Journal of Experimental Psychology: Applied

Can Journalistic “False Balance” Distort Public Perception of Consensus in Expert Opinion?

Derek J. Koehler

Online First Publication, January 11, 2016. http://dx.doi.org/10.1037/xap0000073

Journal of Experimental Psychology: Applied

© 2016 American Psychological Association
1076-898X/16/$12.00 http://dx.doi.org/10.1037/xap0000073

Can Journalistic "False Balance" Distort Public Perception of Consensus in Expert Opinion?

Derek J. Koehler
University of Waterloo

Media critics have expressed concern that journalistic "false balance" can distort the public's perceptions of what ought to be noncontroversial subjects (e.g., climate change). I report several experiments testing the influence of presenting conflicting comments from 2 experts who disagree on an issue (balance condition) in addition to a complete count of the number of experts on a panel who favor either side. Compared with a control condition, who received only the complete count, participants in the balance condition gave ratings of the perceived agreement among the experts that did not discriminate as clearly between issues with and without strong expert consensus. Participants in the balance condition also perceived less agreement among the experts in general, and were less likely to think that there was enough agreement among experts on the high-consensus issues to guide government policy. Evidently, "false balance" can distort perceptions of expert opinion even when participants would seem to have all the information needed to correct for its influence.

Keywords: judgment and decision making, heuristics and biases, perceived consensus, distribution of expert opinion

Citizens in a democratic society are expected to have, and to express in polls, opinions on a broad array of complex topics, from health (should schoolchildren be required to be vaccinated?) to economics (would increasing the minimum wage decrease employment?) to foreign policy (would the Syrian people benefit from Western military intervention?). The typical member of the public, however, lacks the time, knowledge, and access to directly evaluate the relevant evidence on such issues. Instead, they must rely on the opinions of experts who have, in fact, evaluated the evidence. Members of the public, for that matter, do not typically interact directly with the experts, but rather learn about prevailing expert opinion through the news media (e.g., Wilson, 1995).

Journalists, then, play a crucial role in distilling and distributing expert opinion to the public. They face a quandary when, as is often the case, the experts themselves disagree in their interpretation of the evidence pertaining to an issue of interest to the public. In such situations, journalists generally strive to convey a balanced account of the competing expert opinions (Bennett, 1996; Dunwoody, 2005; Entman, 1989). This might entail, for example, obtaining comments from at least one expert with an opinion on either side of a contentious issue.

Media critics have expressed concern that the journalistic standard of balance is now so ingrained in reportage that it is reflexively applied even to issues for which the weight of evidence overwhelmingly supports one "side" (e.g., Boykoff & Boykoff, 2004; Dearing, 1995; Stocking, 1999), a phenomenon that has been labeled *false balance*. Climate change is the most salient contemporary example of such an issue: Although approximately 97% of scientific reports (and the experts themselves) support the claim of anthropogenic global warming (Anderegg et al., 2010; Cook et al., 2013; Doran & Zimmerman, 2009), many TV and print media reports continue to include reference to, or comments by, climate change denialists (Boykoff, 2013; Greenberg, Robbins, & Theel, 2013). Communication science researchers have expressed concern that false balance in media coverage of climate change has contributed to distorted public perceptions (Boykoff & Boykoff, 2004; Corbett & Durfee, 2004). Contrary to the near unanimity among climate scientists regarding man-made global warming, for instance, a substantial segment of the public does not believe it is happening (Gallup, 2014; Leiserowitz et al., 2014; Weber & Stern, 2011).

In the present research, I investigate how "balanced" presentation of conflicting comments from experts influences perceptions of the overall distribution of expert opinion on an issue. The focus here is not on how such presentation affects a person's own opinion on the topic, but rather his or her impression of the extent to which the experts agree or disagree in their opinions on the issue.

Can journalistic false balance distort public perception of the distribution of expert opinion? Although this would seem a ripe topic for psychological research, relatively little empirical work has been conducted to answer this question. Most relevant is that by Dixon, Clarke, and colleagues, who in several studies have investigated participants' perceptions of expert opinion on the now-discredited link between vaccines and autism (Clarke, Dixon, Holton, & McKeever, 2015; Dixon & Clark, 2013; Dixon, McKeever, Holton, Clarke, & Eosco, 2015). Their studies presented participants with newspaper articles on the topic modified either to present only the dominant expert view that there is no link between vaccines and autism, or to present (falsely) balanced coverage that

This research was supported by a grant from the Natural Sciences and Engineering Research Council of Canada.

Correspondence concerning this article should be addressed to Derek J. Koehler, Department of Psychology, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada. E-mail: dkoehler@uwaterloo.ca

includes reference to claims that there is such a link. Compared with the one-sided or other control conditions, participants presented with the balanced articles perceived more disagreement and less certainty among the experts regarding the vaccine-autism link.

Results from two other studies are more mixed. Corbett and Durfee (2004) found that perceptions of scientific certainty were affected when a study result that was the focus of a newspaper article was balanced by reference to other studies with opposite results (what the authors called "context"), but presentation of comments from a disagreeing expert (what the authors called "controversy") did not have a reliable effect. Jensen and Hurley (2012) compared effects of presenting pairs of news articles on a given scientific controversy, which offered either similar or disagreeing viewpoints on the issue, and found mixed results: For one issue, the disagreeing viewpoints increased perceived uncertainty but for another issue it actually decreased uncertainty. As in the studies by Dixon, Clarke, and colleagues, these other two previous studies compared effects of exposure to modified news articles that differed substantially in content, and gave participants no other information on which to base their judgments.

Though somewhat slim, it seems the current available evidence does support the idea that false balance can, at least in some circumstances, exert a distorting influence on public perceptions of expert opinion. Awareness of and concern over the possible distorting influence of false balance has grown within the field of journalism as well. For instance, following a report suggesting that some BBC science coverage had succumbed to false balance (BBC Trust, 2011), the BBC now offers workshops for reporters to raise awareness of the problem (BBC Trust, 2014). Awareness alone, however, does not provide a remedy. How are responsible journalists supposed to maintain standards of accuracy and objectivity when reporting on complex issues such as climate change, while avoiding the problem of false balance? Even if a very large majority of climate scientists has reached a consensus on the issue, for example, can the journalist entirely ignore the small minority of scientists (and the larger minority of the public) who deny the existence of global warming?

