

Experiments 1 and 2 provide good evidence against the claim that caveats alter how interesting participants find news stories. However, it is possible that our design was simply insensitive to changes in interest. For example, participants might have found all the news stories boring or all interesting; they may have been reluctant to change their response from trial to trial; or they may have misunderstood the intentions behind the interest question. In which case caveats might still alter how interesting participants find the stories but we would have not detected the effect. In Experiment 3 we address this criticism by manipulating *a priori* how interesting participants would find the news stories. We selected some news stories that we believed would be *appealing* (e.g. news stories with topics that are relevant to student-aged readers; news stories with clear results and few scientific details) and others that would be *unappealing*. If the judgement of participants coincides with our own, and the methodology is sufficiently sensitive, appealing stories should have higher interest ratings that unappealing stories. We would then have a positive effect of topic on interest against which to compare the potential effect of caveats. A failure to observe effects of topic would suggest that the design was indeed a poor assessment of reader's interest.

We first established that the caveats in the new materials lowered perceived confidence (Experiment 3a), as they did in previous experiments. We then tested whether interest was affected by caveat and news appeal (Experiment 3b).

Method

Participants. All participants were recruited via Prolific. Ninety-eight participants completed Experiment 3a (confidence) (61 females, 37 males; aged 18-25, *M* = 21.68) and 100 Experiment 3b (interest) (62 females, 38 males; aged 18-24, *M* = 21.82). Participants

were randomly allocated to one of four counterbalancing lists within each experiment subject to the constraint that there was an equal number of participants in each list. The age of participants was restricted to < 25 in order that we could better estimate which stories would appeal. No participants completed both experiments nor had any participants completed Experiments 1 or 2. The average completion time was 13 minutes. Participants were paid £1.50.

Design and materials. Participants read and answered questions about four news stories. Two of the stories were appealing and two unappealing, as described below, and two contained caveats and two did not, crossed with the appeal factor. Thus story appeal and caveat were within subject factors. Assignment of caveat to news story was counterbalanced so that all stories appeared equally in the caveat and no-caveat conditions across participants but no participant saw the same story twice. This resulted in two counterbalancing lists per experiment.

Stories were complete press releases containing caveats, chosen from Sumner et al. (2014). The no-caveat condition was formed by removing the caveats from the original. Out of all the press releases with caveats described by Sumner et al. (2014), we selected two of the most appealing and two of the most unappealing according to our own intuitions (see Appendix 4). For the appealing category, we chose stories that had a topic that would appeal to students, and had clear and simple messages that were easily comprehensible. For the unappealing category, we chose less interesting topics with more complex findings and whose overall message was less clear. The caveats themselves covered a wider range of limitations than those used in Experiments 1 and 2.

Procedure. The procedure was similar to Experiment 1. Participants read stories in blocks at their own pace and then answered either a confidence question (Experiment 3a) or an interest question (Experiment 3b). They then completed three memory questions per item.

We asked more memory questions in Experiment 3 than Experiments 1 and 2 because the stories were longer.

Results.

Data screening. Responses to trials in which no memory questions were answered correctly were removed (i.e. 0 out of 3 correct). Thirty-five trials (out of 791; 4%) satisfied this criterion (Experiment 3a: 6, 4, 7, 1; Experiment 3b: 4, 6, 2, 5; [no-caveat, unappealing], [no-caveat, appealing], [caveat, unappealing], [caveat, appealing] respectively). They were evenly distributed over conditions, χ^2's (1) < 0.53, p's > 0.47.

Model specification. Data was analysed using the following mixed model

```
score~caveat*newsapp+(1+caveat+newsapp|subject)
```

where `caveat` (yes, no) and `newsapp` (appealing, unappealing) were fixed factors, and `subject` was a random factor. We removed caveat random slopes because they accounted for a very small amount of variability and were highly correlated with intercepts (see Baayen, Davidson & Bates, 2008). We did not attempt to generalise over the items (i.e. include items as a random factor) because there were only four (two in the appealing and two in unappealing conditions). As in previous experiments, age nor its interactions were significant predictors of score, χ^2's (4) < 1.58, p's > 0.81. We therefore omitted age from the model.

Analysis. As in previous experiments**,** caveats again affected confidence judgements (Experiment 3a), such that caveats lowered confidence (see Figure 4), β = 2.68 SE = 0.71, t = 3.80, d = 0.36, p < .001. Surprisingly, news appeal also altered confidence, β = -2.94 SE = 0.69, t = -4.27, d = 0.40, p < .001, with no significant interaction β = .044, SE = 0.67, t = 0.065, BF = 0.17, p = 0.95. We discuss this further in the Discussion section.

There was a significant effect of news appeal on interest (Experiment 3b), β = 5.15, SE = 0.95, t = 5.44, p < .001, d = 0.54, such that low news appeal caused low interest (see Figure 4). Thus our design is sensitive to changes in the level of interest experienced by participants. As in previous experiments, the presence of a caveat had no effect on interest

ratings, β = 0.25 SE = 1.11, t = 0.23, d = 0.023, p = 0.83, BF = 0.11, nor was there an interaction with news appeal, β = 1.02, SE = 0.87, t = 1.17, BF = 0.19, p = 0.24. Furthermore, the difference between appealing and unappealing content was significantly greater than the difference between caveat and no-caveat content, M = 10.00 (SD = 18.66) *vs* M = 0.17 (SD = 24.17), $t(89)$ = 3.30, p = 0.001, d = 0.35, demonstrating that even if there were a statistically non-detectible effect of the caveat, it is small relative to the effect of other news story content.

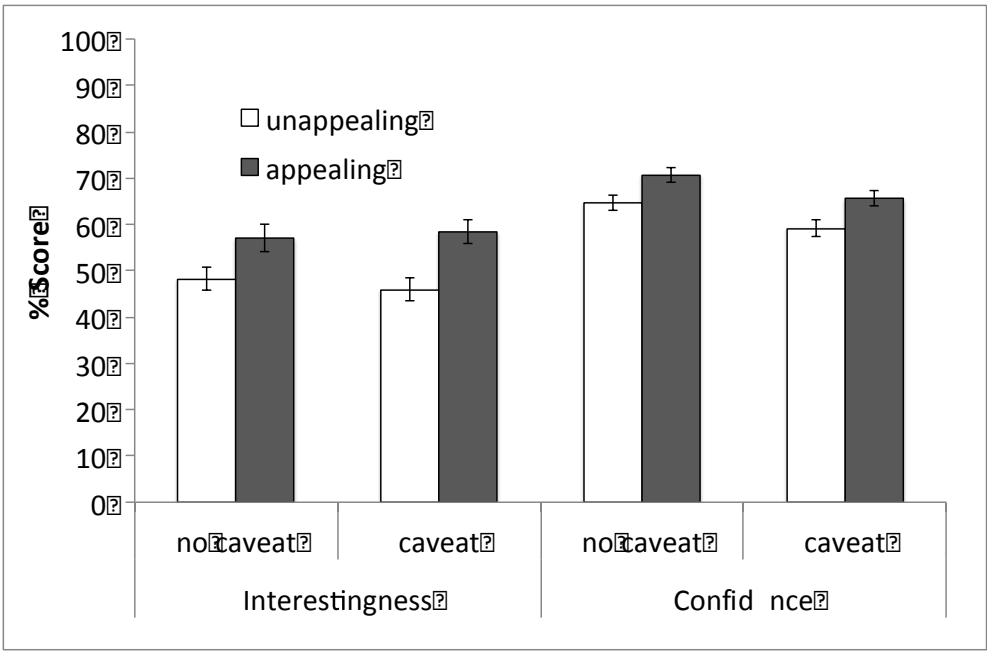

Figure 4. Mean confidence and interest scores for Experiment 3. Error bars refer to the standard error of the mean.

Discussion

The effects of caveats were again found to communicate lower researcher confidence (Experiment 3a) without lowering levels of interest (Experiment 3b). Moreover, we found that participants rated appealing stories as more interesting than unappealing stories. This suggests that the interest question was a sufficiently sensitive and valid measure of engagement for us to have observed the effects of caveats, if indeed they were any.

Surprisingly, we also found that news appeal altered judgements of researcher confidence. When news appeal was high, the researchers were perceived to have high

confidence. Apparently readers perceive simple, clearly stated messages, such as those in the appealing news stories, to come from confident researchers. Caveats might be a part of this, in that caveats add complexity, but they must also be different in some respects because this complexity does not lower interest.

Experiment 3 demonstrates that caveats do not noticeably lower interest, as in previous experiments, but it also puts the size of any non-detectible effect into perspective. We cannot eliminate the possibility that there is an extremely small effect of caveats on interest but even if there is, it must be small relative to the size of content-based factors captured by our appeal manipulation.

Experiment 4

Experiments 1 to 3 found that readers' interest was unaffected by caveats. We therefore argue that science writers can include caveats without fear of losing readership or news uptake. However, the decision to publish a news story is not taken by readers but by journalists and editors, and they may feel differently about caveats than readers. Furthermore, assigning a numerical interest score in the context of an online experiment is quite different from the decision-making process used by media professionals in newsroom environments. Different decision criteria may exist across the two processes. In Experiment 4, we address these concerns by testing a participant sample more familiar with journalism practice than those in Experiments 1 to 3 and by using a paradigm modelled on newsroom procedures.

Participants were postgraduate journalism students taught by media professionals. They were invited to read press releases and accompanying science articles, and to then make a judgement about which press releases were newsworthy[2]. The experiment took place in the "mock newsroom" of the University's journalism school. The newsroom was an attempt to

[2] There was a second part of the experiment in which participants were required to write news stories about two pre-selected press release/article pairs. The aim of this second part was to assess what sort of information journalists chose to include e.g. whether they included study design information. Because the aim was different to that of this paper, we do not report the method and analysis here. A full description is provided in Luke Bratton's PhD thesis (Bratton, in preparation).

emulate the environment in which such judgements would typically be made (Williams & Clifford, 2009). We recorded which press release/article pairs participants chose and the reasons for their choice. There were two manipulations. The first was whether press releases had caveats, similar to Experiments 1 to 3. The second was whether press releases used strong causal language to describe the basic findings of the study (e.g. "Extra testosterone reduces empathy") or whether they used weak language (e.g. "Extra testosterone may reduce empathy"). Qualifying expressions, such as *may* or *might*, signal caution when introduced into headlines (Adams et al., 2016), and so perform a similar function to caveats ("sentence level hedging" *vs* "lexical hedging", Hyland, 1996). For both manipulations the hypotheses were the same. Do journalism students judge press releases with caveats/qualifying expressions as more newsworthy than those without? Do they highlight caveats/qualifying expressions as reasons for their choice?

