How readers understand causal and correlational expressions used in news headlines

Rachel C. Adams[1,2], Petroc Sumner[1,2], Solveiga Vivian-Griffiths[1,2], Amy Barrington[2], Andrew Williams[3], Jacky Boivin[2], Christopher D. Chambers[1,2], Lewis Bott[2]

[1]Cardiff University Brain Research Imaging Centre (CUBRIC), School of Psychology, Cardiff University, CF10 3AT, UK.
[2]School of Psychology, Cardiff University, CF10 3AT, UK.
[3]School of Journalism, Media & Cultural Studies, Cardiff University, CF10 3NB, UK.

Corresponding author: Rachel C. Adams, School of Psychology, College of Biomedical and Life Sciences, Cardiff University, 70 Park Place, Cardiff, CF10 3AT, UK, (adamsrc1@cardiff.ac.uk)

Key words: causal inference; correlation and causation; science and the media; science communication;

DOI: http://dx.doi.org/10.1037/xap0000100

Abstract

Science related news stories can have a profound impact on how the public make decisions. The current study presents four experiments that examine how participants understand scientific expressions used in news headlines. The expressions concerned causal and correlational relationships between variables (e.g. “being breast fed *makes* children behave better”). Participants rated or ranked headlines according to the extent that one variable caused the other. Our results suggest that participants differentiate between three distinct categories of relationship: direct cause statements (e.g. “makes”, “increases”), which were interpreted as the most causal; can cause statements (e.g. “can make”, “can increase”); and moderate cause statements (e.g. "might cause", “linked”, “associated with”), but do not consistently distinguish within the last group despite the logical distinction between cause and association. Based on this evidence, we make recommendations for appropriately communicating cause and effect in news headlines.

How readers understand causal and correlational expressions used in news headlines

Science stories in the media have profound effects on public health. For example, following coverage of the measles, mumps and rubella vaccine scare in the later 1990s, immunisation rates dropped (Health and Social Care Information Centre, 2009), with consequent increases in disease incidence (Ramsey, 2013). It is therefore important that science writers use language that conveys information consistent with the peer-reviewed papers. In this study we systemically test how people understand scientific expressions used in media headlines. Our overall aim is to contribute evidence-based advice for science writers attempting to clearly communicate the conclusions of a study.

There is growing evidence that science stories contain exaggerations of scientific findings (Brechman, Lee, & Cappella, 2009; Cooper, Lee, Goldacre, & Sanders, 2012; Haneef, Lazarus, Ravaud, Yavchitz, & Boutron, 2015; Leveson, 2012; Schwitzer, 2008; Sumner et al., 2014; Sumner et al., under review; Woloshin, Schwartz, Casella, Kennedy, & Larson, 2009). For example, Sumner et al. (2014), found that 33% of press releases and 81% of the associated new stories contained causal claims when the peer-reviewed papers described correlational studies. Exaggeration is problematic because if the public adjust their behaviour in proportion to the extremity and certainty of news stories, behavioural change will be exaggerated relative to the intentions of the peer-reviewed authors. The consequences could be as severe as patients refusing to take prescribed medicine (as with statins, see e.g., Bosely, 2014). Exaggeration in the media also demonstrates that there is a misalignment between science writing and the peer-reviewed articles on which they are based.

While there is a general consensus that exaggeration exists, there is no accepted explanation for why. The problem cannot be attributed solely to journalistic practices because

exaggerations appear in press releases, written by scientists and press officers, not journalists (Brechman et al., 2009; Sumner et al., 2014, under review; Woloshin et al., 2009). This suggests that all of the contributors to science in the media, including scientists, share responsibility for the failure to inform the public. One suggestion for why exaggeration occurs is that science writers are under pressure to make their stories accessible and interesting, and in doing so, they use language that results in exaggeration. For example, writers might try to avoid dry scientific jargon, like "correlates with", and instead use everyday expressions like, "increases." So, "Being breast fed *correlates with* good behaviour" becomes, "Being breast fed *increases* good behaviour." They might also try to vary the language so as not to use the same expression in every headline. Instead of always using, "linked with", say, they may prefer to use, "results in" on occasions. Finally, they may prefer to describe a study using short, succinct expressions instead of longer phrases, so that instead of "is associated with", they use, "causes."

That there are exaggerations in news stories does not mean that science writers intend to exaggerate the scientific claims. Instead, they may simply not know how the reader will understand the expressions they use (or alternatively, the science writer may not understand the specific expressions used in the peer-reviewed articles). Scientific expressions mean different things to different people. For example, "Being breast fed *is linked with* good behaviour" could be interpreted as *being breast fed causes good behaviour,* or as *being breast fed correlates with good behaviour*, depending on the reader's knowledge, prior beliefs and the general context. Furthermore, scientific articles often contain probabilistic expressions, such as "might", which have notoriously variable meanings across samples (Budescu & Wallsten, 1985), and modal verbs, such as "can" or "may", which have many subtly different senses (Kennedy, 2002). Scientific language is detailed and specific, and translating it into language that others outside of the community understand is difficult.

The difficulty of finding appropriate expressions for scientific terms is reflected in the number of best-practice guides for science writers (Science Media Centre, 2012; Straight Statistics and Sense about Science, 2010; Schwitzer, 2010). Services such as HealthNewsReview.org are popular (receiving ~ 5000 hits per day; source: www.semrush.com), and in common with the resources cited above, provide intuitively sound advice and raise awareness of the general difficulties of misinterpreting scientific expressions. However, their detailed suggestions about appropriate vocabulary are generally based on the judgment of only a few individuals. In terms of how the reader understands the relevant expressions there is little evidence-based guidance for science writers except their own personal experience and the general information provided in the above-mentioned resources. In this study we aim to provide the evidence base by systematically testing how people understand scientific expressions used.

We focus on causal and correlational expressions. Scientific studies that employ designs with random assignment to conditions are generally more informative than studies that observe existing relationships between variables; namely, the former allow causal inference whereas the latter do not. Science writers must consequently take particular care in describing studies to make sure they do not conflate correlation with causation. However, exaggeration from correlation to causation in the media is particularly common (Sumner et al., 2014). For example, where an original study makes correlative conclusions ("Being breast fed *is correlated with* good behaviour"), the associated media headline might describe the findings using causal language ("Being breast fed *results in* good behaviour"). The need for guidance on the meaning of causal and correlational expressions is particularly important.

Previous experimental work on understanding cause and correlation (e.g., Bleske-Rechek, Morrison, & Heidtke, 2015; Mueller & Coon, 2013; Norris, Phillips, & Korpan, 2003) has been from an educational perspective, rather than a media perspective, and has

focussed on whether individuals make appropriate scientific inferences from descriptions of experimental and observational study designs. These studies found that participants often confused correlation and causation. For example, Bleske-Rechek et al. (2015) presented a well-educated community sample with descriptions of a causal study (random assignment to conditions) or a correlational study (an observational study) and asked about the causal inferences that could be derived from the reported results. Bleske-Rechek et al. found that participants who read correlational studies made the same inferences as those who read causal studies. Similarly, Norris et al. (2003) found that only a third of psychology undergraduate students could correctly identify causal and correlational statements from media reports. When people are asked to extract and comprehend the relevant information from study descriptions they appear to have great difficulty. While these sorts of studies are very useful for assessing scientific understanding, they address a different question to what we are concerned with here. We sought to identify how strongly different expressions communicate causal relationships, rather than whether people can extract relevant study design information.

Overview of Experiments

Participants in the current study read headlines such as, "Being breast fed *makes* children behave better" and judged how much they thought one variable in the headline caused the other. For the breast-feeding headline, for example, they rated the extent to which being breastfed caused better behaviour in children. We used a variety of relational expressions in the headlines, such as "makes", "increases" or "is linked to", and a range of sentence frames with appropriate independent variables and outcomes. Expressions that imply a strong causal relationship between variables should lead to high causal ratings and vice versa.

We used headlines rather than complete news stories because we believe headlines are particularly important in communicating news. People arguably spend longer looking at headlines than the main text (Dor, 2003), and when they do look at the text, the headline can have a continued influence, such that misleading headlines are resistant to correction despite the subsequent text (Ecker, Lewandowsky, Chang, & Pillai, 2014; Ecker, Swire, & Lewandowsky, 2014; Lewandowsky, Ecker, Seifert, Schwarz, & Cook, 2012). We also wanted to avoid introducing extraneous material that could confound the interpretation of relational expressions in the headline (e.g., inclusion of caveats, quotes from scientists, details of experimental procedures).

The materials for relational expressions were derived from those used in an analysis of Russell Group Health and Life Sciences press releases from 2011 (http://dx.doi.org/10.6084/m9.figshare.903704; Sumner et al., 2014). Sumner et al. (2014) measured exaggerations of scientific findings in press releases and news stories. To quantify the causal inferences in headline claims, they developed a coding scheme in which each relational expression was categorized based on its causal implications (see Table S1 Supplementary Information). Expressions judged as implying the most causality (e.g., "increase", "reduces") were assigned to the *direct cause group*, expressions judged as implying *correlation* (e.g., "relates to") were assigned to the *correlation group*, and expressions of middling causality were assigned to the intervening groups. We used a selection of relational expressions from each group as the basis for our experiments. For example, in Experiment 1, we compared judgements of *direct causal* statements with *ambiguous* statements and *correlation* statements. While the judgements of Sumner et al. might turn out to be incorrect (the coding scheme is based on their intuitions), it is nonetheless a useful starting point because it presents a framework around which we can

make predictions in our task (*direct causal* expressions should be rated as most causal, *can cause* as next most causal etc.).