A sensible prescription for journalists might be to convey to the audience the "weight of evidence" on either side of a contentious issue (Dunwoody, 2005; Wilson, 2000). This way, both sides of the issue are presented (e.g., with comments from Experts A and B, who disagree on the issue), but information about the distribution of expert opinion on the issue is also reported (e.g., Expert A's views are supported by the large majority of fellow experts in the field while Expert B is part of a small dissenting minority on the issue). According to Dunwoody (2005), weight-of-evidence reporting requires journalists:

> . . . find out where the bulk of evidence and expert thought lies on the truth continuum and then communicate that to audiences. Reporters are still responsible for capturing points of view accurately (objectivity) and for sharing with audiences the existence of more than one contrasting point of view (balance). But added to that mix would be information about which point of view has captured the hearts and minds of the majority of experts, information about where they think the truth lies at that moment.

The BBC Trust Conclusions (BBC Trust, 2014) prescribes a similar approach:

> The Trust wishes to emphasize the importance of attempting to establish where the weight of scientific agreement may be found and make that clear to audiences (p. 2).

As does NPR in the Fairness section of its Ethics Handbook (NPR, 2012):

> Our goal is not to please those whom we report on or to produce stories that create the appearance of balance, but to seek the truth. . . . If the balance of evidence in a matter of controversy weighs heavily on one side, we acknowledge it in our reports (p. 19).

These remedies and prescriptions carry with them, implicitly, some psychological assumptions about how the perceptions of the news audience are influenced by, on the one hand, statements regarding the weight of evidence and, on the other, presentation of conflicting views. It is presumed that, ultimately, the audience will use the weight of evidence information to adjudicate between conflicting expert opinions. Psychologically, however, it has been argued that people tend to neglect weight-of-evidence considerations (Griffin & Tversky, 1992), and more generally are less influenced by statistical summaries than by vivid individual cases or stories (e.g., Borgida & Nisbett, 1977). As such, presenting weight of evidence information may not prove to be an effective remedy for the potentially distorting influence of false balance.

With this theoretical starting point, the present research tested the hypothesis that, even in the ideal case of perfect weight-of-evidence information (namely, the number of experts in the population with opinions on either side of the issue), presentation of conflicting opinions (in the form of brief comments from disagreeing experts) can exert a distorting influence on perceptions of the distribution of expert opinion. In particular, it is hypothesized that (false) balance can make it more difficult for people to discriminate between issues where there is higher and lower levels of expert consensus.

Why might presentation of balanced comments representing conflicting expert opinions make it more difficult to distinguish issues with high expert consensus from those with low expert consensus? The present studies are focused on testing whether such an effect exists rather than with identifying its causes, but the hypothesis is motivated by several findings in psychology that might point to potential underlying mechanisms.

First, previous research has shown that people often have difficulty discounting evidence on the basis of knowledge of how it was selected (e.g., Koehler & Mercer, 2009). For instance, when participants are presented with an interview with a sadistic prison guard, they do not appear to adequately take into account knowledge that the guard was selected for the interview precisely because he is unusually cruel in his treatment of prisoners, and instead generalize his characteristics to all prison guards (Hamill, Wilson, & Nisbett, 1980). In the present studies, even when they are informed that a comment is being deliberately selected for presentation because it comes from an expert who disagrees with the large majority of fellow experts on the topic, participants may not adequately discount the weight they place on that comment in light of the method by which it was selected. If so, then conflicting "balanced" comments will receive more equal weight than they should in light of the overall expert consensus, leading to judgments that do not distinguish as sharply as would otherwise be the case between high and low consensus issues.

Second, it is possible that the juxtaposition of conflicting expert comments leads participants to impose a twofold partition on their representation of the distribution of expert opinion regarding the issue (e.g., the expert is either on one side of the issue or the other), which in turn has been shown to bias judgments toward placing equal weight on either side of the partition (e.g., in a probability distribution, toward 50%; Fox & Rottenstreich, 2003). In the present studies, then, this process would be expected to push judgments toward a midpoint between the two conflicting opinions. Because the process is based on the conflict between the two "balanced" opinions that are presented, it is expected to anchor judgments toward the midpoint even when participants attempt to incorporate in their judgments the information about the broader distribution of expert opinion, again with the consequence that, compared with a control condition in which the balanced comments are not presented, judgments will discriminate less sharply between issues with high and low expert consensus.

Third, the mere presence of disagreement in the two balanced expert comments may trigger the perception of conflict that in turn produces a sense of general uncertainty. Links between conflict detection and decreased confidence have been established in previous research (e.g., De Neys, Cromheeke, & Osman, 2011). In particular, conflict between cues can decrease confidence in intuitive predictions even when the conflicting cues have superior validity when compared with redundant but nonconflicting cues (Kahneman & Tversky, 1973). For example, presentation of conflicting, two-sided evidence (e.g., seeing arguments from both sides of a legal case) can diminish confidence in predictions based on that evidence compared to presentation of one-sided evidence (e.g., seeing arguments for just one side of a legal case), even when predictive accuracy is higher in the former condition because the predictions are based on more information (Brenner, Koehler, & Tversky, 1996). Conflicting expert comments might, for instance, make it more difficult to form a coherent representation (i.e., a "good story") of the issue in question, and consequently diminish confidence in any inferences made regarding that issue. If this diminished confidence in turn leads to less extreme judgments, the result would be judgments that distinguish less sharply than would otherwise be the case between the issues with high and low expert consensus.