Method

Participants. Twenty-nine students studying Masters degrees in Journalism were recruited through their course coordinators at Cardiff University and the University of West England (16 females, 13 males; aged 21-29, M = 24.03). Four testing sessions were held, three with separate cohorts of students from the University of West England (27 participants between 2016 and 2017), and one session with students from Cardiff University (2 participants in 2017).

Design and materials. Sixteen health-related press releases and accompanying articles were selected as experimental items (from Sumner et al., 2014). They all described correlational studies (i.e. there was no random allocation of participants to conditions) and none had caveats. Eight used strong, causal language in the headline and the main claim (e.g. "Extra testosterone reduces empathy") and 8 used weaker, associative language ("Extra testosterone may reduce empathy"). Causal and associative language were defined according to Adams et al. (2016).

Caveats were created for each press release (see Appendix 5). They all stated (i) the correlational nature of the study design (ii) the inability of correlational research to provide evidence for cause and effect, and (iii) the type of study design that could conclude cause and effect. These caveats were stronger than those seen in the media (compare Appendix 5 with Appendix 1), in that it is rare for caveats of observational studies to exclude cause and effect, but we wished to maximize the chances of observing effects of the caveats. The caveat was inserted into the press release for the caveat conditions and the press release was unchanged from the original for the no-caveat conditions.

Strong and weak language versions of press releases were created by altering the language appropriately. For press releases that originally used strong language, a weak version was created by inserting a modal verb into the headlines and main claims (e.g. *may* in "Extra testosterone may reduce empathy") or replacing the causal expression with an associative expression (e.g. replacing *reduces* with *linked to lower* in "Extra testosterone linked to lower empathy"); for press releases that originally used weak language, the modal verb was removed or associative expression replaced with a causal expression.

Each participant saw all 16 press releases. The two factors, language and caveat, were manipulated within subject and within item. However, the assignment of items to conditions was counterbalanced so that no single participant saw the same press release twice but across participants, all experimental press releases occurred in all conditions.

There were also four filler items. These were press releases that described causal and not correlational studies. Their purpose was to introduce variability in the range of press releases seen by participants. The four filler items were seen by all participants and were not manipulated.

Two implementation errors were detected after the experiment had been completed. First, for one press release, the strong language version was found to contain only weak language. These trials were therefore recoded as weak language trials and included in the

reported analysis. The biasing effect of including a single item in one condition and not the other was minimized by using a mixed model (glmer in R) with item as a random effect (the model adjusts for different numbers of observations per cell in the design and each item's overall contribution to variability). We also verified that complete removal of the item did not alter the qualitative conclusion of the analysis. Second, there was an error in the counterbalancing such that one group of participants saw five items in the caveat condition and three in the non-caveat condition, and another saw three in the caveat condition and five in the non-caveat condition. Mixed models also minimize the bias from such an error.

Procedure. The experiment was conducted in the school's mock newsroom. Each participant was provided with a storage drive containing a set of press release and journal article pairs. Participants were told that they could use any resources they would normally use to make selections (for example, using the internet to define terminology), but in the interest of time, they should not attempt to contact anyone for further information. They were also not allowed to communicate with other participants.

Participants were given 40 minutes in which they should indicate whether they believed the scientific findings reported in each press release-article pair were newsworthy by indicating "yes" or "no" to the instruction, "Please indicate whether you think this research should be put forward for a news article". Participants were not told to select any particular number but most selected around half as newsworthy (see Results, Figure 5).

Results and Discussion

Selection of press release. The effects of caveat on the choice of press release was analysed using a generalized linear mixed model (glmer in r) with a binomial linking function. The model is shown below.

```
selected ~ caveat*language + (1|subject) + (1|item)
```

where `selected` refers to a binary outcome variable indicating whether the press release was accepted (newsworthy) or rejected (not newsworthy), `caveat` (yes, no) refers to the presence of a caveat in the press release, `language` to whether the press release used

strong language (yes, no), and `subject` and `item` to participant and press release random factors respectively. Random slopes for subject and items were omitted because they were highly correlated with intercepts and contributed little variability overall. We initially included age in the model but as in other experiments there was no significant effect, χ^2 (1) = 0.12, p = 0.79 (we only tested the main effect of age because including the interaction terms failed to converge). We therefore omitted age from the model specification.

There was no effect of caveat (see Figure 5), β = 0.038, SE = 0.98, t = 0.39, p = .70, BF[3] = 0.23, nor was there an interaction between caveat and language, β = 0.083, SE = 0.098, t = 0.85, p = .40, BF = 0.31. Language also had no effect, β = 0.039, SE = 0.10, t = 0.39, p = 0.40, BF = 0.29. Overall the results demonstrate that caveats do not lower the uptake of press releases, consistent with Experiments 1 to 3.

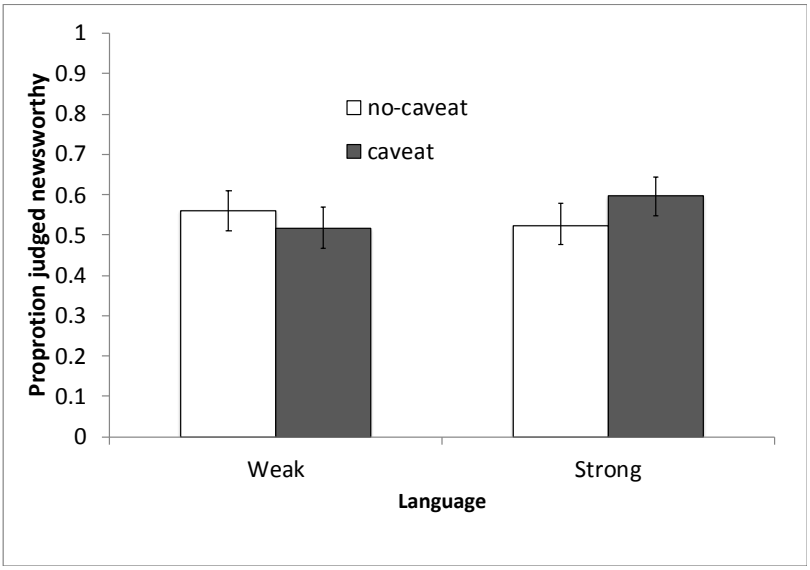

Figure 5. Proportion of press releases judged newsworthy, Experiment 4. Error bars are standard errors of the mean.

Analysis of comments. Participants were encouraged to provide reasons for their choice about whether a press release was newsworthy. Across 29 students, 628 reasons were generated. We performed a post hoc, incremental classification process whereby a single coder followed the first three phases of thematic analysis (Braun & Clarke, 2006) to classify

[3] There are currently no accepted algorithms for deriving mixed model BFs with categorical data. We therefore report BFs for repeated measures ANOVAs with caveat and language as repeated measures factors.

each reason into distinct categories. This resulted in 28 distinct categories. We then assigned a frequency count to each category as a function of our design (see Table 2) and whether the press release was judged newsworthy.

Table 2. Reasons for press releases choice.

Reason category	Not newsworthy				Newsworthy			
	Caveat		No-caveat		Caveat		No-caveat	
	W	S	W	S	W	S	W	S
1. appealing/interesting	7	9	6	10	15	16	13	11
2. size of audience	2	4	5	6	12	17	15	11
3. novelty	7	6	6	7	5	6	7	4
4. complexity	6	11	9	5	0	3	4	3
5. specific target (e.g. specific audience)	6	6	3	5	4	2	5	4
6. helpfulness to reader	2	1	1	0	3	8	6	10
7. impact of research/implications	1	1	0	0	6	5	8	9
8. study quality	6	3	6	0	3	6	5	1
9. topic (popular, current, over-reported)	2	1	0	1	7	0	7	6
10. influence behaviour of readers	3	1	3	2	1	5	3	5
11. relationship strength	8	3	3	3	3	1	0	2
12. importance	1	2	1	2	5	2	4	5
13. common knowledge	3	4	9	2	1	0	0	1
14. attention grabbing ("groundbreaking")	2	0	3	4	2	6	0	1
15. negative (including controversial)	3	2	3	6	2	0	0	1
16. balanced reporting	1	0	0	0	3	5	4	4
17. further research needed	8	4	2	1	0	0	0	0
18. accessibility (ease of understanding)	0	0	1	1	3	5	3	1
19. misleading (potential to be misunderstood)	4	3	3	1	0	0	2	0
20. human interest	1	0	2	0	6	2	1	0
21. debatable (sparks discussion)	0	0	0	0	2	2	1	4
22. inserted caveat	3	2	0	0	0	0	0	0
23. press release quality	1	0	1	2	0	0	1	0
24. source quality (e.g. "Cambridge University")	0	0	0	0	1	1	2	1
25. shareable (social media)	0	0	0	0	0	1	2	1
26. positive	0	0	0	0	1	1	1	0
27. political	1	0	0	0	2	0	0	0
28. entertaining	1	0	0	0	0	0	1	0

Note. Categories are in descending order of frequency. Scores are counts of the respective reasons given. *Not newsworthy* and *newsworthy* columns refer to the selection decision associated with a reason; *Caveat* and *No-caveat* to caveat conditions respectively; and W/S to weak/strong language conditions respectively. For example, the "novelty" reason was given 7 times for the no-caveat, strong language condition, to justify "not newsworthy".

For the effect of caveat, *inserted caveat* (Category 22), which included only those comments that referred to the inserted caveat, and *further research needed* (Category 17) were the most relevant categories. When the press release was rejected, more of these reasons were given in the caveat condition than the no-caveat condition (5 vs 0; 12 vs 4; respectively), suggesting that caveats did indeed play a role in judging press releases as newsworthy. However, it is important to compare this effect with other reasons that were unrelated to caveats. In particular, *size of expected audience* (total count = 72), *novelty* (total count = 48), and *appeal* (total count = 87) were often described for accepting and rejecting press releases, and overall in the caveat condition, there were only 17 examples of reasons related to caveats versus 306 reasons unrelated to caveats.

The pattern for strong/weak language was similar. Here, *relationship strength* (Category 11) and *further research needed* (Category 17) are the categories most related to the hypothesis. *Relationship strength* was stated more often as a reason for rejection of press releases for weak than strong language (11 vs 6), but at the same rate for acceptance of the press release (3 vs 3), and *further research needed* was stated more often for weak language than strong language for rejection of press releases (10 vs 5) but not for acceptance of press releases. Thus there was some weak evidence that strong/weak language played a role in judgements of newsworthiness but overall, factors relating to weak/strong language played a very small role in decision making (total count = 38) when compared to other factors (total count = 628).