Experiment 1

Experiment 1 had two goals. First, we aimed to test whether people are sensitive to the difference between causal and correlational expressions in newspaper headlines. Norris et al. (2003) and Bleske-Rechek et al. (2015) demonstrate that readers often fail to distinguish between causal and correlational study designs and readers might also fail to discriminate between causal and correlational expressions.

Second, given that readers might be sensitive to the difference in strength between causal and correlational expressions, we were interested in how people understand ambiguous expressions, such as "is linked to", in "Being breast fed *is linked to* better behaviour in children." Ambiguous expressions might be understood in several ways. Readers might think that since there is no direct causal expression in the sentence, the writer must mean that there is no causal relationship. Under this reasoning, ratings for ambiguous expressions would be lower than those for causal ratings and similar to those of correlative expressions. On the other hand, readers might think the opposite: since there is no correlative expression and no statement about the absence of a relationship, ambiguous expressions should be read as communicating a strong, and quite possibly causal relationship. Here, there should be little difference between causal and ambiguous expressions, but both should be perceived as stronger than correlative expressions. Finally, readers may sense the ambiguity and rate the sentences somewhere between causal and correlative expressions (as in the coding scheme of Sumner et al., 2014).

We also collected information about the science training of the participants. We expected that participants with more science training might be more likely to derive causal

inferences from causal statements and less likely to derive causal inferences from correlational statements.

Method

Participants. Eighty-eight participants were recruited through social media (using Twitter). Seventeen participants were excluded from all statistical analyses due to study incompletion (19% attrition rate), leaving a final sample of 71 participants (49 female, 22 male; aged 17-63, M=27.72, SE=1.31). Participants were randomly allocated to one of three counterbalancing lists (*n*s = 17, 19 and 35 for each list respectively). All experiments were approved by the School of Psychology Research Ethics Committee, Cardiff University.

Design and materials. Headline topic (science, sport and business/ politics) and relationship category (direct cause, ambiguous, and correlation) were within-subject factors. We recorded participants' science experience and coded this as a between-subject factor (none, A-Level[1] only, degree only, both A-level and degree). Experiment 1 consequently had a 3 x 3 x 4 design. The dependent measure was the causality rating for each headline, this was measured using a visual analogue scale from 0 "definitely does not cause" to 100 "definitely does cause".

The materials were based around nine sentence frames. The frames were news headlines sourced from an online news search. Table 1 shows an example. There were three sentence frames from science headlines, three from sport and three from business/politics. Each of the frames included two variables, one that was more likely to be a causal agent (e.g., being breast fed) and one that was more likely to be an outcome (e.g., good behaviour). Experimental sentences were formed by inserting a relational expression between the two variables, with the causal agent always appearing in subject position and the outcome in object position (i.e. [causal agent] *expression* [outcome]). Wherever the direct cause

[1] A-levels are a national set of UK qualifications typically studied at 16-18 years.

expression specified a direction, we included a directional expression such as “higher” or “lower” in the headlines of the other conditions. For example, since “boost” communicates an increase in the outcome, such as, “Healthier diet boosts childhood IQ,” we added “higher” to the *correlation* and *ambiguous* conditions, such as, “Healthier diet has a relationship with *higher* childhood IQ.” Thus there were no differences across conditions in directional information.

The main independent variable was the relational category. There were three types: *direct cause* relationships, which used the expressions “makes”, “leads to” and “boosts”; *ambiguous* relationships, which used the expressions “is linked to”, “is connected to” and “predicts”; and *correlation* relationships, which used the expressions “is associated with”, “is related to” and “has a relationship with” (see Table 1 for an example of one sentence frame with each category of relationship; see Table 2 for all relational expressions used in Experiments 1-4).

Participants saw nine sentences. Each sentence was based on a different sentence frame and used a different relational expression. Expressions were assigned to topics (science, sport, and business/politics) in such a way that each topic included one expression from each category (consequently topic and relational category formed within-subject factors in the design).

Counter-balancing of expression to sentence frame was achieved by generating 27 sentences, three from each sentence frame, and dividing them into three counter-balancing lists (all materials are provided in the Supplementary Information). Participants were randomly assigned to one of the three lists.

Table 1. Example stimuli from Experiment 1.

Relationship Category	Sentence
Direct cause	Being breast fed **makes** children behave better
Ambiguous	Being breast fed **is linked to** better behaviour in children
Correlation	Being breast fed **is associated with** better behaviour in children

Table 2. Relational expressions used in Experiments 1-4.

	Relationship category					
	Direct cause	Can cause	Conditional cause	Ambiguous	Correlation	Conditional correlation
Experiment 1	*makes* *leads to* *boosts*			*is linked to* *is connected to* *predicts*	*is associated with* *is related to* *has a relationship with*	
Experiments 2 & 3	*makes* *leads to* *boosts* *impacts* *drives* *induces* *heightens* *increases* *influences* *is attributable to* *elevates* *optimises*	*can…*	*might…* (E2 & E3) *may…* (E3 only) *could…* (E3 only)		*is associated with* *is related to* *has a relationship with* *varies with*	
Experiment 4	*boosts* *decreases* *elevates* *increases* *leads to* *lowers* *raises* *reduces* *responsible for* *results in*	*can…*	*might…* *may…* *could…*	*is linked to* *predicts*	*is associated with* *is related to*	*might…* *may…* *could…*

Note. Experiments used only the expressions shown in the relevant sections of the table. *Direct cause*, *ambiguous*, and *correlation expressions* were inserted directly into sentence frames. *Can cause*, *conditional cause*, and *conditional correlation* expressions were formed by combining the listed modal verbs with expressions from the *direct cause* or *correlation* categories, e.g., Experiment 4 *can cause* condition used *can elevate*, and the *conditional correlation* condition used *might be associated with*

Procedure. Participants were informed that they would be taking part in a study of how people interpret information in news headlines. Sentences were presented one after another and responses collected immediately after the presentation of each sentence. Participants were asked "According to the headline, to what extent does [causal agent] cause [outcome]?" Each sentence was presented for a minimum of 5 seconds to ensure that participants read the sentences. Following the experimental questions participants were asked whether they had completed a science-based A-Level or science-based degree.

Statistical analysis

All results are reported with unadjusted p-values. Corrections for multiple comparisons were calculated for all within-test analyses and are only reported where these corrections changed the interpretation of an analysis from statistically significant to non-significant. The alpha level for comparison is shown as the p-value subscript. Departures from sphericity assumptions were corrected as a function of Huynh-Feldt epsilon. We also report sensitivity analyses in the Supplementary Information.

We used Bayes factors to interpret the evidential value of nonsignificant findings (Dienes, 2011, 2014; Rouder, Speckman, Sun, Morey, & Iverson, 2009). With no previous literature to guide an informed prior we used the default JZS prior (Rouder et al., 2009) for all analyses. The JZS prior is a non-informative objective prior that minimises assumptions regarding expected effect size. Bayes factors using the JZS prior were calculated using JASP (r was set *a priori* to the default value, $r = 0.707$; Love et al., 2015). Bayes factors > 3 suggest 'substantial' evidence for the alternative hypothesis and Bayes factors < 0.33 indicate 'substantial' evidence for the null hypothesis (Dienes, 2011, 2014). All study data is available online (https://github.com/SolveigaVG/CausalLanguage.git).

Results

Figure 1 shows mean causality ratings as a function of topic and relationship category. For each topic, *direct cause* sentences were rated as highly causal with scores between ~75-80. *Ambiguous* and *correlation* sentences were rated as much less causal with scores between ~45-55. A 3x3x4 mixed ANOVA revealed a significant overall effect of relationship category (F (2,134)=117.79, p<0.001, η^2_p =0.64), with causality ratings for *direct cause* significantly greater than for both the *ambiguous* (p<0.001; dz[2]= 1.59) and *correlation* (p<0.001; dz= 1.48) conditions. The difference between *ambiguous* and *correlation* conditions was not statistically significant and substantially favoured the null hypothesis over the alternative hypothesis (p=0.42; dz= 0.09; B=0.21). The main effect of topic (F (2, 134)=1.05, p=0.35, η^2_p =0.02; B=0.07) and the interaction between topic and category (F (4, 272)=0.24, p=0.92, η^2_p =0.003; B=0.02) were non-significant and favoured the null.

To assess whether having experience with statistics was related to causality ratings participants were asked whether they had a science-based A-level or degree. Participants were categorised as having no experience (n=18), a science A-level (n=17), a science degree (n=14), or both (n=22). The results of the ANOVA revealed no main effect of experience (F (2,68)=1.59, p=0.21, η^2_p =0.02, B=0.10) and no significant interactions with topic or relationship category (all Fs <1.19, all ps>0.29, all η^2_ps<0.05; all Bs <0.06). These results are consistent with previous findings showing that science education appears to be unrelated to how well students are able to interpret scientific media reports (Norris et al., 2003).

Due to our random assignment method there was an unequal distribution of participants to counterbalancing lists (see Participants section). We therefore conducted an additional analysis on the first 17 participants assigned to each list, i.e., the maximum number of

[2] Cohen's $d_z = \frac{M_{diff}}{\sqrt{\frac{\sum (X_{diff} - M_{diff})^2}{N-1}}}$; suggested values for small, medium and large effects are 0.2, 0.5 and 0.8, respectively (Cohen, 1988).

participants such that there was an equal distribution of participants to lists. The analysis revealed the same pattern of significant effects as the complete analysis (see Supplementary Information for full analysis).