Fourth, if conflicting expert comments are seen as "cancelling" one another and thereby being collectively nondiagnostic, it is possible that presentation of such nondiagnostic evidence could "dilute" the impact of other, more diagnostic weight-of-evidence information regarding the overall distribution of expert opinion. Past research has established such a dilution effect, in which judgments based on a mix of diagnostic and nondiagnostic evidence are less extreme (and hence less discriminating) than judgments based on the diagnostic portion of the evidence alone (Nisbett, Zukier, & Lemley, 1981; Troutman & Shanteau, 1977).

Overview of Studies

The studies reported here shared a common design. For a given topic or issue, participants were presented with a table summarizing the views of a panel of experts, which indicated how many (or what percentage) of the experts gave a positive, neutral, or negative evaluation of the target topic. This information was drawn from real expert panel data. (Specifically, depending on the study, the table showed how many film critics gave positive, mixed, or negative evaluations of a particular movie, or how many economics experts agreed, were uncertain, or disagreed with a statement regarding a particular economic issue.) In a control condition, participants saw only the summary table. In the "balance" condition, the table was supplemented with a comment provided by one expert on either side of the issue, that is, one expert who had given a positive evaluation and one who had given a negative evaluation. Again, actual expert comments were used. It should be emphasized that the method of selecting these comments was made clear to participants, namely that one comment was selected from among the set of experts in the table who gave a positive evaluation and one from among those who gave a negative evaluation.

Within a study, several issues were presented that varied (either within- or between-subjects, depending on the study) in the level of consensus among the experts. Some issues had high consensus, with the majority of experts giving positive evaluations and only a small minority giving negative evaluations; other issues had low consensus, with experts being more divided in their opinions. Participants gave ratings that reflected their perceptions of the distribution of expert opinion on the issue. A major focus in analysis of the results is how sharply their ratings discriminated between high- and low-consensus issues. Because participants were asked to evaluate the distribution of expert opinion (i.e., not to give their own opinion), normatively we should expect their ratings to effectively discriminate between issues which objectively had high versus low expert consensus. The primary hypothesis was that, compared with the baseline level of discrimination achieved in the control (table only) condition, participants in the balance condition (who saw two conflicting expert comments in addition to the table) would give ratings that less clearly discriminated between issues for which there was high and low expert consensus. The current studies offer a strong test of this hypothesis by providing (in the table presented to all participants) precise numerical information regarding the distribution of expert opinion, effectively "stacking the deck" against observing any influence of presenting balanced (conflicting) expert comments.

Study 1

Stimuli

Stimuli were movies selected from the Metacritic web site, which aggregates reviews from professional film critics by assigning each review (and, ultimately, each movie) a numerical score between 0 and 100. Movies were selected based on the following criteria: released in 2013; a minimum of 30 reviews; English; not documentary, animated, children's, or sequel. Two types of movies from this set were selected, based on the review scores: "good" movies for which there was critical consensus in the form of largely positive reviews, and "mediocre" movies for which the critics were more divided and gave mixed or neutral reviews. The "good" movies were selected by rank-ordering the set of 2013 movies by their aggregated Metacritic rating. (The very top film, *12 Years a Slave*, had only very positive reviews and so was excluded from the resulting set.) The aggregated Metacritic ratings for the "good" movie set ranged between 96 and 82. The "mediocre" movies were selected among films in the rank-ordered list with Metacritic ratings just below 50. (Movies that received the most negative ratings, at the very bottom of the list, tended not to have enough reviews to meet that selection criterion.) In this manner, a total of 16 movies, eight good and eight mediocre, were selected for pretesting.

Review excerpts, selected by Metacritic from the original reviews, were rank ordered for each movie on the Metacritic web site by the associated Metacritic score imputed to the film critic who wrote that review. For use in the balance condition, the most positive excerpt was selected from the top of the rank-ordered list and the most negative from the bottom. Occasionally the very top- or bottom-ranked excerpt was passed over, typically because it was not sufficiently evaluative in tone or focused too much on specific details of the movie that might allow participants to identify it. In addition, for use in a comparison condition ("typical" condition, described later), two moderate review excerpts were selected from the median position of the list or from the section of the list with ratings corresponding to the overall Metacritic rating for the movie. (Typically these two locations in the list coincided.) Again, specific excerpts were selected that (a) offered a clear evaluation of the movie and (b) did not rely too heavily on identifying features of the movie. The selected excerpts were stripped of references to the title of the movie and to the names of actors, directors, and writers; in these instances a generic phrase such as Movie Title or Actor Name was inserted (set off in square brackets) into the excerpt to maintain comprehensibility.

In a pretest, each of the excerpts (four per movie) from all 16 movies were then presented to participants in a random order. Participants were instructed:

> A popular web site that collects and combines reviews from dozens of critics has assigned a score between 0 (*very negative*) and 100 (*very positive*), categorized as follows: positive reviews: scores between 61 and 100; neutral/mixed reviews: scores between 40 and 60; negative reviews: scores between 0 and 39. Your task is to estimate the score assigned by the website to each review based on the excerpt you are presented from that review.

The pretest data ($N = 60$ participants recruited from same population as the main studies reported below) were used to select 10 movies (five good, five mediocre) for the main studies. Movies were selected that showed the clearest separation in ratings of the moderate relative to the extreme (positive or negative) review excerpts. The ratings of the moderate excerpts tended, as would be expected, to be more positive for the good than for the mediocre movies. That meant that the main challenge was to find good movies for which the moderate excerpts were at least a little less positive than the positive excerpt, and to find mediocre movies for which the moderate excerpts were at least a little more positive than the negative excerpt. Five movies of each type were selected on this basis.

Procedure

Three similar studies (Studies 1a–1c) were conducted using the movies stimuli. Common aspects of the procedure in the three studies are described here, with the few differences between them noted in the results section. In this and subsequent studies, sample size was determined in advance, with a set target of 100 participants per condition. All data exclusions, all manipulations, and all measures in this and the subsequent studies are reported.

Participants were U.S. residents recruited through Amazon Mechanical Turk, who were paid 25 cents for their participation in the online study, which typically took about 5 min to complete. The studies were completed online.