Discussion

In Experiment 4 we tested whether the effects of Experiments 1 to 3 would generalise to a different paradigm, a different type of caveat (strong or weak language), and a different

type of participant. The experiment used a mock newsroom where postgraduate journalism students made decisions about which press releases to publish. Our results were consistent with previous experiments. Neither the presence of the caveat nor the language used to describe the headline had an effect on whether the press release should be published.

We also analysed the reasons given for choosing or not choosing to publish. Here, participants occasionally highlighted the caveat or type of language as a reason, but this effect was dwarfed by the number of reasons related to novelty, appeal and other references to general content. Much like Experiment 3, reasons other than the caveat were much more important in determining levels of interest.

One limitation of Experiment 4 is that we cannot be sure how representative journalism students are of genuine journalists. The students were all enrolled on a postgraduate media course during which they underwent journalistic training and obtained professional journalistic work experience but there is nonetheless a difference between being a student and an experienced professional. The professional might have gained experience or received informal instruction encouraging them to treat press releases with caveats disfavourably, in contrast to the conclusions we present here. An appropriate generalisation of this experiment then is that there is nothing in the formal training of journalists that leads them to judge caveats disfavourably.

A further criticism is that we did not test how journalism students understood the caveats (e.g. we did not ask how confident the researchers were). Our reasons for this were that we wanted first to make the main task realistic, and second to ask about their reasons in an unconstrained way rather than to lead them to focus on confidence/caution. Nonetheless, it remains possible that journalism students read the caveats differently to lay people and did not derive the same level of caution (although the size and generality of the effects in Experiments 1 to 3 suggest that this is unlikely).

General Discussion

We report four experiments investigating the effect of caveats on reader perceptions. Together, we found that caveats were effective at communicating caution, yet they did not lower perceived interest or make the texts more difficult to understand. We therefore argue that caveats are an effective and useful technique for communicating caution in science-based news stories.

Generalisation across caveats

There are many different sorts of caveats. For example, caveats can be used to indicate a general level of caution ("Further research is needed to confirm our findings"), to highlight a certain design ("The experiment was conducted on animals and we would need to test our theories on humans before making practical recommendations"), or to warn of specific limitations ("The study was conducted on UK participants and so caution is warranted when generalising to participants from other countries). Therefore it is useful to consider to what extent our findings generalise across the range of caveats.

In Experiment 1, we used caveats that expressed a need for further research and that provided an alternative explanation for the observed results. These are representative of many of the caveats often used by writers (see Appendix 1), as are the caveats used in Experiment 3 (which were unaltered caveats taken from genuine press releases), and our conclusions should therefore apply in many real-world situations. In Experiment 2, we used clauses that made little sense as caveats but were prefaced by a caveat opening clause, such as "One limitation…" This experiment widens our claims to caveats that are less comprehensible than those used in Experiments 1 and 3. Given that an opening clause is sufficient to lower confidence, a caveat that did not lower confidence when prefaced appropriately would have to have further content that boosted confidence. Finally, in Experiment 4, the caveats were stronger than those typically used in the media and those from the previous experiments, in that they included information denying cause and effect conclusions. Even these caveats did

not influence judgements of newsworthiness. We also tested the effects of including qualifying expressions (e.g. *may* and *might*; a form of lexical hedging) to headlines and found similar results. The breadth of caveats tested in our study and the consistency of the results suggest that our conclusions should hold for a wide range of sensible caveats.

However, there are some caveats we did not test and that may give rise to different results. One possibility is a caveat that might make the news story more interesting. For example, a caveat could be written to imply controversy or mystery (the scientist as maverick or detective), or a journey into the unknown (the scientist as explorer). Media coverage of climate change provides an example. Here, the media have often been accused of over-playing the level of uncertainty and disagreement among scientists over the existence of climate change (Dearing, 1995; Corbet & Dunfree, 2004; Gelbspan, 1998). The motivation for exaggerating uncertainty might be lobbying by the energy companies (Gelbspan, 1998) but it might also be to improve reader engagement with scientifically dense material.

Effectiveness may also depend on whether the caveat is presented as asserted content ("This was a very small study. Nonetheless, we are optimistic that it will generate advantages…") or as a presupposition ("Whilst this was a very small study, we are optimistic that it will generate advantages… ", "Although this was only a pilot study, it represents a potentially important step in developing new treatments"; see Appendix 1, ID = 04-11-017, for a genuine example). Presuppositions assume the content is known to the reader prior to the statement (backgrounded content; Karttunen, 1974), and if it is not, the content must be accommodated. In the case of caveats expressed as presuppositions, the caveat is unlikely to be known in advance and will therefore require accommodation. Since backgrounded information is less important to the conversation than asserted content, and the process of presupposition accommodation is computationally difficult (see Chemla & Bott, 2013; Schwartz, 2007), participants may choose not to process the presupposed caveat deeply.

Presupposed caveats may therefore be less effective than the caveats used in this study, which were all asserted caveats.

Report length and caveats

We have argued that two important factors in the inclusion of caveats are their perceived effectiveness and their perceived effect on interest. Another possibility, however, is that the added length of a caveat biases writers against their inclusion. Writers do not want long news stories and so perhaps they omit the caveat on grounds of length. To some extent additional length is a necessary consequence of allowing the reader to make a more informed decision about the strength of the science underpinning the claims but we question whether the writer should be concerned about the addition of a few sentences to the press release. Caveats in our experiments were associated with more material (caveat texts were longer than no caveat texts) but equal interest. Thus it is unlikely that more material generally lowers interest. Furthermore, a regression analysis on the Sumner et al. (2014) corpus showed that press release uptake was positively associated with length (albeit with very low effects sizes), suggesting that within reasonable limits, longer press releases should not bias editors against selecting them for publication[4].

Caveats and lexical hedging

The focus of this paper has been on sentence level hedging (caveats) but in Experiment 4, we also tested lexical hedging, such as *could* or *may*. We found that the two types of hedging affected reader perceptions in a similar way. What are the advantages of each? Lexical hedges are appealing because they are short and require no scientific expertise to

[4] We implemented a General Estimating Equation (GEE) regression predicting press release uptake (i.e. whether a press release appeared in the media) as a function of press release characteristics for the Sumner et al. (2014) data. This demonstrated that the incident rate ratio for the body word count of news stories was 1.002 (95% CI: 1.0001-1.004). Thus press release uptake increased 1.002 times for every additional word in the press release.

include, unlike caveats, which can be long and often require the scientist themselves to write. Lexical hedges apply to the main claims of the story, and so may alter the initial encoding of the text, whereas caveats apply after the main text, correcting inferences that may (or may not) have arisen. This property could make caveats less effective than lexical hedging because of the difficulties of correcting pre-existing information (Lewandowsky et al., 2012; although we saw no evidence of this). Nonetheless, caveats also have advantages over lexical hedges. Caveats allow writers to explain the reason for caution. As we show in Experiment 1, specific caveats are more effective than general caveats, and so a more detailed explanation can communicate a greater degree of caution. Providing a reason for the caveat might also prevent readers interpreting the entire scientific source with scepticism (as can happen in legal testimony when one piece of evidence is discredited, see Lagnado & Harvey, 2008). The caveat can also be placed into quotes and ascribed to the scientist, rather than the anonymous writer (the journalist, by default), which could make the hedge more authoritative and consequently more effective (see Jensen, 2008). Furthermore, caveats could be written in such a way that it makes the uncertainty itself appealing, as we discussed above. Choosing between lexical hedges and caveats therefore depends on the audience and the intentions of the writer. Of course, we see little harm in including both, since they fulfil different purposes and might complement each other (the caveat could be the explanation for the earlier lexical hedge)[5].

Power and effect size

In none of our experiments did we observe a significant effect of the caveat on interest. Could this be due to low sensitivity of the experiments? We computed Bayes Factors to answer this question and in all cases the important comparisons resulted in low Bayes Factors (less than 0.33). Furthermore, combining the interest scores across Experiments 1 to 3 to

[5] An analysis of the Sumner et al. (2014) corpus revealed some evidence of an association between lexical hedging and caveats in press releases. There was a trend showing that lexical hedging was more frequent in main statements when the press release contained a caveat compared to those that did not (19.6% vs. 10.3%; $\chi^2(1) = 3.53$, $p = .06$). However, there was no difference for press release titles (10.9% vs. 8.9%; $\chi^2(1) = 0.2$, $p = .66$).

create a more powerful test yields the same result: no significant effect of caveat on levels of interest, $t(273)$, $p = 0.41$, $d = 0.05$, and a low Bayes Factor, BF = 0.095. Thus, using all the available data, the observed pattern is over ten times more likely under the hypothesis that there is no effect of the caveat than it is under the alternative. Our failure to find a significant effect was not due to low sensitivity, therefore, but represents "strong" support for the hypothesis that there is no effect of the caveat (see Dienes 2011, 2014, and Rouder et al., 2009).

Nonetheless, this conclusion assumes that the potential effect sizes are comparable to other observed psychological effects (a consequence of the "non-informative" prior used in the Bayes Factor calculations). If they are extremely small, we would not have had the power to detect them, and so cannot draw conclusions about their absence. However, small effect sizes mean that caveats have small effects on readership. Other factors, such as the topic or the comprehensibility of the material would outweigh many fold the cost to readership of including the caveat (see Experiments 3 and 4). To put this in perspective, the effect size of the content manipulation in Experiment 3 was 10 times that of the (nonsignificant) effect of the caveat (d = 0.54 *vs* d = 0.023). Overall, whether there is a small effect or no effect should be of little concern to editors or writers because in either case other factors will have a much larger impact on readership.

Psychology of caveats

In the Introduction we considered some of the cognitive processes underlying caveats. Here we consider how our results relate to these possibilities and the subsequent effects of caveats on knowledge representation.

We suggested that one effect of caveats would be to make the story less coherent. Previous findings demonstrate that the removal of causal structure from stories lead to reductions in coherence (e.g. Corner & Hahn, 2009; Johnson-Laird, 2012; Pennington & Hastie, 1988, 1992, 1993; Lagnado, 2011; Thagard, 2000). However, we found no evidence

for this. Stories with caveats were reported to be just as interesting and easy to understand as those without. One potential reason for this is that the caveats we tested only addressed causal relationships indirectly, and so didn't weaken the causal structure sufficiently to lower coherence. For example, the caveats in Table 1 refer to how the results generalise rather than to the lack of causality *per se*. Similarly, our caveats often provided reasons for the caution (e.g. "One caveat is that sportspeople with alcohol sponsorship also had more problems in their romantic relationships, which means they were more stressed.") and so did not create an unexplained gap in the causal structure that was difficult to fill (see Johnson & Seifert, 1994; Tenney, Cleary & Spellman, 2009; for similar arguments from the misinformation literature). Of course, caveats that directly contradict causal implications of the story ("There is no causal relationship between Vitamin D and cancer"), and that provide no explanation, might indeed lower the coherence. Our evidence suggests that more sensible caveats, such as those we have used here, should not.