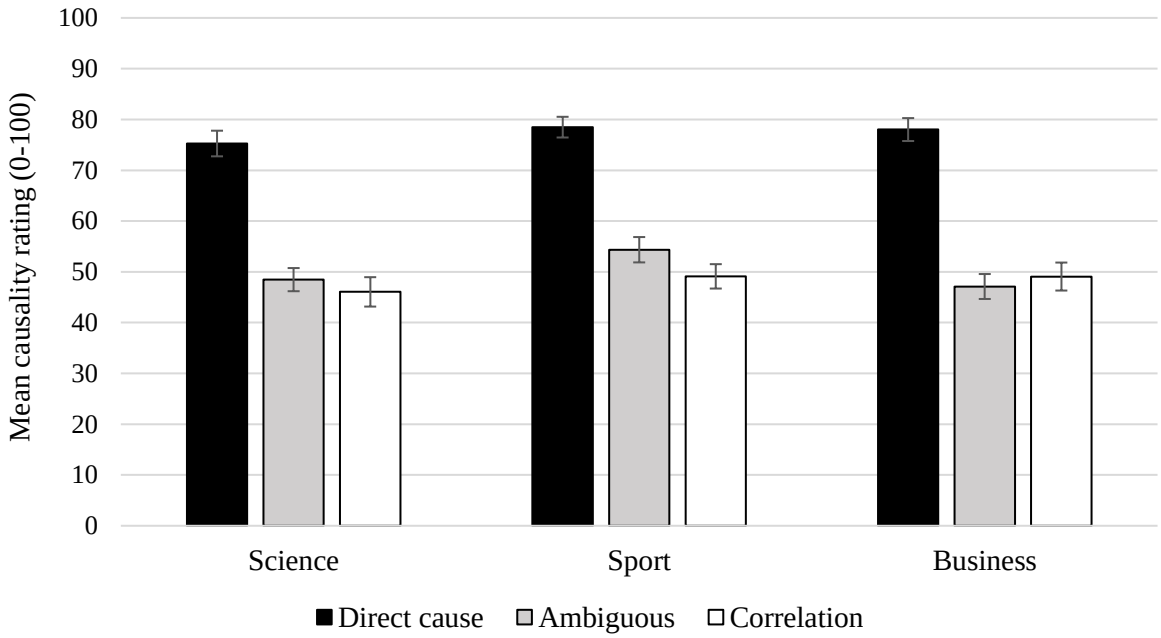

Figure 1. Mean causality ratings for the three categories of relationship as a function of topic in Experiment 1. Error bars show ±1 within subject standard error (Cousineau, 2005; Morey, 2008).

Discussion

Participants overwhelmingly rated directly causal sentences as more causal than correlational and ambiguous sentences. While this result corresponds with our own intuitions, previous studies have shown that students have difficulty distinguishing between causal and correlational designs when the studies are presented as vignettes (e.g., Bleske-Rechek et al., 2015; Norris et al., 2003) and it would not have been surprising if students were also impervious to the distinction when claims were made in sentences, as in our study.

Interestingly, however, there was no significant difference between the *ambiguous* and *correlation* conditions, and the Bayes factor demonstrated substantial support for the null hypothesis rather than a general insensitivity of our experiment. People perceive causality in

ambiguous expressions and correlational expressions equivalently, contrary to the coding scheme of Sumner et al. (2014).

Finally, although participants rated directly causal statements as most causal, they still rated correlational and ambiguous sentences as moderately causal (around 50%). This suggests that our participants were either uncertain about these phrases or that they believe even the weakest relational expressions imply causality. Results from our pilot work provide evidence for the latter. When we presented questions in which the likely cause and direction of outcome were reversed, relative to a presented news headline, causal ratings were low (~20%). This demonstrates that participants were comfortable with providing low ratings when deemed appropriate. These results also suggest that participants in the current experiment judged correlational and ambiguous expressions to imply a moderately causal relationship.

Experiment 2

In Experiments 2 and 3, we consider how modal verbs, such as *can*, *may* and *might*, alter causal inferences. Modal verbs are used when writers want to express uncertainty or doubt about the truth of their statement (in this context). For example, *might* in "Being breast fed *might* make children behave better," suggests that the writer is uncertain about the relationship between breast feeding and behaviour. However, there are different sorts of modal verbs and there may be differences in the type of uncertainty each conveys. This variation could result in differing degrees of causal implications across verbs.

In Experiment 2 we tested four categories of relational expression: *direct cause, can cause, might cause* and *correlation* (see Table 3). We had three goals. First, we wanted to confirm that the apparent uncertainty introduced by the modal verbs resulted in fewer causal implications relative to direct cause expressions. Intuition suggests they do but we can find no previous studies about this. Second, we wanted to discover whether "can" and "might"

were perceived to express different degrees of uncertainty in news headlines. “Might” conveys a possibility of an event (e.g., “John might give a good talk”), whereas “can” conveys an ability, (“John can give a good talk”) or a conditionality (“John can give a good talk if he prepares well enough”). This would suggest that “can” generates more causal implications than “might”, and reflecting this, Sumner et al. (2014) coded “can” as generating more causal implications than “might” (see Table S1, Supplementary Information).

Finally, we wanted to know whether modal verbs modifying causal expressions implied more causal implications than simple correlational expressions, such as “associated with”. Since correlational expressions do not explicitly express causality, it might be expected that modal causal expressions generate more causal implication than simple correlational expressions. This is the view of the HealthNewsReview.org website, which explicitly suggests using “associated with” instead of qualified causal expressions, such as “might boost”. More generally, a writer wishing to communicate uncertainty about a causal relationship might prefer correlational expressions for stylistic reasons, such as the added length of modal verbs, or the need to be direct without qualification.

Method

Participants. One hundred and sixty five psychology undergraduate students at Cardiff University participated for partial course credit. Five participants were excluded from the analysis for incomplete data. The remaining participants (N=160; 137 female, 21 male, 2 missing values; aged 17-30, *M*=19.36, *SE*=0.13) were randomly distributed to one of four counterbalancing lists (*n*s = 26, 24, 79 and 31).

Design, materials and procedure. Experiment 2 had a 3 x 4 x 2 design. Topic (science, sport and business/ politics) and relationship category (direct cause, can cause, might cause, correlation) were within-subject factors and Year of study, (Year 1 or 2) was a between subject factor.

The materials were based around 12 sentence frames. Nine were taken from Experiment 1 and three more were sourced using an online news search, one each for science, sport and business/politics. Construction of the experimental sentences was similar to Experiment 1.

Four categories of relationship were used: *direct cause, can cause, might cause* and *correlation*. There were 12 causal expressions and four correlational expressions (see Table 2; the assignment of expressions to relationship categories was based on Sumner et al., 2014; see Table 3 for examples). *Can cause* and *might cause* sentences were formed by inserting the words "can" or "might" prior to the direct cause expression. For *direct cause, can cause*, and *might cause* conditions, the more likely causal agent was presented in subject position and the more likely outcome in object position (i.e. [causal agent] *expression* [outcome]). Because this ordering is not consistent with many correlational headlines in the media, we reversed the order for the *correlation* condition, that is, the more likely outcome was in subject position and the more likely causal agent was in object position ([outcome] *expression* [causal agent]).

Participants saw 12 sentences. Each sentence was based on a different sentence frame and used a different expression. Three of the sentence frames were assigned *direct cause* expressions, three *can cause* expressions, three *might cause* expressions, and three *correlation* expressions. Counter-balancing of expression to sentence frame was achieved using a method similar to Experiment 1, except that there were now a pool of 48 sentences and four counter-balancing lists (see Supplementary Information).

The procedure was identical to that used in Experiment 1 with the exception that participants were asked to report their year of academic study in the debrief as opposed to whether or not they had completed a scientific degree.

Table 3. Example stimuli from Experiment 2.

Relationship Category	Sentence
Direct cause	Being breast fed **makes** children behave better
Can cause	Being breast fed **can make** children behave better
Might cause	Being breast fed **might make** children behave better
Correlation	Better behaviour in children **is associated with** being breast fed

Results

Figure 2 shows that participants rated *direct cause* statements as more causal than *correlation* statements, consistent with Experiment 1. As expected, *can cause* was rated as less causal than *direct cause* but more causal than *correlation*. More interestingly, *might cause* was rated as less causal than *correlation*.

This pattern was confirmed with a 3x4x2 mixed ANOVA. We observed a main effect of relationship category, (F (2.9, 458.77)= 84.61, p<0.001, η^2_p =0.35), with all pairwise comparisons reaching statistical significance (all $ps_{.008}$< 0.007, all dz> 0.22) including the comparison between *might cause* and *correlation* (p<0.001; dz=0.38). There was a large effect size for the comparison between *direct cause* and *might cause* (dz= 1.24) and medium-large effect sizes for the comparisons between *direct cause* and both *can cause* and

correlation conditions (*dz*= 0.58 and 0.78, respectively) and between *can cause* and *might cause* (*dz*= 0.77). The effect size was small for the comparison between *can cause* and *correlation*.