In all three studies, participants were presented with the 10 movies in a randomized order. Participants were instructed that the movies were selected from a web site that aggregates reviews by professional film critics, and that each review is assigned a score from 0 to 100, with scores between 0 and 39 categorized as negative, scores between 40 and 60 categorized as neutral or mixed, and scores of 61 to 100 categorized as positive. Participants were told that the web site assigns an overall score on this scale by combining the individual reviews, and that their task would be to estimate the score for each in a series of movies. For that purpose, they were told they would see a summary indicating how many positive, mixed/neutral, and negative reviews the movie had received. They were instructed, "Because you will not be given the exact reviewer scores, and because not all the reviewer scores are equally weighted, it is not possible to simply calculate the overall web site score assigned to the movie from the reviewer summary. But your task is to give your best estimate of that overall web site score based on the information provided." It was noted that movie titles and any other identifying information had been removed so that participants would have to base their estimates exclusively on the film critic rating information provided in the study.

Participants in the balance condition were further informed, "In case you find it helpful, you will also be provided with example excerpts from two reviews of the movie in question. **One excerpt is taken from the most negative review of the movie; the other excerpt is taken from the most positive review**."

For each movie, a table was presented showing a count of how many reviews that movie had received in each of the three categories (positive, mixed/neutral, negative). In the balance condition, the table was followed by an "excerpt from the most positive review" and then an "excerpt from the most negative review" of that movie. Participants in the control condition saw only the table. Based on this information, the participant provided an estimate of the movie's overall web site score on a sliding scale that ran from 0 (*very negative*) to 100 (*very positive*) with the slider initially positioned at the scale's center value of 50 (*mixed/neutral*). Appendix A shows an example screenshot of the task.

Once they had provided estimates of the web site score for all 10 movies, in Studies 1a and 1b participants completed a brief set of follow-up questions asking about their movie-watching and review-reading habits, along with one item measuring self-reported numeracy and another serving as an attention check. Other than the attention-checking question, the other measures are not discussed further here as they did not appear to qualify any of the primary results. In Study 1c, these follow-up items (with the exception of the attention-checking question) were replaced with items measuring perceptions of agreement among the expert reviewers, as discussed below.

Results

In all three studies, initial data were filtered to exclude that from participants who appeared not to be paying adequate attention to the task. To this end, data from participants who failed the attention check question were excluded. Data were also excluded from participants who gave ratings that did an unusually poor job of distinguishing the good from the mediocre movies. Specifically, using the primary dependent variable z-diff that is described in the next paragraph, data were excluded from participants whose ratings of the good movies did not fall, on average, at least one standard unit above the ratings of the mediocre movies (i.e., z-diff <1). Collectively, application of these two filtering criteria led to the exclusion of data from 11 of 195 participants in Study 1a, 25 of 300 participants in Study 1b, and 20 of 301 participants in Study 1c.

To test the hypothesis that presentation of balanced excerpts reduces discrimination between good and mediocre movies, a measure of discrimination accuracy (similar to d-prime in signal detection theory) was calculated. For each participant, ratings assigned to the 10 movies (five good and five mediocre) were standardized within subject, and then the mean difference between standardized scores assigned to the good and the mediocre movies was calculated by summing the scores for the good movies, subtracting from that the sum of the scores for the mediocre movies, and then dividing by five (effectively, the number of good–mediocre movie pairs). The resulting score, denoted z-diff, indicates how much more highly, on average, the participant rated the good movies than the mediocre movies, in standard units (i.e., relative to the standard deviation of his or her ratings). The main advantage of using this dependent variable over a simpler mean difference score in the unscaled ratings themselves is that it corrects for idiosyncrasies in scale use (represented by the mean and standard deviation of each participant's ratings). The unscaled mean ratings for good and mediocre movies are shown, by condition for each study, in Table 1.

A second dependent variable, calculated as the mean absolute deviation between the participant's rating of a movie and the actual score as published on the Metacritic web site, was calculated to more directly assess the accuracy of the ratings. This variable was used to determine whether any observed differences between conditions in discrimination (using z-diff) were associated with differential accuracy. Otherwise, it would be difficult to say whether a condition showing greater discrimination resulted in greater accuracy or simply reflected excessive extremity.

Study 1a compared the balance condition with a control (no excerpt) condition.[1] The difference in means on the z-diff variable was in the hypothesized direction, with ratings from the balance condition showing less discrimination of good from mediocre movies than the control condition, $F(1, 182) = 3.95$, $p = .048$, partial $\eta^2 = .021$. The analysis of mean absolute deviation showed that the reduced discrimination in the balance condition was also reflected in lower accuracy relative to the control condition, $F(1, 182) = 4.92$, $p = .028$, partial $\eta^2 = .026$. Figure 1 shows means (and associated standard errors) by condition on both variables.

Study 1b included an additional "typical" condition (presenting the moderate excerpts rather than the extreme positive and negative excerpts as in the balance condition) but was otherwise identical to Study 1a. In the typical condition, the two moderate excerpts were presented and labeled as "excerpt from a typical review" and "excerpt from another typical review." Analysis of the z-diff variable indicated a significant difference among conditions, $F(2, 272) = 3.51$, $p = .031$, partial $\eta^2 = .025$. Ratings in the balance condition discriminated between good and mediocre movies less well than in the other conditions. Tukey post hoc tests on the z-diff variable showed that the balance condition differed significantly from the control condition, $p = .046$; scores in the typical condition fell in between and did not differ significantly from either of the other conditions. On the mean absolute deviation measure, again ratings in the balance condition were on average the least accurate (had the highest mean absolute deviations), though this difference was not significant, $F(2, 272) = 2.51$, $p = .083$, partial $\eta^2 = .018$. Means of both variables are shown in Figure 2.