We also considered whether participants would omit the caveat from their representations of the news story. The motivation for this prediction was previous work on misinformation and retractions (see Lewandowsky, Ecker, Seifert, Schwarz, and Cook, 2012), which found that updating initial beliefs with new information is difficult. In our study, however, we found large effects of the caveat on judgements of researcher confidence (e.g. $d = 0.90$ and $d = 1.18$ in Experiment 1a). Participants clearly had no difficulty recognising that their initial beliefs (represented by the no caveat condition) needed to be adjusted to reflect the caution in the caveat.

What could explain the disparity across studies? We offer three explanations. First, our caveats refuted inferences rather than explicit assertions. Inferences may be easier to disregard than assertions or they may not be fully formed (although there is evidence that some inferences are just as resistant as explicit text, see O'Brien, Shank, Myers & Raynor, 1988; Garrod, O'Brien, Morris, & Rayner, 1990). For example, the caveats in Table 1 caution

against generalizing the results, but there was never an assertion in the text that the results should be generalized. In contrast, the refutations used in the misinformation literature refer to earlier assertions.

A second possible explanation is that the test question was different across paradigms. We asked about confidence/caution whereas the retraction literature typically asks about causality. The different degrees of abstraction required for these questions could focus attention on different parts of the text. Finally, there are differences in the manner in which the information is presented. We presented the caveats as part of single text (a news story) whereas the retraction literature typically presents information as a sequence of bulletins, or disconnected messages. Retractions may have less of an effect when they are not integrated with the main text.

Limitations and future directions

The data from this study show that caveats are effective at communicating caution. However, what we do not yet know is how effective caveats are at influencing decision-making and behaviour. For example, does a caveat highlighting the associative nature of a study influence a person's decision to stop taking their medication? Just because people understand and recognise the pragmatic intent of a caveat, does not mean that they will necessarily alter their behaviour as a consequence.

We are optimistic about the link between caveats and behaviour, however. First, there are strong associations in general between science news and health behaviour (e.g. Matthews et al., 2016; Ramsay, 2013), so caveats would have to be an exception to the general pattern of influence if people ignored them. Second, there is evidence from work on decision making and source credibility that suggests people take into account uncertainty of the type communicated by caveats when they make decisions (e.g., Corner & Hahn, 2009). This work tested manipulations similar to the insertion of caveats, such as argument structure and source reliability in science news stories but using dependent measures that were closer to decision

making than those we used. For example, in one of their tasks, Corner and Hahn asked participants to rate the "truth of a claim" from a science-based news story, and found that the source reliability and the structure of the evidence influenced responses. Similarly, Adams et al. (2016) showed that lexical hedging influenced whether one variable was perceived to "cause" another. If source reliability and lexical hedging alters the perceived truth of scientific claims in the news, it seems likely that caveats would do also. However, whether these effects generalise to life threatening changes in behaviour, such as taking medication, is unknown (and difficult to test).

The effect of memory on caveats is another important question. Relevant behaviour is rarely required immediately after reading the science story and memory processes might distort the caveat. Much research demonstrates that surface information is quickly forgotten and only the gist of the story or the important information is retained (e.g., Johnson, 1970; Sachs, 1967). Assuming that participants do not perceive the caveats as central to the theme of the story, lasting effects will depend on the representation of the main claim rather than remembering the caveat itself. Future research needs to identify whether the remembered gist represents any of the cautious intent of the caveat, to what extent forgetting limits the applicability of caveats, and how caveats can be written to mitigate these effects (e.g. positioning the caveat in text positions most resistant to forgetting; using lexical hedging, which alters the easily-remembered headline rather than the text). More positively, long term memory effects are of little consequence for writers concerned about the effects of caveats on story appeal. Editors will make a judgment about the story's interest immediately after reading the story, much like in our experiment, and it is unlikely that caveats will become more important over time with respect to interest.

We tested two sorts of individual difference covariates, age and education. We thought the effectiveness of caveats might be reduced in older people because they are cognitively impaired relative to their younger counterparts (Park, Lautenschalegr, Hedden, Davidson,

Smith AD & Smith PK, 2002) and are particularly vulnerable to forgetting the context and source of the material (Skurnik, Yoon, Park & Schwartz, 2005; Wilson & Park, 2008). However, contrary to the hypothesis, age failed to be a significant predictor of confidence/caution or interest/newsworthiness. Nonetheless, we had a small sample for investigating individual differences (80-100 per experiment) and we especially lacked participants at the older end of the scale (e.g. N = 5 aged over 65 in Experiment 1a). The null effect may simply be due low power and non-uniform sampling. We therefore think it prudent to conduct a dedicated study before concluding that caveats are similarly effective at all ages.

Practical recommendations

The goal of our study was to investigate whether caveats are a useful technique for introducing caution. In this respect we provide the following advice for science writers (including scientists, press officers and journalists). (1) An effective approach is to use caveats that declare the need for further research and explain why the further research is necessary. If space is limited then even expressing the need for further research alone is useful to the reader. (2) A caveat can be effective even if it is complex - to the point that many readers may not understand its relevance - provided that the material is signalled appropriately (with e.g. “One limitation…”). (3) Caveats like those we used here do not appear to make a story less interesting or less newsworthy and any undetected effect is negligible compared to judicious choice of source material.

Conclusion

Our study makes two important contributions to understanding how readers perceive caveats in news stories: (1) general, specific, and even incomprehensible caveats can be effective at altering the perceived confidence or caution of scientists in their conclusions; (2) such caveats do not cause noticeably lower levels of reader interest, nor do they make the text more difficult to understand. It remains to be seen whether caveats alter people’s behaviour

as well as their cognition, by, for example, preventing patients from stopping their medication, or whether the communicated caution persists over the long term. On balance our findings support and extend guidelines from the Academy of Medical Sciences, HealthNewsReview.org and others encouraging the use of caveats in news reporting. Moreover, they provide reassurance to writers and editors that inclusion of a caveat will not be detrimental to their readership.

References

Academy of Medical Sciences (2017). Evidence reporting in the media. In *Enhancing the use of scientific evidence to judge the potential benefits and harms of medicines* (4.2). Retrieved from https://acmedsci.ac.uk/file-download/44970096

Adams, R. C., Sumner, P., Vivian-Griffiths, S., Barrington, A., Williams, A., Boivin, J., Chambers, C. & Bott, L. (2017). How readers understand causal and correlational expressions used in news headlines. *Journal of Experimental Psychology: Applied, 23(1),* 1-14.

Angell, M., & Kassirer, J. P. (1994). Clinical research – what should the public believe? *The New England Journal of Medicine, 331,* 189-190.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*(4), 390-412.

Bates, D., Mächler, M., Bolker, B. M., & Walker, S. C. (2015). Fitting linear fixed-effects models using lme4. *Journal of Statistical Software, 67*, 1-48.

Boivin, J., Bunting, L., Koert, E., ieng U, C., & Verhaak, C. (2017). Perceived challenges of working in a fertility clinic: a qualitative analysis of work stressors and difficulties working with patients. *Human Reproduction*, *32*(2), 403-408.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology. 3,* 71-101.

Chemla, E., & Bott, L. (2013). Processing presuppositions: dynamic semantics vs pragmatic enrichment. *Language and Cognitive Processes, 28*, 241-260.

Corbett, J. B., & Durfee, J. L. (2004). Testing public (un)certainty of science – Media representations of global warming. *Science Communication, 26*, 129-151.

Collins, H. M. (1987). Certainty and the public understanding of science: Science on television1. *Social Studies of Science, 17*(4), 689-713.

Corner, A., & Hahn, U. (2009). Evaluating science arguments: evidence, uncertainty, and argument strength. *Journal of Experimental Psychology: Applied*, *15*(3), 199.

Cram, P., Fendrick, A. M., Inadomi, J., Cowen, M. E., Carpenter, D., & Vijan, S. (2003). The impact of a celebrity promotional campaign on the use of colon cancer screening - the Katie Couric effect. *Archives of Internal Medicine, 163*, 1601-1605.

Crismore, A., & Vande Kopple, W. J. (1988). Readers' learning from prose: The effects of hedges. *Written Communication*, *5*(2), 184-202.

Crismore, A., & Vande Kopple, W. J. (1997). Hedges and readers: Effects on attitudes and learning. In R.Markkanen & H. Schro¨der (Eds.), *Hedging and discourse: approaches to the analysis of a pragmatic phenomenon in academic texts* (pp. 83–114). Walter de Gruyter.

Dearing, J. W. (1995). Newspaper coverage of maverick science: Creating controversy through balancing. *Public Understanding of Science, 4*, 341–61.

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science, 6*, 274–290.

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology, 5*, 1–17.

Ecker, U. K., Lewandowsky, S., Swire, B., & Chang, D. (2011). Correcting false information in memory: Manipulating the strength of misinformation encoding and its retraction. *Psychonomic Bulletin & Review*, *18*(3), 570-578.

Garrod, S., O'Brien, E. J., Morris, R. K., & Rayner, K. (1990). Elaborative inferencing as an active or passive process. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*(2), 250.

Gelbspan, R. (1998). *The heat is on: The climate crisis, the cover-up, the prescription*. New York: Perseus.

Gilbert, R., Martin, R. M., Beynon, R., Harris, R., Savovic, J., Zuccolo, L., ... & Metcalfe, C. (2011). Associations of circulating and dietary vitamin D with prostate cancer risk: a systematic review and dose–response meta-analysis. *Cancer Causes & Control*, *22*(3), 319-340.

Glover, M (2017). Media reports of science: caveats reduce audience perceptions of causality and researcher confidence. *Cardiff University Undergraduate Thesis*. Available on request from the first author.

Graesser, A. C., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological review*, *101*(3), 371.

Grilli, R., Ramsay, C., & Minozzi, S. (2002). Mass media interventions: effects on health services utilization. *Cochrane Database of Systematic Reviews*, *1*, CD000389.

Guéraud, S., Harmon, M. E., & Peracchi, K. A. (2005). Updating situation models: The memory-based contribution. *Discourse Processes*, *39*(2-3), 243-263.

Hahn, U., & Oaksford, M. (2007). The rationality of informal argumentation: A Bayesian approach to reasoning fallacies. *Psychological review*, *114*(3), 704.

Hahn, U., Harris, A. J., & Corner, A. (2009). Argument content and argument source: An exploration. *Informal Logic*, *29*(4), 337-367.

Health News Review. *Observational studies – does the language fit the evidence? – Association versus causation*. Retrieved from http://www.healthnewsreview.org/toolkit/tips-for-understanding-studies/does-the-language-fit-the-evidence-association-versus-causation/

Hyland, K. (1996). Talking to the academy: Forms of hedging in science research articles. *Written Communication, 13*, 251-281.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLOS Medicine, 2(8),* e124.