While the same general pattern holds across all three topics, there was a significant interaction between relationship category and topic (F (5.66, 893.91)=2.89, p=0.01, η^2_p =0.02). Business/ politics received lower causality ratings for the *direct cause* statement than the other topics, M=62 *vs* M = 68 and M = 71 for science and sport (p=0.003, *dz*=0.24; p<0.001, *dz*=0.33), respectively).

Participants were also asked their year of academic study to see whether experience with statistics was related to causality ratings. Eighty-five participants reporting being in the first year and 75 participants reported being in the second year of their undergraduate psychology degree. Consistent with Experiment 1 and previous research exploring the role of statistical experience (Norris et al., 2003), the results of the mixed ANOVA revealed that there was no significant main effect of year of study (F (1, 158)=0.07, p=0.79, η^2_p <0.001, B=0.12) and no significant interactions with either topic or relationship category (all Fs <1.68, all ps>0.19, all η^2_ps<0.01; all Bs<0.06).

As in Experiment 1, we conducted an additional analysis to avoid uneven counterbalancing groups using the first 24 participants from each group. The pattern of significant effects was very similar to the complete analysis except that the pairwise comparison between *can cause* and *correlate* (p=0.18, *dz*=0.13; B=0.26) and the interaction between topic and relationship category (F(5.61, 527.21)=1.69, p=0.12, η^2_p =0.02; B=0.02) were both nonsignificant (see Supplementary Information for the full analysis).

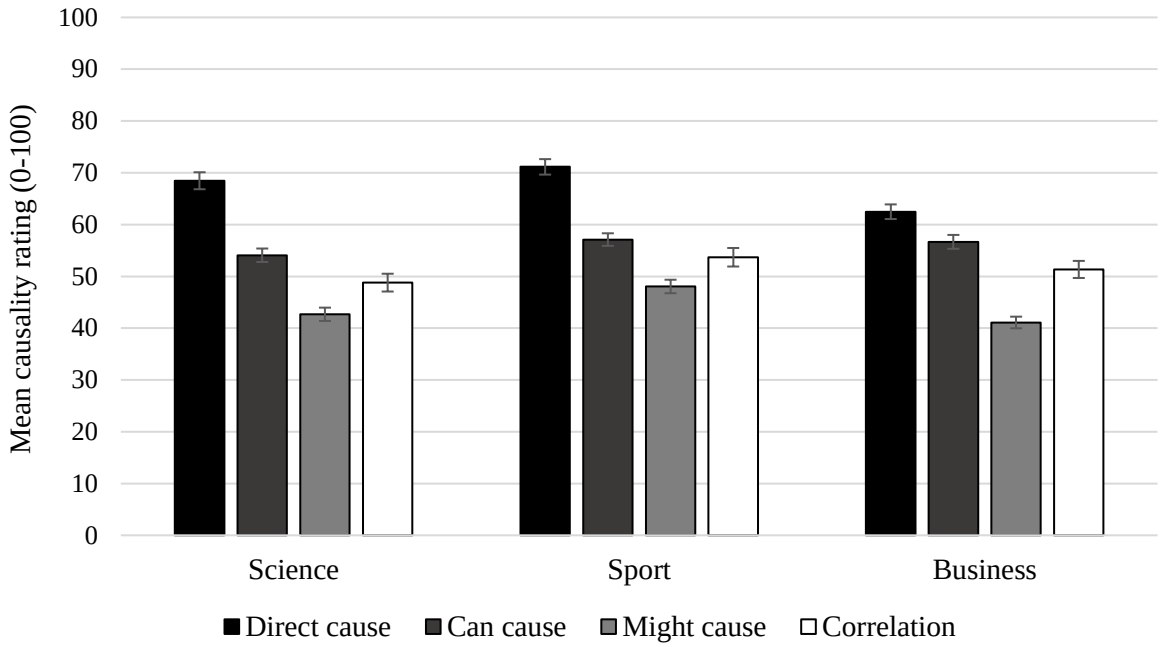

Figure 2. Mean causality ratings for the four categories of relationship as a function of topic in Experiment 2. Error bars show ±1 within subject standard error (Cousineau, 2005; Morey, 2008)

Discussion

Modal verbs reduced causality ratings relative to expressions without modal verbs. The degree depended on the particular modal verb. Contrary to Sumner et al.'s (2014) coding scheme and the advice of HealthNewsReview.org, "might cause" was rated as less causal than even simple correlational expressions. In other words, expressions such as "associated with" were perceived as more causal than "might cause", despite the intuition that "associated with" ought to convey correlation not causation.

Experiment 3

Experiment 3 continued the investigation of the modal verbs and tested *might cause*, *may cause*, *could cause* and *correlation* expressions. "May" is argued by many usage guides to express greater likelihood than "might" (e.g., the BBC world service English guide[3]). For example, "John may go to the party" implies that John attending the party is more likely than "John might go to the party". If so, "may" should lead to higher causal implications than

[3] http://www.bbc.co.uk/worldservice/learningenglish/grammar/learnit/learnitv162.shtml

"might" in the context of newspaper headlines. We were also interested in replicating the finding from Experiment 2 that *might cause* was rated less strongly than correlational expressions, and whether other modal verbs would also lead to lower causality ratings than *correlation* expressions. We therefore included *could cause* and *may cause*. Sumner et al. (2014) grouped all three modal expressions together as "conditional cause," and assumed they should generate more causal implications than *correlation* expressions.

Method

Participants. Ninety-nine psychology undergraduate students from Cardiff University participated in Experiment 3 for partial course credit. Two participants were excluded from all statistical analyses because they did not provide information regarding their academic year of study. The remaining participants (N=97; 78 female, 19 male; aged 18-46, *M*=19.86, *SE*=0.32) were randomly allocated to one of four counterbalancing lists (*n*s = 25, 25, 23 and 24).

Design, materials and procedure. Experiment 3 had a 3 x 4 x 2 design. Topic (science, sport and business/ politics) and relationship category (might cause, may cause, could cause, and correlation) were within-subject factors and Year of study, (Year 1 or 2) was a between subject factor.

We used the same sentence frames as those in Experiment 2. The relational categories were different, however. Here we used *might cause, may cause, could cause* and *correlation* (see Tables 2 and 4, and Supplementary Information). Sentence construction and counterbalancing was the same as Experiment 2, except that the ordering of causal agents and outcomes was consistent across all four categories.

The procedure was identical to that used in Experiment 2.

Table 4. Example stimuli from Experiment 3.

Relationship Category	Sentence
Might cause	Being breast fed **might make** children behave better
Could cause	Being breast fed **could make** children behave better
May cause	Being breast fed **may make** children behave better
Correlation	Being breast fed **is associated with** better behaviour in children

Results

Consistent with the results of Experiment 2, ratings for modal and correlational expressions were in the moderately causal range (between ~45-60), although causality ratings of the modal conditions (*might cause, may cause, could cause*) were lower than those of the *correlation* condition (Figure 3). A 3x4x2 mixed ANOVA revealed a significant main effect of relationship category (F (2.31, 219.71) =9.38, p<0.001, η^2_p =0.09). Pairwise comparisons showed that each modal condition was rated as significantly less causal than the correlation condition (all ps <0.004, all dzs > 0.3 and were in the small-medium range). The modal conditions did not significantly differ from one another (all ps$_{.008}$ >0.039, all dzs < 0.21; all Bs <0.91). This was confirmed with an additional exploratory ANOVA when the *correlation* condition was removed (main effect of expression: F (1.85, 175.9) =1.94, p=0.15, η^2_p =0.02; B= 0.15).There was no main effect of topic (F (2,190) =1.59, p=0.21, η^2_p =0.02; B=0.04) and no significant interaction between topic and relationship category (F (5.82, 552.7) =1.12, p=0.35, η^2_p =0.01; B=0.007).

To explore the relationship between statistical experience and causal ratings, participants were again asked whether they were in the first (n=51) or second (n=46) year of their degree. The mixed ANOVA showed no main effect of year of study (F (1, 95) =0.27, p=0.59, η^2_p =0.003, B=0.23) and no significant interactions with either topic or relationship category (all Fs <0.81, all ps>0.49, all η^2_ps<0.01; all Bs <0.5).