Study 1c differed from Studies 1a and 1b only in that, in Study 1c, participants first gave a rating associated with each excerpt they read (in the balance and typical conditions) before estimating the overall web site rating for the movie. Participants were instructed to estimate the review score assigned by the website based on the excerpt. The purpose of eliciting these ratings was to direct attention to the excerpts, to ensure they were being read by all participants. As such, these ratings were not analyzed and instead the focus remains on the final rating assigned to the movie based on all the critics' reviews, for comparability with Studies 1a and 1b. The only other difference in Study 1c was that supplementary measures were collected after the main rating task that gauged global perceptions of the reviewers' ratings across the entire set of movies (e.g., the extent to which the reviewers were seen, in general, to agree with one another).

Analysis of the z-diff measure in Study 1c showed that, once again, ratings in the balance condition discriminated good from mediocre movies the least well of the three conditions, $F(2, 278) = 8.96$, $p < .001$, partial $\eta^2 = .061$. Tukey post hoc tests showed a significant difference between the balance and control conditions, $p < .001$, but not between the balance and the typical conditions, $p = .092$, or between the typical and control conditions, $p = .097$. Figure 3 shows the means by condition.

Analysis of the mean absolute deviation measure did not reveal a significant effect of condition, $F(2, 278) = 1.66$, $p = .192$, partial $\eta^2 = .012$; unexpectedly, the mean absolute deviation in the control condition was actually higher than that in the balance condition, contrary to both of the previous studies. This unexpected result prompted a closer examination of the raw ratings estimates, which revealed four participants for whom the mean rating was less than 40. This is highly discrepant from the remaining participants in this and the previous studies, as the typical good movie had a rating around 90 and the typical mediocre one a rating around 50; a mean across all movies of 40 suggests these participants were using the rating scale in an unusual way that would likely exert a disproportionate influence on the mean absolute deviation measure (which depends on raw rather than standardized ratings, in contrast to the z-diff measure). All subsequent analyses reported here exclude those four outliers, all of whom were in the control condition. When the analysis of the mean absolute deviation variable was rerun excluding the outliers, the means in the control condition look more like that of the typical condition, both of which were lower than that in the balance condition. However, the overall effect of condition was still not significant, $F(2, 274) =$

[1] Due to a programming error, a third condition that was intended to present moderate excerpts instead presented the extreme positive and negative excerpts but labeled them as typical rather than extreme. Data from this condition were excluded from all analyses.

Table 1
Mean (and Standard Deviation) of Raw Movie Ratings for Good and Mediocre Movies in the Balance, Control, and Typical Conditions of Studies 1a–c

	Study 1a		Study 1b		Study 1c	
	Good	Mediocre	Good	Mediocre	Good	Mediocre
Balance	88.4 (7.7)	54.1 (9.7)	87.1 (9.4)	54.3 (9.2)	84.0 (8.4)	53.5 (9.3)
Control	90.0 (6.4)	54.6 (8.6)	89.6 (8.0)	56.4 (8.6)	88.9 (12.1)	53.8 (11.4)
Typical	—	—	89.8 (6.7)	55.1 (8.0)	87.1 (7.0)	55.0 (7.8)

1.63, $p = .199$, partial $\eta^2 = .012$. Figure 3 shows the mean absolute deviations by condition after excluding those additional four outliers.

After making their movie ratings in Study 1c, participants were asked to evaluate several general statements tapping their global perceptions (i.e., across the 10 movies) of the review information they had been presented with. They were instructed as follows: “Please rate your agreement with the statements below based on the review information for the movies you were presented with in the previous section of this survey.” Two of the statements concerned perception of agreement among the reviews: “The reviews of a movie often disagreed with one another” and “The film critics writing the reviews usually had similar opinions about a movie.” These two items correlated $r = -.57$ and were combined into a single “agreement” measure after reverse scoring the former item. Table 2 shows mean responses by condition for this and the other supplementary measures. Perceptions of agreement among the reviews differed by condition, $F(2, 274) = 92.6$, $p < .001$, partial $\eta^2 = .40$. Participants in the balance condition perceived less agreement among the reviews than those in either of the other conditions, $p < .001$ for both comparisons. Participants in the typical condition perceived more agreement among the reviews than did those in the control condition, $p < .001$, presumably because they were presented with two moderate excerpts with associated ratings quite close to one another.

Two other statements concerned the certainty or confidence with which the movie’s overall rating could be estimated based on the reviews: “It was difficult to be certain how good a movie was based on the reviews,” and “I was confident in my estimates of the movie’s overall website score.” These two items correlated $r = -.29$ and were combined into a single “certainty” measure (with the former item reverse scored) such that higher values indicated greater certainty or confidence in estimating the movie rating. The certainty measured also differed by condition, $F(2, 272) = 4.60$, $p = .011$, partial $\eta^2 = .033$. Certainty was significantly lower in the balance condition than in the typical condition, $p = .012$, and also somewhat though not significantly lower than in the control condition, $p = .064$.

The final supplementary item was: “The reviews gave me a good sense of whether I personally would enjoy the movie.” Because this “personal confidence” measure goes beyond the task of estimating the overall web site rating, which could be determined fairly directly from the summary of reviewer scores, it stands to reason that the excerpts would be potentially more informative for assessing whether the movie would be personally enjoyable. Although responses on this measure did not differ significantly between conditions, $F(2, 274) = 2.01$, $p = .136$, partial $\eta^2 = .014$, participants in the typical condition did express somewhat greater confidence in predicting their own personal preferences from the information provided than did those in either the control or the balance conditions.