Jensen, J. D. (2008). Scientific uncertainty in news coverage of cancer research: Effects of hedging on scientists' and journalists' credibility. *Human communication research, 34*(3), 347-369.

Jensen, J. D., Pokharel, M., Scherr, C. L., King, A. J., Brown, N., & Jones, C. (2017). Communicating uncertain science to the public: How amount and source of uncertainty impact fatalism, backlash, and overload. *Risk Analysis, 37(1),* 40-51.

Johnson, R. E. (1970). Recall of prose as a function of the structural importance of the linguistic units. *Journal of Memory and Language*, *9*(1), 12.

Johnson-Laird, P. N. (1980). Mental models in cognitive science. *Cognitive science, 4*(1), 71-115.

Johnson, H. M., & Seifert, C. M. (1994). Sources of the continued influence effect: When misinformation in memory affects later inferences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(6), 1420.

Karttunen, L. (1974). Presupposition and linguistic context. *Theoretical Linguistics, 1*, 181-194.

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological review*, *95*(2), 163.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2016). lmerTest: Tests in Linear Mixed Effects Models (R package version 2.0-33) [Software]. Available from https://CRAN.R-project.org/package=lmerTest

Lagnado, D. A. (2011). Thinking about evidence. In *Proceedings of the British Academy* (Vol. 171, pp. 183-223). Oxford, UK: Oxford University Press.

Lagnado, D. A., & Harvey, N. (2008). The impact of discredited evidence. *Psychonomic bulletin & review*, *15*(6), 1166-1173.

Lewandowsky, S., Ecker, U. K., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest*, *13*(3), 106-131.

Lewis, J. (2001). *Constructing public opinion: how political elites do what they like and why we seem to go along with it*. Columbia University Press.

Lewis, J., Williams, A., & Franklin, B. (2008). A compromised fourth estate? *Journalism Studies, 9*, 1–20.

Matthews, A., Herrett, E., Gasparrini, A., Staa, T. V., Goldacre, B., Smeeth, L., & Bhaskaran, K. (2016). Impact of statin related media coverage on use of statins: interrupted time series analysis with UK primary care data. *BMJ, 353*, i3283.

Morey, R. D., & Rouder, J. N. (2015). BayesFactor: Computation of Bayes Factors for Common Designs (R package version 0.9.12-2) [Software]. Available from https://CRAN.R-project.org/package=BayesFactor.

Nelkin, D. (1995). *Selling science*. New York: Freeman.

Oaksford, M., & Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford University Press.

Olausson, U. (2009). Global warming - global responsibility? Media frames of collective action and scientific certainty. *Public Understanding of Science, 18*, 421-436.

O'Brien, E. J., Rizzella, M. L., Albrecht, J. E., & Halleran, J. G. (1998). Updating a situation model: A memory-based text processing view. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*(5), 1200.

O'Brien, E. J., Shank, D. M., Myers, J. L., & Rayner, K. (1988). Elaborative inferences during reading: Do they occur on-line? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(3), 410.

Parascandola, M. (2000). Health in the news: What happens when researchers and journalists collide. *Research Practitioner, 1*, 1–29.

Park, D. C., Lautenschlager, G., Hedden, T., Davidson, N. S., Smith, A. D., & Smith, P. K. (2002). Models of visuospatial and verbal memory across the adult life span. *Psychology and Aging, 17(2),* 299–320.

Pellechia, M. G. (1997). Trends in science coverage: A content analysis of three U.S. newspapers. *Public Understanding of Science, 6,* 49-68.

Pennington, N., & Hastie, R. (1988). Explanation-based decision making: Effects of memory structure on judgment. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(3), 521.

Pennington, N., & Hastie, R. (1992). Explaining the evidence: Tests of the Story Model for juror decision making. *Journal of personality and social psychology*, *62*(2), 189.

Pennington, N., & Hastie, R. (1993). *The story model for juror decision making* (pp. 192-221). Cambridge: Cambridge University Press.

Phillips, D. P., Kanter, E. J., Bednarczyk, B., & Tastad, P. L. (1991). Importance of the lay press in the transmission of medical knowledge to the scientific community. *New England Journal of Medicine, 325(16),* 1180–1183.

Ramsay, M. E. (2013). Measles: the legacy of low vaccine coverage. *Archives of Disease in Childhood, 98,* 752-754.

Rapp, D. N., & Kendeou, P. (2007). Revising what readers know: Updating text representations during narrative comprehension. *Memory & Cognition*, *35*(8), 2019-2032.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16,* 225-237.

Sachs, J. S. (1967). Recognition memory for syntactic and semantic aspects of connected discourse. *Perception & Psychophysics*, *2*(9), 437-442.

Schäfer, M. S. (2011). Sources, characteristics and effects of mass media communication on science: a review of the literature, current trends and areas for future research. *Sociology Compass*, *5*(6), 399-412.

Schat, J., Bossema, F. G., Numans, M. E., Smeets, I., & Burger, J. P. (2018) Overdreven gezondheidsnieuws. Relatie tussen overdrijving in academische persberichten en in nieuwsmedia, *Nederlands Tijdschrift voor Geneeskunde 162*, 13-17.

Schonbrodt, F. D., Wagenmakers, E., Zehetleitner, Z., & Perugini, M. (2017). Sequential Hypothesis testing with Bayes factors: Efficiently testing mean differences. *Psychological Methods, 22*, 322-339.

Schwarz, F. (2007). Processing presupposed content, Journal of Semantics, 4, 373–416 https://doi.org/10.1093/jos/ffm011.

Schwartz, L. M., & Woloshin, S. (2004). The media matter: a call for straightforward medical reporting. *Annals of Internal Medicine*, *140*(3), 226-228.

Schwartz, L. M., Woloshin, S., & Andrews, A. (2012). Influence of medical journal press releases on the quality of associated newspaper coverage: retrospective cohort study. *BMJ, 344*, d8164.

Schwitzer, G. (2008). How do US journalists cover treatments, tests, products, and procedures? An evaluation of 500 stories. *PLoS Medicine, 5*, e95.

Science Media Centre. (2012). *10 best practice guidelines for reporting science and health stories*. Science Media Centre, London. http://www.sciencemediacentre.org/wp-content/uploads/2012/09/10-best-practice-guidelines-for-science-and-health-reporting.pdf

Skurnik, I., Yoon, C., Park, D. C., & Schwarz, N. (2005). How warnings about false claims become recommendations. *Journal of Consumer Research, 31*, 713–24.

Straight Statistics and Sense About Science. (2010). *Making sense of statistics*. Straight Statistics and Sense About Science, London.

Stryker, J. E, Moriarty, C. M., & Jensen, J. D. (2008). Effects of newspaper coverage on public knowledge about modifiable cancer risks. *Health Communication, 23,* 380–390.

Sumner, P., Vivian-Griffiths, S., Boivin, J., Williams, A., Venetis, C. A., Davies, A., Ogden, J., Whelan, L., Hughes, B., Dalton, B., Boy, F., & Chambers, C. D. (2014). The association between exaggeration in health related science news and academic press releases: Retrospective observational study. *BMJ, 349*, g7015.

Sumner, P., Vivian-Griffiths, S., Boivin, J., Williams, A., Bott, L., Adams, R., ... & Chambers, C. D. (2016). Exaggerations and caveats in press releases and health-related science news. *PLoS One*, *11*(12), e0168217.

Tenney, E. R., Cleary, H. M., & Spellman, B. A. (2009). Unpacking the doubt in "Beyond a reasonable doubt": Plausible alternative stories increase not guilty verdicts. *Basic and Applied Social Psychology*, *31*(1), 1-8.

Thagard, P. (2000). *How scientists explain disease*. Princeton University Press.

Tenney, E. R., Cleary, H., & Spellman, B. A. (2009). The other dude did it: A test of the alternative explanation defense. *Jury Expert*, *21*, 37.

Travis, S. (2010). CNN poll: Quarter doubt Obama was born in US. *CNN Politics*.

Sumner, P., Vivian-Griffiths, S., Boivin, J., Williams, A., Bott, L., Adams, R., Venetis, C., Whelan, L., Hughes, B., & Chambers, C. D. (2016). Exaggerations and caveats in press releases and health-related science news. *PLOS ONE, 11(12),* e0168217.

Williams, A., & Clifford, S. (2009). Mapping the field: Specialist science news journalism in the UK national media. Science and the Media Expert Group, Department of Business, Innovation and Skills.

Wilkes, A. L., & Leatherbarrow, M. (1988). Editing episodic memory following the identification of error. *The Quarterly Journal of Experimental Psychology*, *40*(2), 361-387.

Wang, M. T., Bolland, M. J., & Grey, A. (2015). Reporting of limitations of observational research. *JAMA internal medicine, 175*(9), 1571-1572.

Yavchitz, A., Boutron, I., Bafeta, A., Marroun, I., Charles, P., Mantz, J., & Ravaud, P. (2012). Misrepresentation of randomized controlled trials in press releases and news coverage: A cohort study. *PLoS Medicine, 9*, e1001308.

Acknowledgements

We thank Cameron Dunlop, Madelaine Glover and Sofia Grammenos for the collection of pilot data and Chris Chambers for helpful advice. We also thank Michael Hill from Cardiff School of Journalism and Anne Harbin from UWE School of Film and Journalism for their role as gatekeepers to participants. The corpus used in this study was supported by ESRC Grant ES/M000664/1 and ESRC grant ES/M500422/1. We thank the following for contributing to the database or other work leading up to this study: S. Vivian-Griffiths. C. Chambers, Caitlin Argument, Amy Barrington, Laura Benjamin, Hannah Coulson, Eleanor Corney, Bethan Dalton, Cecily Donnelly, Cameron Dunlop, Rebecca Emerson, Rose Fisher, Oliver Gray, Bethan Hughes, Katie John, Laura Jones, Sarah Mann, Olivia Manship, Hannah Maynard, Hannah McCarthy, Jack Ogden, Amy Parfitt, Naomi Scott, Lauren Stead, Eliza Walwyn-Jones, Claire Weeks, Leanne Whelan, Joe Wilton.