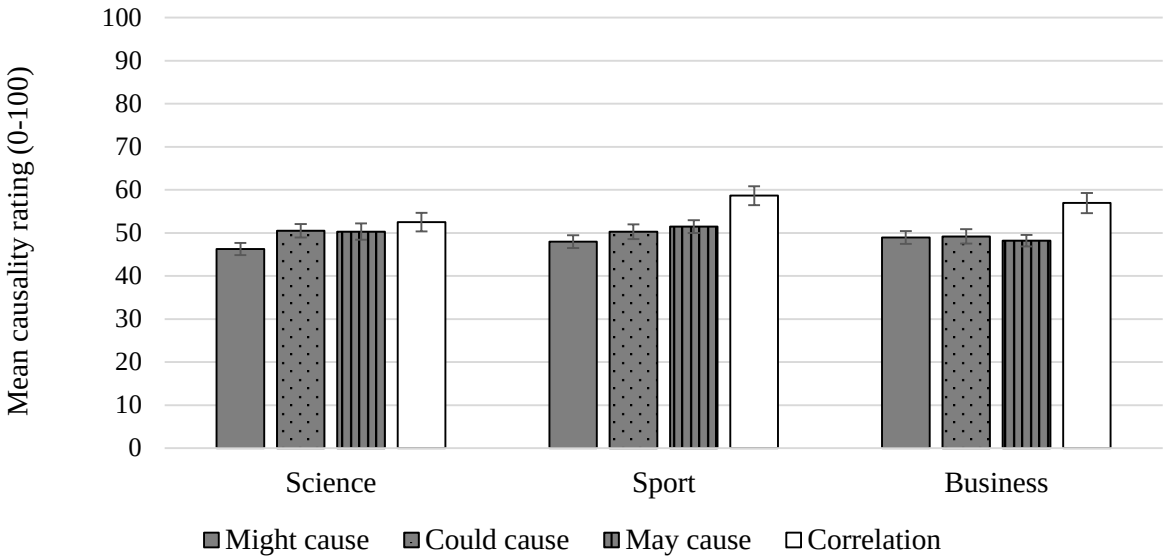

Figure 3. Mean causality ratings for the four categories of relationship as a function of topic in Experiment 3. Error bars show ±1 within subject standard error (Cousineau, 2005; Morey, 2008)

Discussion

Participants rated *might cause, may cause,* and *could cause* as significantly less causal than simple *correlation* statements. Thus, in general, modal verbs combined with causal expressions reduce the causal implications of statements, and they do so to such a degree that the resulting causal implication is less than that of correlation expressions. In other words, correlational phrases must carry some causal implication, as indicated by the moderately high rating scores. This result contradicts the coding scheme of Sumner et al. (2014) and the advice in the HealthNewsReview.org website, which suggest that simple correlational expressions are less causal than modified causal expressions. It also contradicts the predictions of usage guides which suggest that "may" should lead to more causal implications than "might".

Experiment 4

In Experiment 4 we tested all of the relational categories that we used in Experiments 1 to 3, and an additional category, *conditional correlation*. The conditional correlative condition used expressions that were correlative, such as "is associated with," combined with

a modal verb, such as “may”, as in “may be associated with.” This condition was included to test the hypothesis that the effects of the modal verbs seen in Experiment 3 generalised to correlative relationships as well as causal relationships.

For converging evidence, we used a different method of assessing causal implications compared to our previous experiments. In Experiment 4 participants ranked six forms of a given headline presented simultaneously, one form for each category of relationship, according to the degree of causal implication generated by each expression. Table 5 shows an example. Participants saw only two questions. The changes to the design were introduced to: (1) establish that our previous results generalised using other methods; (2) eliminate any carry-over effects arising from participants rating many headlines; (3) test the coding scheme described in Sumner et al. (2014) using a method analogous to its intended use (i.e., a method of ranking statements into one of six distinct categories).

Method

Participants. Five hundred and fifty-six participants were recruited using an online crowdsourcing platform (Prolific Academic). Fifty-seven participants were excluded from all statistical analyses because they failed to complete the task and 119 participants met the exclusion criterion (see below). The final sample size was 380 participants (225 females, 152 females (3 missing values); aged 16-67, M=28.65, SE=0.52). Sample size was determined according to an *a priori* power analysis based on the results of Experiments 1- 3 (using G*Power; Faul, Erdfelder, Lang, & Buchner, 2007). The smallest significant effect size was used (dz= 0.22; the comparison between *can cause* and *correlation* in Experiment 2); to achieve 90% power this gave a required sample size of N= 373 (with α=0.003; after correction for 15 comparisons).

Design and materials. Relationship category was the only factor in Experiment 4. There were six levels: direct cause, can cause, conditional cause, ambiguous, correlation and conditional correlation. The dependent measure was the causality ranking for each headline.

Thirty sentence frames were constructed. They covered a range of health and lifestyle-related topics (e.g. diet, pregnancy, mental health). Each frame was used in six forms corresponding to the six categories of relationship: *direct cause*, *can cause*, *conditional cause*, *ambiguous*, *correlation* and *conditional correlation* (see Table 5 and Supplementary Information). Across the 30 sentence frames each modal verb, *ambiguous* expression and *correlation* expression was used an equal number of times, and expressions were approximately counterbalanced. Because causal expressions are more varied in the news (see Table S1 Supplementary Information) we used more examples of *direct cause* expressions (consequently each causal expression was presented less frequently than the other expressions). Where *direct cause* expressions specified the direction of relationship (e.g. boosts, reduces) the same expression was used in the *ambiguous, correlation* and *conditional*

correlation sentences. For example "Dietary advice reduces saturated fat intake" was changed to "Dietary advice predicts *reduced* saturated fat intake" for the *ambiguous* sentence.

Table 5. Example stimuli from Experiment 4.

Relationship Category	Sentence
Direct cause	Being breast fed **results in** better behaviour in children
Can cause	Being breast fed **can result in** better behaviour in children
Conditional cause	Being breast fed **may result in** better behaviour in children
Ambiguous	Being breast fed **is linked to** better behaviour in children
Correlation	Being breast fed **is associated with** better behaviour in children
Conditional correlation	Being breast fed **may be associated with** better behaviour in children

Procedure. Participants were given instructions to rank sentences according to the degree of causal implication. They were told to place them in order from most causal at the top to least causal at the bottom. No feedback was provided on their responses.

The instructions contained two examples. Each used a single sentence frame expressed in four versions. The first sentence frame was, "Eating baked beans are [*expression*] to cause large elbows," and the second, "Cycling is [*expression*] to cause headaches". The expressions were: *very likely*, *likely*, *unlikely* and *highly unlikely*. Participants were told to position the statements so that the most causal headline (i.e., "baked beans are very likely to cause large elbows") was at the top, the next most causal headline below (i.e., "baked beans are likely to cause large elbows"), and the least causal (i.e. "baked beans are highly unlikely to cause large elbows") at the bottom. Following the example questions, participants were randomly assigned to two sentence frames, one for the first question and one for the second.

At the end of each question participants confirmed that they had ordered their statements from *most* causal to *least* causal. To ensure that participants read all of the headlines, each question was presented for a minimum of 90 seconds (there was no time limit on the first example question and 60 seconds for the second example question).

Exclusion criterion. We decided upon an exclusion criterion post hoc to remove participants who were ranking the statements at random. We reasoned that while statements ranked in the middle of the scale might differ across sentence frames (and therefore questions), those at the extremes would not. We therefore excluded participants who were inconsistent in their rankings of the most causal, or the least causal, across questions. For example, a participant who ranked *direct cause* as most causal in the first question but *ambiguous* as the most causal in the second question was removed (results of the analysis for the full sample is provided in the Supplementary Information).

We also reversed ranks for participants who appeared to have misread instructions and ranked statements from least to most causal, rather than *vice versa* (n=65). Ranks were reversed where participants had consistently placed *direct cause* as the least causal item (based on the results of Experiments 1-3 showing that *direct cause* is consistently rated as the most causal statement of relationship).

Results and discussion

To avoid carry-over effects from answering multiple questions we only analysed causality rankings for the first question (rankings for the second question were used purely for the exclusion criterion, see above). Figure 4 shows the mean causality rankings for each condition. A Friedman test revealed a significant overall effect of relationship category ($\chi^2(5)=900.92$, $p<0.001$). Follow-up Wilcoxon Signed-Rank tests revealed significant differences between all conditions (all $ps_{0.003} <0.001$; all rs > 0.2), with the exception of the comparisons between *conditional cause* and *ambiguous* ($p_{0.003} =0.08$; r = 0.09; B = 0.17) and *conditional cause* and *correlate* ($p_{0.003} =0.01$; r = 0.14; B= 2.5). Effect sizes were large for comparisons between direct cause and all other conditions and between conditional correlation and all other conditions (all rs >0.57). All other comparisons were in the small-medium or medium range.

The results generally support the ordering we observed in the previous experiments. However, there were two exceptions. The first is that *ambiguous* statements, such as "linked to", were significantly more causal than *correlation* statements, such as "associated with" ($p<0.001$; $r=0.22$; and Bayes factors showed decisive evidence in favour of the alternative hypothesis, B=5054), unlike Experiment 1, in which we did not find a difference (and the Bayes Factor from Experiment 1 indicated substantial evidence in favour of the null hypothesis [B=0.21] suggesting that this was not due to low power). The second is that *conditional cause* statements, such as "may result in", were not ranked lower than *correlation* statements, unlike Experiments 2 and 3, where we did find a difference. We suggest that the inconsistency across experiments can be explained by differences in the methodology (free choice *vs* ranking), different samples of participants across experiments (undergraduate Psychology students vs online recruitment) and different materials (see Supplementary Information). We discuss this further in the General Discussion. In the next section we use our results to redefine exaggeration and apply this definition to the reanalysis of Sumner et al. (2014, under review).