The primary finding from Studies 1a–c was that discrimination of good (high expert consensus) from mediocre (low expert consensus) movies, as indexed by the z-diff measure, was reduced in the balance condition relative to the control condition. Presentation of balanced, conflicting comments generally made it more difficult to distinguish movies the critics agreed were good from those the critics did not agree were good. It should be acknowledged that the effect size was not large, and in some cases the differences did not achieve statistical significance at $p = .05$. That said, the studies

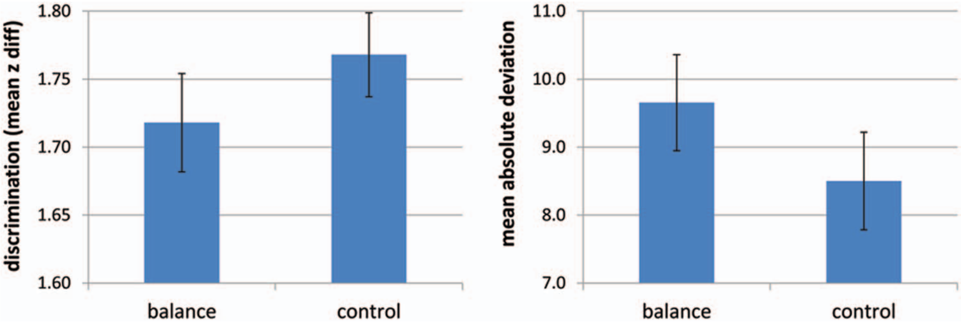


Figure 1. Mean discrimination (separation in standardized estimated ratings between good and mediocre movies) and absolute deviation (distance between estimated and actual web site movie ratings) by condition in Study 1a. Note: Error bars indicate 95% confidence intervals. See the online article for the color version of this figure.

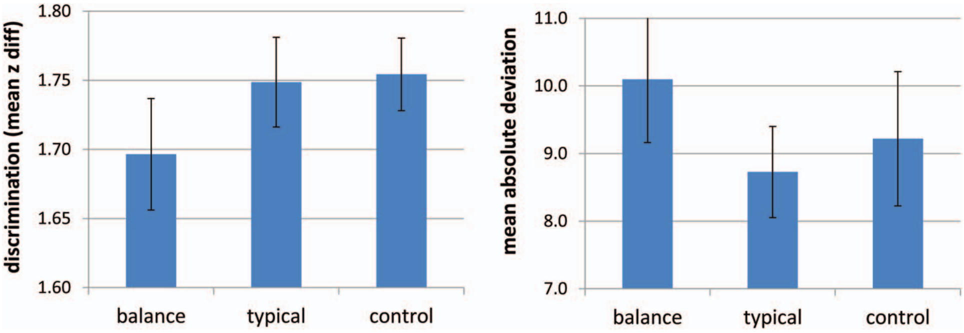


Figure 2. Mean discrimination (separation in standardized estimated ratings between good and mediocre movies) and absolute deviation (distance between estimated and actual web site movie ratings) by condition in Study 1b. Note: Error bars indicate 95% confidence intervals. See the online article for the color version of this figure.

provided a strong test of whether balance can reduce discrimination, as the expert consensus information was provided explicitly in a precise numerical form that inevitably led participants to be very responsive to it in both the balance and control conditions. It is also worth noting the effect size was larger in Study 1c (d = .59), which effectively forced participants to attend to the review excerpts by requesting a rating of each, compared with that of Studies 1a and 1b (d = .29 and .35), which did not. This observation suggests that, in more naturalistic settings in which the excerpts might tend to draw more attention than "background" weight-of-evidence statements (regarding the overall distribution of expert opinion), balance might have a stronger impact.

Study 2

Generalizability of the results from Study 1 was tested through a conceptual replication in a new domain. Participants were presented with a statement regarding an economic issue that had been evaluated by a panel of economics experts, each of whom had indicated whether or not they agreed with the statement. Participants saw the proportion of experts who agreed, disagreed, or were uncertain. Several measures of perceived consensus among the experts were taken. Participants were presented with one of four issue statements, two of which had high expert consensus (most of the experts agreed with the statement) and two of which had low consensus (experts were more evenly divided between agreeing and disagreeing with the statement). In contrast to Study 1, then, the target topic/issue was varied between rather than within subjects.

In addition to the new domain, Study 2 differed in some other important respects from Study 1. First, in contrast to the more continuous assessment of movie quality in Study 1, in Study 2 the issues were presented in a more dichotomous fashion in the form of statements that experts either agreed or disagreed with. Arguably, this dichotomous structure may map on more directly to the contexts (e.g., global warming, vaccines) in which concerns have been raised about false balance in media coverage, when experts with discrete opposing opinions about the truth value of a claim are pitted against one another. Second, in Study 2, participants' own opinions could potentially play a role in their judgments, as they could evaluate the economic issue for themselves, in contrast to Study 1 in which the movies were not identified so participants were unable to draw on their own knowledge or opinions. Third, possible order of presentation effects were examined in Study 2 by varying, in the balance condition, (a) whether the relative frequency table summarizing the distribution of expert opinion was presented before or after the balanced excerpts; and (b) whether the positive excerpt (i.e., from the expert who agreed with the statement) was presented before or after the negative excerpt (from the expert who disagreed).

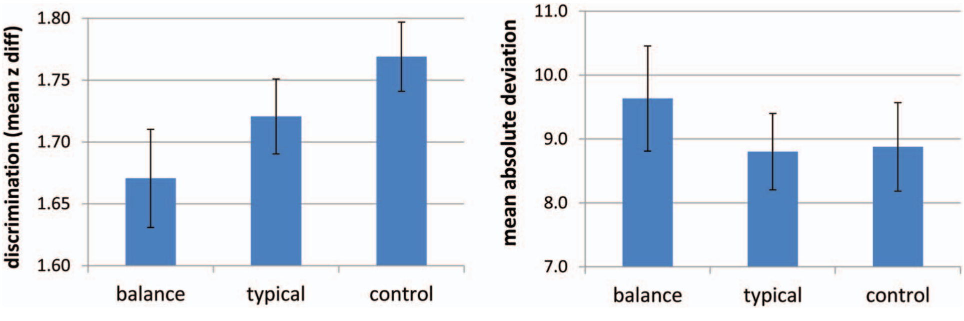


Figure 3. Mean discrimination (separation in standardized estimated ratings between good and mediocre movies) and absolute deviation (distance between estimated and actual web site movie ratings) by condition in Study 1c. Note: Error bars indicate 95% confidence intervals. See the online article for the color version of this figure.