Appendices

Appendix 1. Table of caveats retrieved from Sumner et al. (2014)

Press release ID	Caveat text
02-11-030	However, the children in this study often had access to at least five different devices at any one time, and many of these devices were portable. This meant that children were able to move the equipment between their bedrooms and family rooms, depending on whether they wanted privacy or company. This suggests that we need to work with families to develop strategies to limit the overall time spent multi-screen viewing wherever it occurs within the home.
03-11-014	We now need to see an extension of this study, one which tests larger numbers of people, and then take it out of the hospital and in to the home setting.
03-11-026	However, they conclude that, given the health benefits of eating chocolate, initiatives to reduce the current fat and sugar content in most chocolate products should be explored.
04-11-016	The reasons for this poorer health are not clear. There is an argument for the health, social and criminal justice agencies to work together to limit post-adolescent offending, reducing the risk of illness in later life and the cost to society.
04-11-017	Whilst this was a very small study the key aim was to establish whether this technique may be feasible for sufferers. The scientists now hope to take this method further in formal clinical trials in order to establish whether it holds promise for patients.
05-11-023	We have not found the actual genetic differences that cause some intelligence differences, but we now have evidence that some of the genetic causes are linked to those genetic factors that we tested. This gives us leads that we are now planning to follow.
06-11-018	However, due to natural variation in telomere length from person to person, the test is only effective at a population level, and will not provide useful information on how long an individual can expect to live.
07-11-025	Studies like this are making really important progress and whilst we must always be cautious when taking findings from rodents into humans, these are very interesting and potentially important results.
08-11-014	Researchers say although it needs validation, this test could improve… The researchers caution that it is not yet known why these factors are associated with a lower risk in this study… We now need to validate the test in further studies,…
08-11-018	We acknowledge however, that our finding represents only a small part of the genetic risk for depression and more and larger studies will be required to find the other parts of the genome involved.
08-11-021	What we need to do now is to find out exactly how Bmi1 and Hoxa9 proteins sustains the growth of cancer cells in order to develop an effective treatment to stop the disease returning
11-11-001	Of course, our well-being isn't determined by this one gene – other genes and especially experience throughout the course of life will continue to explain the majority of variation in individual happiness.

12-11-030	However, further validation work is needed before the scales can be recommended for use in routine clinical practice, they conclude.
13-11-009	One limitation, however, was that the study used weight and height information reported by the women and not measured by health professionals. The study used information collected routinely during the women's antenatal visits, and so could not examine whether lifestyle factors such as diet, exercise, alcohol and caffeine consumption influenced pregnancy risks.
13-11-010	What we need to examine further is why some people are more susceptible to developing diabetes than others. Despite being a very small trial, we look forward to future results…
14-11-006	The researchers warned that as the figures only show the number of diagnoses at GUM clinics, rather than the total number of infections of illnesses which can in some cases be asymptomatic.
14-11-007	More research is needed into this issue although we already know that smoking does have an impact on sperm development,
14-11-018	This was an unexpected finding, and so further research using other data sources is needed to confirm these findings as well as provide more evidence on the benefits of different antidepressants in this group of people. They also caution that differences between patients prescribed different antidepressant drugs may account for some of the associations seen in the study, underlining the need for further research to confirm the findings.
14-11-021	Further studies are now required to explore how and when solids should be introduced alongside breastfeeding to aid protection against eczema and other allergic diseases. The size of this study means that its findings are very significant, although the authors recognise that further studies are required.
15-11-001	The researchers only examined data at a national level and they are now examining data at an individual level to try to establish what drives people to overeat.
15-11-006	Although this is a significant and promising result, there are a number of steps to be taken before this new form of drug delivery can be tested in humans in the clinic. He also notes that other steps would be needed before exosomes could be tested in humans, including safety tests and scaling up the procedures.
15-11-007	while the numbers involved were small
15-11-011	However, these two groups of mothers and children are very different across a number of measures, such as mother's age, education and socio-economic position. It could be that breastfeeding is serving as a proxy for something else causing the difference in rates of behavioural problems among the children.
15-11-016	The study did not look at the avoidability of adverse outcomes in different settings, any effect of staffing levels or the configuration of maternity services, or provide detailed analysis of transfers. The study does not provide any quantitative data to address the different ways of organising service provision and any association with quality of care.
16-11-006	While this combination treatment still has to go to phase two of trials…
18-11-013	More research is needed to assess the extent of unnecessary treatment and its impact on quality of life.
19-11-010	Of course, babies cannot tell us how they feel, so it is impossible to know what babies actually experience. We cannot say that before this

	change in brain activity they don't feel pain.
20-11-005	Our findings highlight the need for prospective follow-up studies of regulatory disturbed infants and require reliable assessments of crying, sleeping, or feeding problems.
20-11-019	Our results suggest the need to focus on preventing factors that contribute to child maltreatment…
01-11-009	… although the conclusions are tentative owing to the non-randomised nature of the studies.
02-11-013	There's still a way to go before we fully understand the link between a person's vitamin D levels and their risk of cancer. But we still need more research to clarify whether vitamin D directly prevents bowel cancer or if people with higher levels are generally healthier. There's no convincing evidence to suggest that vitamin D offers any protection against other types of cancer developing.
17-11-005	However, as 20% of the patients were still being prescribed antibiotics, the research could not judge the impact of the NICE guidelines on this group. However, it does not rule out the possibility that antibiotics may be beneficial in certain circumstances and further research is needed to look into these in more detail.
02-11-028	It is unclear how the health risks compare between a woman whose blood pressure rises a lot during pregnancy
07-11-040	This was only a small study
08-11-022	They emphasised that this is an early study, and larger studies will need to be conducted to verify these results.
01-11-020	It is difficult to assess the size of the additional clinical benefit, because these patients were well nourished, and had the highest quality standard therapy anyhow.
14-11-015	The next stage of the research will be to assess whether the group therapy approach works equally well in other centres through a larger study The next stage will be to find if this approach is as effective in other areas of the country. Future research will greatly benefit from the MS Society-supported MS Register project.
03-11-034	However, before this medication can be used for the treatment of stimulant-dependent individuals in clinical practice, more research would be needed using multiple doses over longer period of time.
06-11-026	Our pilot study suggests… … although the long term effects remain to be elaborated. … This study was small and did not allow the testing of all proteomics data so we need larger, more in-depth studies to develop this potential further, and we need longer term studies to link patterns to disease outcomes.
02-11-024	The researchers add there could be a number of reasons for the apparent lack of effect of increased activity: that it was not intense enough, or that it was too early in the disease process for exercise to show an effect. It is also possible that those in the diet and exercise group modified their behaviour and diluted the effect of both interventions, for example, rewarding themselves with extra food due to increased exercise. Further research is needed to clarify whether more intensive or different types of activity, or activity advice offered at a later stage of diabetes will add benefits to diet interventions, or whether benefits of activity interventions will become more apparent after one year.
03-11-043	not found

02-11-052	However, before this gene therapy approach can be trialled in patients, additional pre-clinical studies need to be performed in order to verify not only the efficiency and the safety of AAVs-mediated NGF in type 1 diabetes, but also to find the most efficient AAV serotype, as well as the optimal dose and delivery route to be used.
05-11-033	Although it has been known for some time that DNA mutations predispose individuals to the development of schizophrenia, it has remained a puzzle as to how these genes cause behavioural problems.
06-11-027	This was a small study, and we need more research to confirm its findings, but it does give us a clue to how some of the benefits of exercise might take place.
15-11-003	these results have been seen in a small, healthy group of volunteers, and that these are short-term, not sustained, manipulations of the participants' beliefs about the treatment.
15-11-022	Much larger clinical studies would be needed to show that brain stimulation had a lasting effect in producing clinical benefits for stroke patients. 'This was a study in a small group. Large-scale trials would be needed before concluding that the approach benefits those recovering from strokes

Note. We were unable to find caveats in one of the press releases (03-11-043) listed as containing them. We presume this is an error in the initial coding.

Appendix 2. Material for Experiment 1.

Story 1

Extra testosterone reduces empathy

A new study from Utrecht and Cambridge Universities has for the first time found that an administration of testosterone under the tongue in volunteers reduces a person's ability to 'mind read' which is an indication of empathy.

The researchers used the 'Reading the Mind in the Eyes' task as the test of mind reading, which tests how well someone can infer what a person is thinking or feeling from photographs of facial expressions from around the eyes. Mind reading is one aspect of empathy, a skill that shows significant sex differences in favour of females.

The researchers found that administration of testosterone led to a significant reduction in mind reading. Given that people with autism have difficulties in mind reading, and that autism affects males more often than females, the study provides further support for the 'extreme male brain' theory of autism.

Specific caveat

However, the scientists emphasize the need for more research before generalizing their results. They note that the task was quite simple and that there are other components of mind reading that were not captured by their study.

General caveat

However, the scientists emphasize the need for more research before generalizing their results.

Story 2

Low income and poor diet results in accelerated ageing

A new study looking at the DNA of people living in Glasgow suggests that earning less than the average wage and eating an unhealthy diet accelerates the ageing process. The study, conducted by the University of Glasgow compared the length of telomeres in blood samples taken from 382 Glaswegians, from the most and least deprived parts of the city.

Telomeres, the tails on the ends of chromosomes, shorten throughout a person's life and can be used as a measure of the ageing process. This study is a first for the city in that it shows that adverse social conditions influence the biology of ageing and therefore disease.

It is hoped that the findings will help to create a test that can be used for faster feedback on the effects of public health improvement measures.

Specific caveat

However, the scientists still believe that more work needs to be done. For example, telomeres can be shortened by many factors aside from poor diet, such as a sedentary lifestyle and smoking, and so they hesitate to draw strong causal conclusions from their study.

General caveat

However, the scientists still believe that more works needs to be done.

Story 3

TV advertising of unhealthy food increases children's preference for high-fat and high-sugar foods

Researchers at the University of Liverpool have found that watching adverts for unhealthy food on television encourages children to want to eat high-fat and high-sugar foods. The study by researchers in the Institute of Psychology, Health and Society examined

the food preferences of a group of 281 children aged six to 13 years old from the North West of England.

The children were shown an episode of a popular cartoon preceded by five minutes of adverts showing either toys or fast food. They were then asked to choose what they would like to eat from a list of unhealthy and healthy food items.

The study found that viewing the fast food adverts led to children preferring high-fat and high-sugar foods.

Specific caveat

The scientists, however, argue that more work is required before firm conclusions can be drawn. They note that the adverts for fast food were generally more interesting than those for toys, and this could have explained the results.

General caveat

The scientists, however, argue that more work is required before firm conclusions can be drawn.

Story 4

Alcohol sponsorship of sport increases the number of sportspeople drinking

New research from the University of Manchester has provided fresh evidence that alcohol industry sponsorship of sport facilitates hazardous drinking in sportspeople compared to non-alcohol sponsorship.

Previous research had already established higher levels of drinking in sportspeople with alcohol sponsorship. What sets this study apart is that it is the first to compare alcoholic with non-alcoholic sponsorship in order to rule out the financial gain of sponsorship as an explanation.