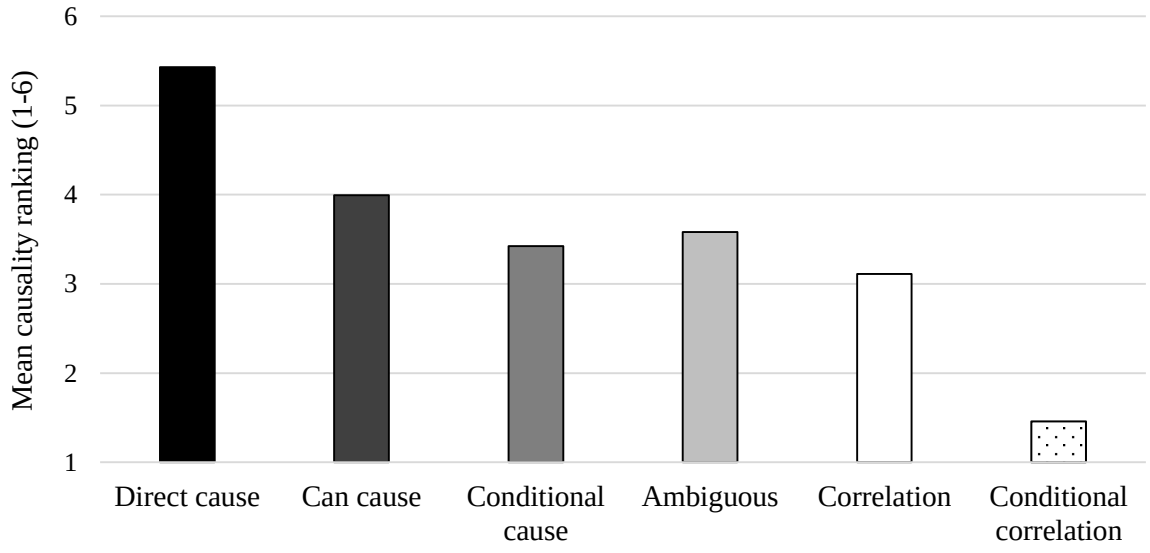

Figure 4. Mean causality rankings for each category of relationship in Experiment 4.

Reanalysis of Sumner et al. (2014, under review)

Sumner et al. (2014) analysed the association of exaggerations in news and health-related press releases issued by leading UK Universities. One of the foci for analysis was causal statements referring to research with correlational designs. Their results showed that 33% of press releases contained exaggerated causal claims and suggested that the majority of exaggeration in the news is already present in the preceding press release. However, Sumner et al. analysed exaggerated causal statements from correlational research using a seven level scale: *direct cause*; *can cause*; *conditional cause*; *ambiguous*; *correlation*; *statement of no relationship*; *no causal claim* (see Table S1, Supplementary Information). "Exaggerations" were defined as any increase in this scale relative to what was stated in the journal article. For example, a press release that used an ambiguous expression where the journal article used a correlation expression was classed as an exaggeration.

Our results suggest that Sumner et al. (2014, under review) over-estimated the rate of exaggeration, however. We found no evidence to support a categorical difference between *conditional cause*, *ambiguous* and *correlational* statements. We did not find that "may cause" is more causal than ambiguous statements such as "predicts" (Experiment 4), or consistently more causal than statements of correlation (Experiments 2, 3 and 4). Similarly, correlational statements were considered just as, or almost as, strongly causal as ambiguous phrases (Experiments 1 and 4). If readers do not reliably distinguish between these expressions, then one could argue that changes from one expression to another within these categories should not be considered exaggerations. We therefore re-analysed the data from both Sumner et al. (2014) and Sumner et al. (under review) using a scale in which *conditional cause*, *ambiguous* and *correlation* categories were grouped together into a single *moderate cause* category. All other aspects of the analysis were identical to that described in Sumner et al.

Results (Sumner et al., 2014)

Reducing the number of causal categories from seven to five necessarily reduced the calculated rate of exaggerations: 19% (95% CIs 14% to 25%) of press releases and 32% (95% CIs 24% to 41%) of news contained more strongly causal main statements about correlational results than those present in the associated journal article. This compares to 33% of press releases and 39% of news under the original analysis (95% CIs 26% to 40% and 31% to 49%, respectively). However, the association between exaggerated news and exaggerated press releases remained clear (see Figure 5.A). The odds of exaggerated statements in the news was 12 times higher (95% CIs 4.7 to 29.7) when press release statements were exaggerated; 72% of causal claims in the news were exaggerated when the press release contained exaggeration (95% CIs 54% to 85%) compared to 19% when it did not (95% CIs 12% to 27%; difference 53%, 95% CI 49% to 78%).

The second main result – that there was no clear evidence for an association between exaggeration and improved news uptake – also remained (Figure 5.B): 72/146 (49%) press releases without exaggeration had news uptake compared with 22/35 (63%) press releases with exaggeration (95% CIs of the difference -5% to 31%). For the press releases that did generate news, non-exaggerated main causal claims were associated with 2.9 news stories per press release, whereas exaggerated causal claims were associated with 2.4 news stories per press release (95% CIs of the difference -0.7 to 1.4).

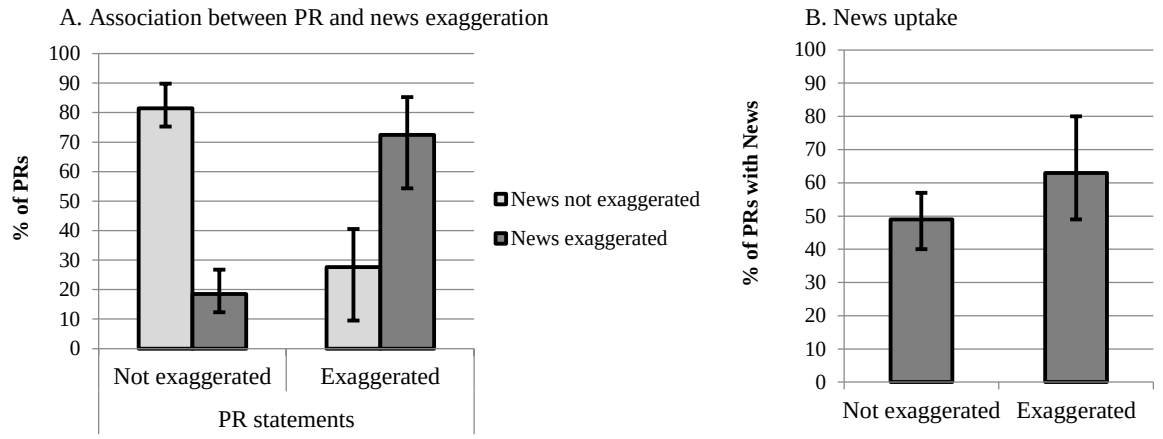

Figure 5. Reanalysis of causal claims from Sumner et al. (2014) with five categories of relationship. Panel A shows the association between exaggeration of statements in the news and press releases. Panel B shows news uptake of press releases with and without exaggerated statements. Error bars are bootstrapped 95% confidence intervals.

Results (Sumner et al., under review)

We also reanalysed Sumner et al.'s (under review) latest results in which they replicate their previous findings for press releases from eight prominent science and medical journals (Lancet, British Medical Journal (BMJ), Science, Nature, Nature Neuroscience, Nature Immunology, Nature Medicine, and Nature Genetics). Similar to academic press releases, Sumner et al. showed that exaggerated causal statements in journal press releases predicted exaggerated statements in the news (odds ratio 10.9, 95% CIs 3.9 to 30.1) but were not associated with increased news coverage.

Reducing the number of causal categories reduced the calculated rate of exaggerations: 13% (95% CIs 8% to 19%) of press releases contained exaggerated causal claims, compared to 21% when causal claims were split into 7 categories. Likewise, the proportion of news stories with exaggerated causal claims dropped from 38% with 7 categories to 31% (95% CIs 21% to 42%). Again, the association between exaggerated news and exaggerated press releases remained clear. The odds of exaggerated statements in the news was 7.3 times higher (95% CIs 2.5 to 21.4; Figure 6.A) when press release statements

were exaggerated (74%, 95% CIs 52% to 89%) than when they were not (25%, 95% CIs 17% to 35%).

There was still no evidence that exaggeration was associated with improved news uptake (Figure 6.B). With five categories of relationship, 71/142 (50%) press releases without exaggeration had news compared to 15/22 (68%) press releases with exaggeration (95% CIs of the difference -4% to 38%). For the press releases that did generate news, the average number of news stories per press release was 3.2 per non-exaggerated press release, and 2.1 for an exaggerated press release (95% CIs of the difference -2.0 to -0.1).

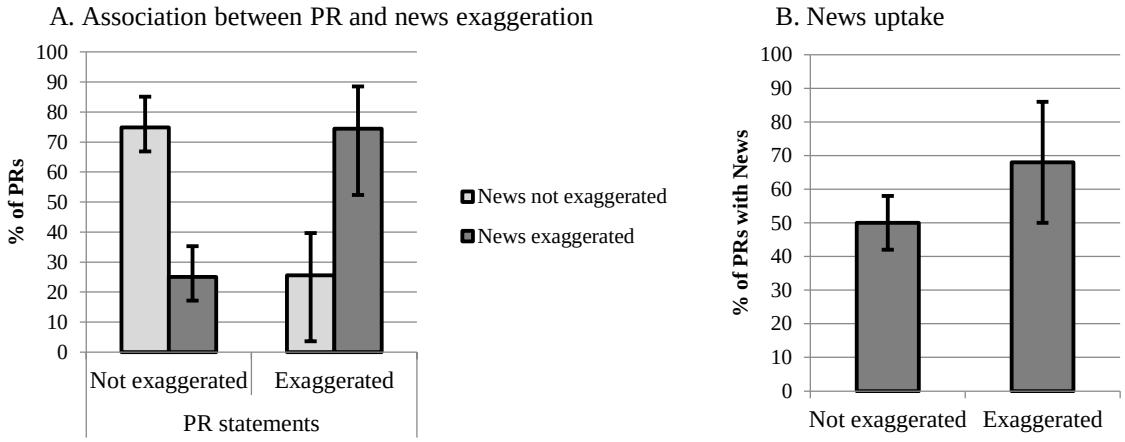

Figure 6. Reanalysis of causal claims from Sumner et al. (under review) with five categories of relationship. Panel A shows the association between exaggeration of statements in the news and press releases. Panel B shows news uptake of press releases with and without exaggerated statements. Error bars are bootstrapped 95% confidence intervals.