Table 2
Mean Supplementary Ratings (Perceived Agreement Among Reviewers, Certainty With Which Good Movies Could be Distinguished From Mediocre Based on the Reviews, and Confidence That Reviews Could be Used to Determine Whether the Participant Personally Would Enjoy the Movie) by Condition in Study 1c

Measure	Condition	Mean	*SD*	95% CI
Agreement	Balance	2.53	1.06	[2.31, 2.75]
	Typical	4.48	.89	[4.29, 4.66]
	Control	3.63	.95	[3.44, 3.82]
Certainty	Balance	3.52	.96	[3.32, 3.72]
	Typical	3.94	1.04	[3.72, 4.16]
	Control	3.84	.96	[3.65, 4.04]
Personal confidence	Balance	3.60	1.40	[3.31, 3.90]
	Typical	3.98	1.34	[3.70, 4.26]
	Control	3.64	1.44	[3.34, 3.93]

Stimuli

The IGM Economic Experts Panel, administered by the University of Chicago Booth School of Business, is a group of expert economists with interests in public policy, drawn from elite U.S. research universities and selected for diversity in areas of expertise, geographical location, and political affiliation. Each month, the panel is presented with a statement on an economic issue, and they are asked to indicate if they agree with the statement, disagree with it, or are uncertain. (Stimuli in this study collapsed over indications of *strongly agree* vs. *agree* and likewise over *strongly disagree* and *disagree*. Experts may also indicate *no opinion* or may fail to respond to that month's issue; these cases were excluded in calculating the proportion of expert agreement.) In addition to indicating whether or not they agree, panel members have the option of providing a comment elaborating on or justifying their opinion. Excerpts from those comments were provided in the balance condition of the study.

Four issue statements (see Table 3) were selected for inclusion in the study, two of which (carbon tax and surge pricing) had high levels of agreement within the panel and two of which (robots and minimum wage) did not. The issues were selected as follows. Starting with the most recent statement evaluated by the panel (at the time the study materials were developed) and working backward, issues were sought that (a) were expected to be understandable by nonexpert participants; (b) had either high consensus levels (approximately 90%) or lower consensus levels (such that the percentage of the panel agreeing with the statement was just slightly higher than the percentages who disagreed or were uncertain); and (c) included at least one comment each from an expert who agreed with the statement and one who disagreed. For each of the four selected issue statements, the percentage of panel experts who agreed, disagreed, or were uncertain was calculated from the published results (http://www.igmchicago.org/igm-economic-experts-panel), excluding nonparticipating and abstaining experts on that issue.

Procedure

Participants ($N = 402$) were recruited from Amazon Mechanical Turk, subject to the condition that they be U.S. residents, and

Table 3
Economic Issues Presented in Study 2

Issue/statement	Agree	Uncertain	Disagree	Agree excerpt	Disagree excerpt
Carbon Tax: A tax on the carbon content of fuels would be a less expensive way to reduce carbon-dioxide emissions than would policies such as "corporate average fuel economy" requirements for automobiles.	93	5	2	This is as clear as economics gets; provides incentives to find minimally costly ways to reduce emissions.	This compares two ineffective approaches. The magnitude of this problem is so great that no sufficient carbon tax is feasible worldwide.
Surge Pricing: Allowing taxicabs to increase prices when demand is high—during peak hours or when the weather is bad—raises consumer welfare by increasing the supply of those services and allocating them to people who desire them the most.	86	10	4	The alternative is standing in the rain or waiting forever at rush hour, sometimes paying the premium is just much better.	Efficiency is NOT the same as welfare! This is probably a good policy, but some people will lose.
Robots: Information technology and automation are a central reason why median wages have been stagnant in the U.S. over the past decade, despite rising productivity.	40	35	24	Unskilled jobs have been lost which may well be a factor, although not the only one, behind stagnant median income and increasing inequality.	Rising health care costs may actually be more important for the median worker.
Minimum Wage: Raising the federal minimum wage would make it noticeably harder for low-skilled workers to find employment.	38	27	36	Unemployment among low-skilled workers is already high by historic standards, indicating that wages are already too high for market-clearing.	The empirical evidence now pretty decisively shows no employment effect, even a few years later.

Note. The agree, uncertain, and disagree columns list the percentage of expert panel members holding that position on the issue statement. The first two issues listed were considered to have high expert consensus and the last two, low consensus.

received 25 cents for their participation. The study was completed online.

After completing a consent form, participants were instructed:

> In this study, you will be asked to evaluate the extent to which a panel of economic experts agree with each other on an economic issue. The experts are mostly economics professors, drawn from the best universities in the U.S., who have agreed to participate in a monthly opinion poll. Specifically, the experts are provided with a statement regarding an economic issue, and asked to indicate whether they agree with the statement, disagree with the statement, or are uncertain. You will be told the percentage of experts who agreed with the statement, were uncertain, or disagreed with the statement.

Participants in the balance condition were further informed:

> Optionally, the experts can also provide a brief comment explaining their opinion. An example comment from one expert who agreed with the statement and one who disagreed with the statement will be shown.

Participants were randomly assigned to be presented with one of the four economic issue statements, in either the control or balance condition. Within the balance condition, furthermore, were two presentation order variables: The table summarizing the percentage of experts agreeing or disagreeing with the statement was presented either before or after the excerpts; and the excerpt from the expert who agreed with the statement was presented either before or after the excerpt from the expert who disagreed with the statement.

Participants were presented with the summary table and (in the balance condition) the excerpts from the two experts on a single screen along with all the dependent measures, which were administered in a fixed order. The table showed the percentage of experts who agreed, were uncertain, or who disagreed with the statement. In the balance condition, the two excerpts were labeled as "example comment from an expert who [dis]agreed with this statement" and were set next to head-and-shoulders silhouettes to emphasize they came from two individual experts. Appendix B provides an example screenshot showing how this information was presented to participants.