The results showed exactly that, with 68% of sportspeople with alcoholic sponsorship meeting the World Health Organization classification for hazardous drinking. Health campaigners argue this study shows that alcohol sponsorship in a sport leads to hazardous drinking within the sport.

Specific caveat

The researchers broadly agree, but admit that the picture is still incomplete. For instance, teams with alcohol sponsorship might win more games, and so fans would have more cause for celebration.

General caveat

The researchers broadly agree, but admit that the picture is still incomplete.

Story 5

Boredom results in dangerous driving

Research at the University of Newcastle has found that acts of dangerous driving, such as speeding and overtaking, are attributable to boredom behind the wheel. The 1,563 drivers surveyed were placed into four groups based on their self-reported driving habits, ranging from ‘easily bored, nervous, and dangerous’ to ‘safe and slow’.

They were then asked to rate how fast they would go in various road conditions. It was found that boredom whilst driving determines how dangerously someone might drive.

The implications of this finding are that road planners could make roads safer by increasing their difficulty to drive on by, for example, building obstacles.

Specific caveat

However, the researchers emphasize that it is still early for road changes to be made and additional studies are needed. They note that their study did not directly test the effects of

boredom on real driving, they relied entirely on questionnaires about driving. This could mean that people exaggerated or under-reported the effects of boredom.

General caveat

However, the researchers emphasize that it is still early for road changes to be made and additional studies are needed.

Story 6

Having slim parents lowers childhood obesity

Children with thinner parents are three times more likely to be thin than children whose parents are overweight, according to a new study by UCL researchers.

Between 2001 and 2006, trained interviewers recorded the heights and weights of parents and up to two children in 7,000 families, and used this information to calculate their BMI. In the case of the children, the international obesity task force criterion was used to predict what their BMI would be in adulthood.

The results showed that having slim parents does lead to thinness in children. This suggests that parents should be mindful of their weight when bringing up children.

Specific caveat

However, the scientists highlight that this is only one of the many studies that need to be done to find the cause of obesity. For instance, the study did not take into account environmental factors, such as sport and diet, and these would have a bigger impact on childhood obesity.

General caveat

However, the scientists highlight that this is only one of the many studies that need to be done to find the cause of obesity.

Story 7

High confidence boosts women's spatial skills

Boosting a woman's confidence makes her better at spatial tasks, University of Warwick scientists have found. The researchers tested spatial ability through a series of four computer-based experiments that women had previously been found to perform poorly on.

At the same time, the women's confidence was artificially varied by giving feedback that they were above or below average on a prior judgement task. The results showed that performance on spatial tasks was significantly enhanced by increasing confidence.

This suggests that the difference between men and women on spatial tasks may actually be manufactured from stereotypical jokes which adversely affect women's confidence rather than a true gender difference.

Specific caveat

However, the researchers admit that there is a lot still to do in order to understand gender effects on cognition. They add that all the participants had university degrees and it could be that the results only apply to those with high levels of education.

General caveat

However, the researchers admit that there is a lot still to do in order to understand gender effects on cognition.

Story 8

Low levels of self-control leads to physical health problems and financial difficulties.

Children as young as three with low levels of self-control have more physical health problems and financial difficulties in later life, according to a new King's College London.

Those taking part in the studies completed a range of physical tests and interviews to assess a range of genetic and environmental factors that can shape children's lives.

Self-control skills such as conscientiousness, self-discipline and perseverance were assessed by teachers, parents, observers and the participants themselves. Scientists observed that having low self-control at a young age results in health problems and difficulties with money management in later life.

Early intervention to make small improvements to a child's self-control could not only reap benefits to individual lives but also reduce societal costs.

Specific caveat

Nonetheless, the researchers admit that they are still far from understanding the exact causes of health and financial problems, highlighting the need for follow-up studies. They remark upon other environmental factors that could play a crucial role, such as life changing events, quality of relationships and family upbringing, which were not tested in their study.

General caveat

Nonetheless, the researchers admit that they are still far from understanding the exact causes of health and financial problems, highlighting the need for follow-up studies.

Story 9

Music enhances productivity in the workplace

Listening to music at work helps office workers relax, improve their mood and make them feel happier, according to research from the University of Sheffield.

These benefits have the knock on effect of improving concentration, and therefore productivity. This was formally tested on over 300 participants with a wide range of musical preferences including classical, rock, and pop music.

The results showed that the most commonly reported benefit of music at work was an improvement in concentration due to the blocking of background noise. This suggests that employers should seriously consider introducing music in some format around the workplace.

Specific caveat

However, the scientists caution that the research is in its early stages and further studies are required to verify the conclusions. For example, they note that they only tested easy tasks, and difficult tasks might be affected differently.

General caveat

However, the scientists caution that the research is in its early stages and further studies are required to verify the conclusions.

Appendix 3. Material for Experiment 2.
The body text was the same as that for Experiment 1. Caveats are listed below.

Nonsense caveats and exciting sentences presented in each story

Story	Nonsense caveat	Nonsense exciting
1	One caveat to the findings, however, is that testosterone also led to increased activation in the fusiform face area, part of the fusiform gyrus, an area of increasing debate about its function.	In an exciting development, researchers also found that testosterone led to increased activation in the fusiform face area, part of the fusiform gyrus, an area of increasing debate about its function.
2	However, the scientists also warn that telomeres are truncated during cell division, their presence stopping the genes before them on the chromosome from being truncated.	The scientists were also intrigued to find that the telomeres were truncated during cell division, their presence stopping the genes before them on the chromosome from being truncated.
3	However, one limitation is that both the healthy and unhealthy food items contained large biomolecules consisting of one or more long chains of amino acid residues.	Interestingly, both the healthy and unhealthy food items contained large biomolecules consisting of one or more long chains of amino acid residues.
4	One caveat to the findings, however, is that sportspeople with alcohol sponsorship also had more problems in their romantic relationships, which means they were more stressed.	In an exciting development, researchers also found that sportspeople with alcohol sponsorship had more problems in their romantic relationships, which means they were more stressed.
5	However, the scientists also warn that driving requires the integration of information from multiple visual and auditory sources, leading to increased activation in both the occipital and the parietal lobes.	The scientists were also intrigued to find that driving requires the integration of information from multiple visual and auditory sources, leading to increased activation in both the occipital and the parietal lobes.
6	However, one limitation of the study is that according to the international obesity task force criterion, an obese child has a body max index that is 2 standard deviations above the	Interestingly, according to the international obesity task force criterion, an obese child has a body max index that is 2 standard deviations above the World Health Organization growth standard

	World Health Organization growth standard median.	median.
7	One caveat to the findings, however, was that spatial visualization is characterized as complicated multi-step manipulations of spatially presented information, which involves the parietal lobe.	In an exciting development, researchers also found that spatial visualization could be characterized as complicated multi-step manipulations of spatially presented information, which involves the parietal lobe.
8	However, the scientists also warn that the ability to control one's impulses and modulate one's emotional expressions is the earliest and most ubiquitous demand that societies place on children.	The scientists were also intrigued to find that the ability to control one's impulses and modulate one's emotional expressions is the earliest and most ubiquitous demand that societies place on children.
9	However, one limitation of the study is that people who preferred listening to rock music were often introverted, less hard-working and passive.	Interestingly, people who preferred listening to rock music were often introverted, less hard-working and passive.

Appendix 4. Materials for Experiment 3.

Note that caveats are enclosed by asterisks in each story.

Unappealing (Story 1)

Dietary advice improves blood sugar control for recently diagnosed type 2 diabetes patients

New research from academics at the University of Bristol shows that, in patients with recently diagnosed type 2 diabetes, 6.5 hours of additional dietary advice sessions leads to improvement in blood sugar control compared with patients who receive usual care. However, increased activity conferred no additional benefit when combined with the diet intervention.

The study, published online first by *The Lancet*, is led by Dr Rob Andrews, Consultant Senior Lecturer in Diabetes and Endocrinology in the University of Bristol's School of Clinical Sciences.

The study assessed 593 adults aged 30—80 years in whom type 2 diabetes had been diagnosed five to eight months earlier. Of these, 99 were assigned to usual care, 248 to diet advice only, and 246 to diet advice plus exercise. Usual care patients received an initial dietary consultation plus follow-up every six months. Diet-only group patients were given a dietary consultation every three months with additional nurse support each month. Diet and exercise patients received the same as diet only patients but were also asked to do 30 minutes of brisk walking five times a week (with activity assessed by pedometers that showed good adherence).

The researchers found that in the usual care group, blood sugar control had worsened, with mean HbA1c (a method of assessing blood sugar control) levels increasing from 6.72per cent to 6.86 per cent over six months, before falling back to 6.81 per cent at 12 months. In the diet advice group, HbA1c fell from a mean 6.64 per cent pre-intervention to 6.57 per cent at six months and 6.55 per cent at 12 months. Exercise did not confer additional benefit on top of the diet advice, apart from in those patients with the highest HbA1c, insulin-resistance, or body-mass index at baseline.

Dr Andrews said: "These findings suggest that intervention at this early stage should focus on improving diet, since the additional cost of training health-care workers to promote activity might not be justified."

*The researchers add there could be a number of reasons for the apparent lack of effect of increased activity: that it was not intense enough, or that it was too early in the disease process for exercise to show an effect. It is also possible that those in the diet and exercise group modified their behaviour and diluted the effect of both interventions, for example, rewarding themselves with extra food due to increased exercise.

Dr Andrews concluded: "Further research is needed to clarify whether more intensive or different types of activity, or activity advice offered at a later stage of diabetes will add benefits to diet interventions, or whether benefits of activity interventions will become more apparent after one year."*

Unappealing (Story 2)

Study sheds light on late phase of asthma attacks

New research led by scientists from Imperial College London explains why around half of people with asthma experience a 'late phase' of symptoms several hours after exposure to allergens.

The findings, published in the journal *Thorax*, could lead to better treatments for the disease.

An estimated 300 million people suffer from asthma, and the prevalence is rising.

Symptoms are commonly triggered by allergens in the environment, such as pollen and dust mites.

These stimuli can cause the airways to tighten within minutes, causing breathing difficulties which range from mild to severe.

Many sufferers also experience a 'late asthmatic response' three to eight hours after exposure to allergens, causing breathing difficulties which can last up to 24 hours.

In the early asthmatic response, the allergen is recognised by mast cells, which release chemical signals that cause the airways to narrow.

In contrast, the mechanism behind the late phase has remained unclear.

In research on mice and rats, the Imperial team have now found evidence that the late asthmatic response happens because the allergen triggers sensory nerves in the airways.

These nerves activate reflexes which trigger other nerves that release the neurotransmitter acetylcholine, which causes the airways to narrow.