Discussion

Analysis using five categories of relationship necessarily reduced the number of exaggerations present in the press releases and news articles but it did not change Sumner et al.'s (2014, under review) main findings: there remained a strong association between exaggerations present in the news and press release statements, and there also remained no support for the intuitive idea that exaggerations in press releases should increase news uptake.

General Discussion

The science writer is faced with the difficult task of conveying scientifically accurate information while at the same time making the language interesting, varied, and appealing to readers. Our study was conducted to test how readers understand the diverse range of causal expressions currently employed in the media (see Table S1, Supplementary Information) and to provide evidence-based advice about the consequences of using those expressions.

Causality ratings showed that different relational expressions communicate different degrees of causal implications. The most causal were those that were direct, such as "increases", and "makes". These expressions were rated consistently higher than the other expressions. When these same expressions were modified by "can", however, causality ratings dropped significantly. Other expressions reduced causality ratings even further. Modifying the causal expression with "may", "might", or "could" lowered ratings, as did ambiguous or correlational expressions such as "linked to*"* and "associated with". The causal judgements for these expressions, however, were not consistently different from one another across experiments. Finally, while there were differences in degrees of causality, the absolute value of even the lowest rated expression ("might cause") was considerably above floor level. This suggests that participants thought that all expressions were at least moderately causal.

Taken together, the results of this study indicate that readers distinguished between three categories of expression: *direct cause*; *can cause*; and *moderate cause*. Table 2 lists the expressions associated with each category. *Direct cause* and *can cause* expressions are as described and *moderate cause* expressions are shown by the conjunction of *conditional cause, ambiguous,* and *correlation* categories. We next discuss explanations of our findings before turning to the practical implications.

Differences across experiments

We observed some inconsistency across experiments in how people understand weakly causal and correlational expressions. One explanation is that we collected ratings in Experiments 1 to 3 but rankings in Experiment 4 and that the difference between these procedures gave rise to the difference in findings. There are two main differences. The first is that ranking prevents participants from assigning the same score to multiple expressions, whereas rating does not. Ranking therefore requires participants to process the sentences sufficiently deeply to make a choice between expressions, whereas rating does not. This implies that ranking is more sensitive at detecting small differences in interpretation than rating. The second difference is that ranking uses a non-parametric scale, whereas rating uses a parametric scale. Thus large differences in interpretation would be curtailed in the ranking procedure, which might make it less sensitive. In short, the differences in procedure work against each other in terms of sensitivity and it is not possible to say that one method is more sensitive overall than the other. Furthermore, the pattern of our findings could not easily be explained by differences in sensitivity across paradigms. While a more sensitive rankings procedure might explain why we observed a difference between *ambiguous* and *correlation* expressions in Experiment 4 but not Experiment 1 (although the Bayes factor from Experiment 1 suggests otherwise), it cannot explain why we observed a difference between *conditional cause* judgements and *correlate* conditions in Experiments 2 and 3 but not in Experiment 4.

The inconsistency across experiments might therefore be explained by other factors, such as different sentence frames and participant samples. Prior knowledge regarding the relationship between two variables plays some role in causal inference judgments and its effects will vary across individuals and sentence frames. In the extreme, judgements of highly plausible or implausible causal relationships will be insensitive to changes in the relational

expression, since prior knowledge will override the new information. For example, “High fat food *is linked to* weight loss” would be insensitive to the relational expression because it is strongly inconsistent with prior knowledge. Since judgements about the plausibility of particular relationships will vary across individuals, using different sentence frames or different samples of participants across experiments will lead to variability in causal judgements, as we observed.

Interestingly, we observed differences across experiments in the weakly causal expressions, such as “associated with”, but not the strongly causal expressions, such as “boosts”. This could be because weak expressions convey a large range of potential relationships (i.e. they are uninformative), leaving participants with no option but to use their own knowledge to make a judgement, whereas strong expressions convey a very narrow range, allowing participants to abandon their prior knowledge and use the new information contained in the expression. For example, “High fat food *is linked to* weight loss” provides very little information about the strength of the causal relationship between high fat food and weight loss, and so the participant must rely on their knowledge to judge how strong the relationship is likely to be. “High fat food *boosts* weight loss”, on the other hand, convinces the reader that this a strong and directly causal relationship, therefore removing the necessity to use prior knowledge in order to interpret the sentence. Consequently, weakly causal expressions are more sensitive to variation in prior knowledge across individuals than strongly causal expressions, and so are much more prone to cross-experimental differences.

Educational background

While we observed robust differences in causal ratings across relational expressions, we did not we did not find effects of educational background. In this respect our results are similar to previous studies. Bleske-Rechek et al. (2015) found no association between education level and the likelihood of selecting correct statements when they split participants

according to whether or not they had a Bachelor's degree. Similarly, Norris et al. (2003) found that the number of science courses taken was not predictive of performance, and reported that undergraduate students did not perform better on their tasks compared to a sample of high school students (Norris & Phillips, 1994).

One explanation for this is that formal science education played no role in how participants understood the headlines. Instead, they might have been using more general, folk notions of causality and correlation. People might not know that random assignment of participants to conditions is the *sine qua non* of an experiment, say, but they nonetheless understand the difference between events that are causally linked and those that are merely associated. This is shown by work in other areas of cognitive psychology. For example, young children use causality in their representation of folk biology and physics (e.g., Inagaki & Hatano, 2002), causality underpins concepts and categories (Murphy & Medin, 1985), and people use knowledge differently when they believe it is causal compared to when they believe it is correlational (e.g., Rehder & Hastie, 2001). The absence of an effect of scientific education could be because the knowledge that was used to make responses was not grounded in science.

Linguistic sources of causal meaning

The variation in causal inference arose because different language was employed across conditions. We suggest that participants extracted the meaning using three linguistic sources of information.

The first is the interaction between lexical content and syntactic construction. When any verb is used actively the resulting meaning involves causation (e.g., Pickering & Majid, 2007). For example, "John kicked Bill" means that John was the cause of the kicking action on Bill. When verbs that express particular changes in state are used, such as "increases" or "boosts", together with appropriate subjects and objects, such as "high-fat food" and "weight

loss", the result communicates a strong causal relationship between subject and object. The causal inference in these cases is a combination of the lexical content of the verb, particular predicates, and an active voice construction.

Causal meaning of ambiguous or correlational expressions, such as "is linked to", was likely derived from a slightly different source, however. The weak or non-existent causal relationship cannot be lexically specified because cause and correlational relationships are equally consistent with the literal meaning of correlational expressions. For example, since cause and correlation are both links, "is linked to" cannot preclude a causal relationship. What might be happening instead is that the non-causal relationship arises through a conversational implicature (Grice, 1989). Since the writer chose to use a weak expression, such as "is linked to", and they were in a position to utter a stronger expression, such as "results in", the reader is licensed to infer that the stronger expression does not apply, that is, the writer meant that it is *not* the case that "results in" is an appropriate description of the relationship. An implicature account is given extra weight by noting that it is possible to defease the meaning of the correlational expressions without generating unacceptable utterances (the hallmark of conversational implicatures). For example, "High fat food *is linked to* weight loss; in fact, it is causally linked to weight loss," is acceptable. In contrast, direct cause expressions (a literal meaning) cannot be defeased in the same way: "High fat food *boosts* weight loss; in fact, it is not causally linked to weight loss" is infelicitous.

Finally, compositional mechanisms could also have contributed to causal meaning. In our experiments, participants judged "can VERB" expressions to be weaker than simple "VERB" expressions. Intuitively the "can VERB" construction weakens any kind of epistemic claim, not just those associated with causality. To see this, compare "Nitrosyl chloride can mollitate benzene," with "Nitrosyl chloride mollitates benzene." In these examples, the "can VERB" statement feels weaker than the simple statement, even though

"mollitate" is a nonword and so cannot be lexically associated with causation. Thus, knowledge of this linguistic construction might have caused participants to rate *can cause* expressions as less causal than *direct cause* expressions. Exactly why "can" conveys this meaning is difficult to say. Literally, "can" adds no relevant meaning to the unmodified verb form (both sentences communicate that the subject is able to perform some action on the object). As with the lexical communication of causality, it is possible that the additional meaning arises from a conversational implicature. In this case, a manner implicature, in which the addition of unnecessary material ("can") makes the reader question why the writer did not use the unmodified verb form (the reason being that the writer was not confident enough about the relationship).

We have suggested three linguistic sources of causal meaning for the statements we presented. This list is far from exhaustive but we hope it presents a starting point for other researchers to identify the psycholinguistic mechanisms behind inferring causal and correlation (future work may be able to link this study to more established research on causality in language, such as implicit causality effects, e.g., Stewart, Pickering, & Sanford, 1998, or causal connectedness, e.g., Myers, Shinjo, & Duffy, 1987).

Practical implications

The current research suggests that readers distinguish three groups of causal expressions: *direct cause*; *can cause*; and *moderate cause*. These results have implications for science writers. We make the following recommendations: (1) writers should use *direct cause* expressions when conveying findings from rigorous experimental designs, (2) insert the word "can" prior to *direct cause* expressions when conveying uncertainty about experimental findings (e.g., where an intervention uses a small sample size or a new drug is only tested on healthy participants), and (3) use *moderate cause* expressions when discussing observational findings. A caveat to these recommendations, however, is that whatever the relational

expression, juxtaposing two variables in a headline implies at least a moderate degree of causality between them. Writers should be aware that this is the likely effect of their headlines and consider taking appropriate steps to mitigate the potential problems (e.g., including statements in the news story that explicitly deny evidence of a causal relationship).