If the findings translate to humans, it would mean that drugs that block acetylcholine - called anticholinergics - could be used to treat asthma patients that experience late phase responses following exposure to allergens.

Steroids are the main treatments for asthma prescribed now, but they are not effective for all patients.

A recent clinical trial involving 210 asthma patients found that the anticholinergic drug tiotropium improved symptoms when added to a steroid inhaler, but the reason for this was unexplained.

"Many asthmatics have symptoms at night after exposure to allergens during the day, but until now we haven't understood how this late response is brought about," said Professor Maria Belvisi, from the National Heart and Lung Institute at Imperial College London, who led the research.

"Our study in animals suggests that anticholinergic drugs might help to alleviate these symptoms, and this is supported by the recent clinical data.

We are seeking funding to see if these findings are reproduced in proof of concept clinical studies in asthmatics."

The researchers hypothesised that sensory nerves were involved after observing that anaesthesia prevented the late asthmatic response in mice and rats.

They succeeded in blocking the late asthmatic response using drugs that block different aspects of sensory nerve cell function, adding further evidence for this idea.

After establishing that sensory nerves detect the allergen, the researchers tested the effect of tiotropium, an anticholinergic drug that is used to treat chronic obstructive pulmonary disease.

Tiotropium blocks the receptor for acetylcholine, which is released by nerves in the parasympathetic nervous system.

Tiotropium also blocked the late asthmatic response, suggesting that parasympathetic nerves cause the airways to constrict.

The study was funded by the Medical Research Council (MRC).

Professor Stephen Holgate, MRC funding board chair and an expert on asthma, said: "Unravelling the complex biology of asthma is vitally important, as it is an extremely dangerous condition which exerts lifelong damaging effects.

The Medical Research Council is committed to research that opens doors to improving disease resilience, particularly in conditions which attack our body over the long-term.

Studies like this are making really important progress and whilst we must always be cautious when taking findings from rodents into humans, these are very interesting and potentially important results."

Appealing (Story 3)

Babies distinguish pain from touch at 35-37 weeks

Babies can distinguish painful stimuli as different from general touch from around 35-37 weeks gestation – just before an infant would normally be born – according to new research.

In a study published online in the journal Current Biology, scientists show that neural activity in the brain gradually changes from an immature state to a more adult-like state from 35 weeks of development.

This change may indicate that neural circuitry allows babies to process pain as a separate sensation from touch.

Dr Rebeccah Slater, UCL Neuroscience, Physiology and Pharmacology, said: "Premature babies who are younger than 35 weeks have similar brain responses when they experience touch or pain.

After this time there is a gradual change, rather than a sudden shift, when the brain starts to process the two types of stimuli in a distinct manner."

Scientists looked at the brain activity of 46 babies at the University College Hospital Elizabeth Garrett Anderson Wing.

21 babies in the study were born prematurely, giving scientists the opportunity to measure activity at different stages of human brain development, from babies at just 28 weeks of development through to those born 'full term' at 37 weeks.

Using electroencephalography (EEG), the scientists measured the babies' electrical brain activity when they were undergoing a routine heel lance – a standard procedure essential to collect blood samples for clinical use.

In the premature babies the EEG recorded a response to the heel lance of non-specific 'neuronal bursts' – general bursts of electrical activity in the brain.

After 35-37 weeks the babies' response changed to localised activity in specific areas of the brain, indicating that they were now perceiving painful stimulation as separate to touch.

Dr Lorenzo Fabrizi, lead author of the paper from UCL Neuroscience, Physiology and Pharmacology, said: "We are asking a fundamental question about human development in this study – when do babies start to distinguish between sensations?

In very young brains all stimulations are followed by 'bursts' of activity, but at a critical time in development babies start to respond with activity specific to the type of stimulation."

Dr Fabrizi added: "Of course, babies cannot tell us how they feel, so it is impossible to know what babies actually experience. We cannot say that before this change in brain activity they don't feel pain."

Previous studies have shown that there is a similar shift from neuronal bursts to evoked potentials in the visual system at this time, suggesting that 35-37 weeks is a time when important neural connections are formed between different parts of the brain.

Dr Slater said: "It is important to understand how the human brain develops so that we can provide the best clinical care for hospitalised infants."

Appealing (Story 4)

The placebo effect: expecting the best, fearing the worst

Poor expectations of treatment can override all the effect of a potent pain-relieving drug, a brain imaging study at Oxford University has shown.

In contrast, positive expectations of treatment doubled the natural physiological or biochemical effect of the opioid drug among the healthy volunteers in the study.

The study of the placebo effect – and its opposite the nocebo effect – is published in Science Translational Medicine.

The findings suggest that doctors may need to consider dealing with patients' beliefs about the effectiveness of any treatment, as well as determining which drug might be the best for that patient.

'Doctors shouldn't underestimate the significant influence that patients' negative expectations can have on outcome,' says Professor Irene Tracey of the Centre for Functional Magnetic Resonance Imaging of the Brain at Oxford University, who led the research. '

For example, people with chronic pain will often have seen many doctors and tried many drugs that haven't worked for them.

They come to see the clinician with all this negative experience, not expecting to receive anything that will work for them.

Doctors have almost got to work on that first before any drug will have an effect on their pain.'

The placebo effect describes the improvements seen when patients – unknowingly – are given dummy pills or sham treatments but believe it will do them good.

This is a very real physiological effect; it is not just about patients 'feeling' better.

The nocebo effect is the opposite: patients see poorer outcomes as the result of doubts about a medical treatment.

Previous studies have investigated the basis of the placebo effect, when using sugar pills or saline injections for example, and confirmed it can elicit a real response.

This new research, funded by the Medical Research Council and German research funders, goes a step further by examining how manipulating participants' expectations can influence their response to an active drug.

The Oxford University team, along with colleagues from the University Medical Center Hamburg-Eppendorf in Germany, Cambridge University, and the Technische Universität München, set out to investigate these effects among 22 healthy adult volunteers by giving them an opioid drug and manipulating their expectations of the pain relief they might receive at different points.

The volunteers were placed in an MRI scanner and heat applied to the leg at a level where it begins to hurt – set so that each individual rated the pain at 70 on a scale of 1 to 100.

An intravenous line for administration of a potent opioid drug for pain relief was also introduced.

After an initial control run, unknown to the participants, the team started giving the drug to see what effects there would be in the absence of any knowledge or expectation of treatment.

The average initial pain rating of 66 went down to 55.

The volunteers were then told that the drug would start being administered, although no change was actually made and they continued receiving the opioid at the same dose.

The average pain ratings dropped further to 39.

Finally, the volunteers were led to believe the drug had been stopped and cautioned that there may be a possible increase in pain.

Again, the drug was still being administered in the same way with no change.

Their pain intensity increased to 64.

That is, the pain was as great as in the absence of any pain relief at the beginning of the experiment.

The researchers used brain imaging to confirm the participants' reports of pain relief.

MRI scans showed that the brain's pain networks responded to different extents according to the volunteers' expectations at each stage, and matching their reports of pain.

This showed the volunteers really did experience different levels of pain when their expectations were changed, although the administration of pain relief remained constant.

*Professor Tracey notes that these results have been seen in a small, healthy group of volunteers, and that these are short-term, not sustained, manipulations of the participants' beliefs about the treatment. *

But she says it's important not to underestimate the strength of the effect of such expectations on any treatment, and that clinicians need to know how to manage that.

Professor Tracey says there may also be lessons for the design of clinical trials.

These are often carried out comparing a candidate drug against a dummy pill to see if there is any effect of a drug above and beyond that of the placebo.

We should control for the effect of people's expectations on the results of any clinical trial.

At the very least we should make sure we minimise any negative expectations to make sure we're not masking true efficacy in a trial drug.'

Appendix 5. Materials used in Experiment 4.

Press release ID	Caveat text
15-11-011	As this was an observational study we cannot conclude that breastfeeding directly affects behaviour, other factors may have been involved and would need to be investigated with an experimental study aimed specifically at uncovering cause and effect.
19-11-014	This observational study contributes to the evidence showing that exposure to family violence is related to brain function, but we cannot rule out other factors with this type of research, we cannot make conclusions about cause and effect - for that, we would need to conduct an experimental trial.
14-11-006	With observational studies such as this, we cannot control for other potential causes of STIs. We can show a relationship with the morning after pill, but we cannot say that this is causal. We would need to perform an experimental study for that.
02-11-027	However, this was an associative study which means that we cannot rule out other factors that may explain the relationship between genetic markers and AS. A study using an experimental method would need to be utilised to show a causal relationship between the two."
03-11-026	With observational studies, we cannot show that a single thing, such as chocolate, actually causes a reduction in heart disease risk – it is possible that other factors may explain the results. Experimental trials such as randomised controlled trials would need to be conducted to establish cause and effect.
06-11-015	As this was a cohort study we cannot conclude that obesity is the primary cause of premature death, other variables may be involved. Only an experimental study would be able to demonstrate cause and effect.
14-11-018	With such observational studies we can see that the use of new generation antidepressants are related to the adverse outcomes listed in the study, but we cannot say that newer antidepressants cause such adverse effects as other factors could be involved. For causal evidence we would have had to have taken an experimental approach.
19-11-011	This study was observational, so it can increase understanding of possible links between mortality and bypass vessel density, but it cannot demonstrate cause and effect because of the possibility of contribution of other variables. The next step would be an experimental trial to establish whether this is a causal relationship.
02-11-028	Associative studies like this cannot establish direct cause and effect as we cannot rule out other explanations. Experimental trials are needed to allow the inference of cause and effect between blood pressure and health risks during pregnancy.
07-11-043	It is not possible to control for other potential causes of clots in associative cohort studies like this because this type of study does not have the same type of power to uncover causal relationships as an experimental study.
07-11-040	As this research was associative, we cannot go as far as to say that premature birth causes poor health in later life outright. Rather, we must conduct more rigorous experimental trials to make that connection.
06-11-012	It is important to remember that our study we cannot infer causation as our methods were primarily cross-sectional and observational.

	Experimental studies are required in order to show whether socioeconomic status really does affect rates of cancer.”
08-11-022	As this was an associative study there is no way to infer cause and effect between vitamin D levels and asthma. To obtain such causal evidence an experimental method would need to be used where the researchers can have more control over the factors which may contribute to asthma.
03-11-010	It is important to consider that this type of longitudinal study is purely observational, and as such, no inference can be made about causation – there may be other factors involved. We would need to conduct a more rigorous experiment to infer causality.
05-11-033	As this comparison was made using observational methods rather than experimental methods, it cannot be said that specific mutations are the cause of schizophrenia.
02-11-018	However, as this study was observational, it can be said that bone size and physical activity are related, but it cannot be concluded that one factor causes the other; for this, an experimental study needs to be conducted.