This advice applies not only to journalists and those writing the headlines (such as sub-editors) but also to other science writers, such as press officers and academics. Press officers may be particularly important for accurately conveying the findings of health research to the public. Press releases have become a dominant link between health research and the media (Kiernan, 2006; Taylor et al., 2015; Williams & Clifford, 2009) and exaggeration in the media appears to be strongly associated with exaggeration in the preceding press release (Sumner et al., 2014, under review). Academics should also be aware of conflating correlation with causation. Although academic journal articles are peer-reviewed, they have been shown to frequently contain misleading information, with up to 53% of abstracts containing exaggerated causal language (Cofield, Corona, & Allison, 2010; Gonon, Bezard, & Boraud, 2011; Lazarus, Haneef, Ravaud, & Boutron, 2015; Yavchitz et al., 2012). We therefore recommend that science writers follow the above advice to ensure that the causal language they use is not exaggerated.

Conclusion

The results of the current study show that readers distinguish between three categories of relational expression: *direct cause*; *can cause*; and *moderate cause*. Based on these results we suggest that journalists, editors, press officers and academics, modify their causal language, using these categories, to suit the study design of the research being discussed. Although we cannot claim that accurately reported science headlines are sufficient for the public to make well-informed choices related to their health (audience responses are complex and multiply determined; Kitzinger, 2004; Sturgis & Allum, 2004), we do argue that they are

a necessary starting point. Following the guidelines we present here should reduce the ambiguity present in press releases and, of most concern, news stories.

Acknowledgements

This project was supported by ESRC Grant ES/M000664/1 and ESRC grant ES/M500422/1. R. Adams was principally responsible for all parts of the paper. S. Vivian-Griffiths conducted the reanalysis of the data in Sumner et al. (2014, under review). P. Sumner, A. Williams, J. Boivin, C. Chambers., and L. Bott made substantial contributions to all parts of the paper. A. Barrington contributed to the design and data collection for Experiments 2 and 3. L. Bott was senior author and oversaw the project. We thank the following undergraduate students for contributions to Experiment 1 and pilot work leading up to the project: Laura Benjamin, Cecily Donnelly, Cameron Dunlop, Rebecca Emerson, Rose Fisher, Laura Jones, Olivia Manship, Hannah McCarthy, Naomi Scott, Eliza Walwyn-Jones, Leanne Whelan, Joe Wilton.

References

Bleske-Rechek, A., Morrison, K. M., & Heidtke, L. D. (2015). Causal Inference from descriptions of experimental and non-experimental research: Public understanding of correlation-versus-causation. *The Journal of General Psychology*, *142*, 48-70.

Bosely, S. (2014, May 15). BMJ rejects scare stories on statins following plea from Oxford professor. Retrieved from http://www.theguardian.com/society/2014/may/15/statins-bmj-statement-professor-collins-side-effects

Brechman, J., Lee, C. J., & Cappella, J. N. (2009). Lost in translation? A comparison of cancer-genetics reporting in the press release and its subsequent coverage in the press. *Science Communication, 30,* 453-74

Budescu, D. V., & Wallsten, T. S. (1985). Consistency in interpretation of probabilistic phrases. *Organizational Behavior and Human Decision Processes, 36,* 391-405.

Cofield, S. S., Corona, R. V., & Allison, D. B. (2010). Use of causal language in observational studies of obesity and nutrition. *Obesity Facts, 3,* 353-356.

Cooper, B. E. J., Lee, W. E., Goldacre, B. M., & Sanders, T. A. B. (2012).The quality of the evidence for dietary advice given in UK national newspapers. *Public Understanding of Science, 21,* 664–673.

Cousineau, D. (2005). Confidence intervals in within-subject designs: A simpler solution to Loftus and Masson's method. *Tutorials in Quantitative Methods for Psychology*, *1*, 42–45.

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. doi:10.3389/fpsyg.2014.00781

Dor, D. (2003). On newspaper headlines as relevance optimizers. *Journal of Pragmatics*, *35*, 695-721.

Ecker, U. K., Lewandowsky, S., Chang, E. P., & Pillai, R. (2014). The effects of subtle misinformation in news headlines. *Journal of Experimental Psychology: Applied*, *20*, 323– 335.

Ecker, U. K., Swire, B., & Lewandowsky, S. (2014). Correcting misinformation — A challenge for education and cognitive science. In *Processing Inaccurate Information: Theoretical and Applied Perspectives from Cognitive Science and the Educational Sciences*. MIT Press.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175–91

Gonon, F., Bezard, E., & Boraud, T. (2011). Misrepresentation of neuroscience data might give rise to misleading conclusions in the media: The case of attention deficit hyperactivity disorder. *PLoS One*, *6*, e14618.

Grice, H. P. (1989). *Studies in the Way of Words*. Cambridge, MA: Harvard University Press.

Haneef, R., Lazarus, C., Ravaud, P., Yavchitz, A., & Boutron, I. (2015). Interpretation of results of studies evaluating an intervention highlighted in Google health news: A cross-sectional study of news. *PloS One*, *10*, e0140889.

Health and Social Care Information Centre. (2009). NHS Immunisation Statistics - England, 2008-09. Health and Social Care Information Centre.

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/

Inagaki, K., & Hatano, G. (2002). *Young Children's Naïve Thinking about the Biological World*. New York: Psychology Press.

Kennedy, G. (2002). Variation in the distribution of modal verbs in the British National Corpus. In *Using Corpora to Explore Linguistic Variation*, Edited by: Reppen, R., Fitzmaurice, S., & Biber, D. 73–90. Amsterdam: Benjamins.

Kiernan, V. (2006). *Embargoed Science*. University of Illinois Press.

Kitzinger, J. (2004). *Framing Abuse: Media Influence and Public Understanding of Sexual Violence Against Children*. Pluto Press.

Lazarus, C., Haneef, R., Ravaud, P., & Boutron, I. (2015). Classification and prevalence of spin in abstracts of non-randomized studies evaluating an intervention. *BMC Medical Research Methodology*, *15*.

Leveson, B. (2012). *An Inquiry into the Culture, Practices and Ethics of the Press*. (pp 22, 80, 690-91, 1803).

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*, 106-131.

Love, J., Selker, R., Marsman, M., Jamil, T., Dropmann, D., Verhagen, A. J., … & Wagenmakers, E.-J. (2015). JASP (Version 0.7)[Computer software].

Morey, R. D. (2008). Confidence intervals from normalized data: A correction to Cousineau (2005). *Tutorials in Quantitative Methods for Psychology*, *4*, 61–64.

Mueller, J. F., & Coon, H. M. (2013). Undergraduates' ability to recognize correlational and causal language before and after explicit instruction. *Teaching of Psychology, 40*, 288-293.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review, 92*, 289-316.

Myers, J. L., Shinjo, M., & Duffy, S. A. (1987). Degree of causal relatedness and memory. *Journal of Memory and Language, 26,* 453-465.

Norris, S. P., & Phillips, L. M. (1994). Interpreting pragmatic meaning when reading popular reports of science. *Journal of Research in Science Teaching*, *31*, 947-967.

Norris, S. P., Phillips, L. M., & Korpan, C. A. (2003). University students' interpretation of media reports of science and its relationship to background knowledge, interest, and reading difficulty. *Public Understanding of Science, 12,* 123-145.

Pickering, M. J., & Majid, A. (2007). What are implicit causality and consequentiality? *Language and Cognitive Processes*, *22*, 780-788.

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

Rehder, B., & Hastie, R. (2001). Causal knowledge and categories: The effects of causal beliefs on categorization, induction, and similarity. *Journal of Experimental Psychology: General*, *130*, 323.

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.

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

Schwitzer, G. (2010). Covering medical research: A guide to reporting on studies. *Association of Healthcare Journalists*, Columbia MO.

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

Stewart, A. J., Pickering, M. J., & Sanford, A. J. (1998). Implicit consequentiality. In Proceedings of the Twentieth Annual Conference of the Cognitive Science Society (pp. 1031-1036). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Straight Statistics and Sense About Science. (2010). Making sense of statistics. *Straight Statistics and Sense About Science,* London. http://www.senseaboutscience.org/data/files/resources/1/MSofStatistics.pdf

Sturgis, P., & Allum, N. (2004). Science in society: Re-evaluating the deficit model of public attitudes. *Public Understanding of Science, 13,* 55–74.

Sumner, P., Vivian-Griffiths, S., Boivin, J., Williams, A., Bott, L., Adams, R. C., …& Chambers, C. D. (under review). Exaggerations and caveats in press releases and health-related science news.

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

Taylor, J. W., Long, M., Ashley, E., Denning, A., Gout, B., Hansen, K., ... & Wojtowicz, A. (2015). When medical news comes from press releases—A case study of pancreatic cancer and processed meat. *PloS One, 10*, e0127848.

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.

Woloshin, S., Schwartz, L. M., Casella, S. L., Kennedy, A. T., & Larson, R. J. (2009). Press releases by academic medical centers: Not so academic? *Annals of Internal Medicine, 150,* 613-618.

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.