In the balance condition, an initial item asked, "Of the two example comments above, did one seem to reflect greater knowledge than the other of this economic issue?" and participants were asked to indicate whether it was the comment from the expert who agreed with the statement, the one from the expert who disagreed, or if both comments seemed equally knowledgeable. This item, results of which are not analyzed further, was included to ensure that participants in the balance condition read the two example comments.

The next three items were the key dependent variables designed to measure perceived agreement among the experts. The first of these (*agree*) asked "To what extent is there agreement among the panel of experts on this economic issue?" with responses given on a 7-point scale (with endpoints labeled *very little agreement* and *a lot of agreement*). The second (*2-experts*) was "Suppose we chose two experts at random from the panel whose opinions are shown in the table above. What is the probability that those two experts would share the same opinion on this issue?" with responses made on a slider that ran from 0% to 100%. The third (*20-experts*) was "Suppose 20 additional economic experts, with similar qualifications, were surveyed on this same issue. How many do you predict would agree with the statement above?" with responses made on a slider that ran from 0 to 20.

The next item (*policy*) was designed to gauge consequences of the perceived level of agreement among the experts on the issue in question. Participants were asked, "Does there seem to be enough agreement in expert opinion on this economic issue to use it as a basis for guiding government policy?" with responses made on a 5-point scale (labels: *definitely not*; *probably not*; *maybe*; *probably yes*; *definitely yes*).

A final item asked for the participant's own opinion on the issue statement, to which they could indicate (as did the experts) either agree, uncertain, or disagree (scored as 1, 2, 3, respectively). This measure was collected primarily for possible use as a covariate in analyses of the other dependent measures, though it could also be used to measure the impact of the expert panel information given that this measure was administered following its presentation.

Results

Of the 402 participants, nine had missing data on one or more of the key dependent variables, so their data were excluded from further analysis, leaving 393 participants in the final sample.

The three key dependent measures of perceived agreement among the experts, as expected, were positively correlated with one another: $r = .82$ between *2-experts* and *20-experts*, and $r = .72$ and .68 between *agree* and *2-experts* and *20-experts*, respectively. As an aggregate test of the primary hypothesis, these three measures were combined by first standardizing them and then taking the mean of the standard scores for each participant.

Recall that each participant was presented with one of four economic issue statements, two of which had high consensus among the experts and two of which had low consensus. A 2 (condition: control vs. balance) × 2 (issue consensus: high vs. low) factorial ANOVA was conducted with the combined agreement measure as the dependent variable. The interaction between these two factors was of primary interest, as it tests the hypothesis that the balance condition leads to poorer discrimination in ratings of perceived agreement between what are in fact high and low consensus issues. The condition by consensus interaction was statistically significant for the combined agreement measure, $F(1, 389) = 11.01$, $p = .001$, partial $\eta^2 = .028$. The interaction was also statistically significant for the *agree* ($p < .001$) and *2-experts* ($p = .033$) measures separately, but not for the *20-experts* measure ($p = .107$). Means on the combined agreement measures are displayed by condition in Figure 4. As in the previous studies, participants in the balance condition gave ratings that generally discriminated less well between the high and low consensus issues. In particular, the difference between the balance and control conditions is quite large for the high consensus issues but quite small for the low consensus issues. In other words, the primary effect of balance was to reduce perceived consensus among experts on the high consensus issues.

A corresponding ANOVA with *policy* as the dependent variable also revealed a significant condition by consensus interaction, $F(1, 389) = 7.67$, $p = .006$, partial $\eta^2 = .019$. As shown in Figure 5, ratings of the extent to which there was enough expert agreement to guide government policy discriminated less strongly the high-

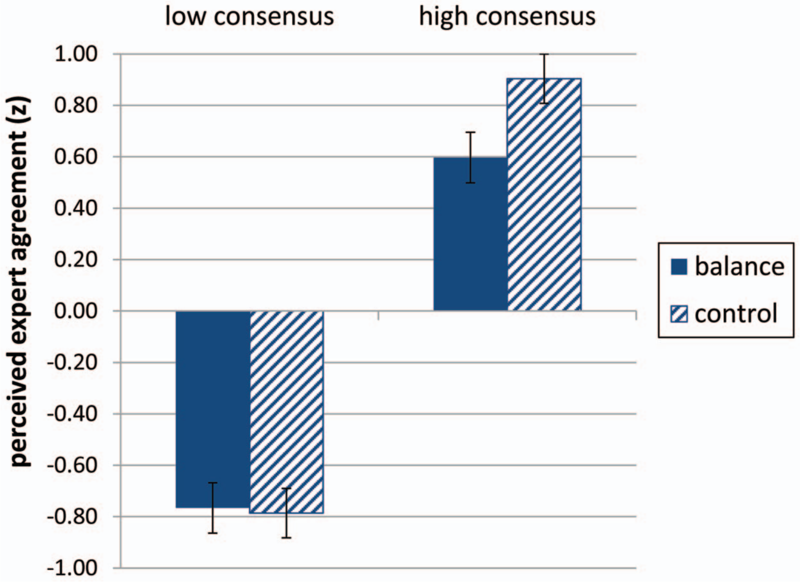


Figure 4. Perceived agreement (combined measure) among experts by level of expert consensus on the issue and condition (balance vs. control) in Study 2. Error bars indicate 95% confidence intervals. See the online article for the color version of this figure.

from the low-consensus issues in the balance condition than they did in the control condition.

Including the participant's own opinion on the issue as a covariate did not produce any qualitative differences to the ANOVA results reported above. There was a main effect of issue consensus on participants' own opinion ratings, as might be expected: Participants were more likely to disagree with the economic issue statements that had low consensus (*M* = 1.94, 95% CI [1.83, 2.05]) than with those that had high consensus on that statement (*M* = 1.68, 95% CI [1.58, 1.78]), *F*(1, 389) = 11.32, *p* = .001, partial η^2 = .028. This could be due to participants seeing the experts' opinions and adjusting their own opinions accordingly, or could simply reflect preexisting agreement between the participants and the experts on these issues.

The impact of the presentation order variables (table before vs. after comments; comment from expert who agreed vs. disagreed

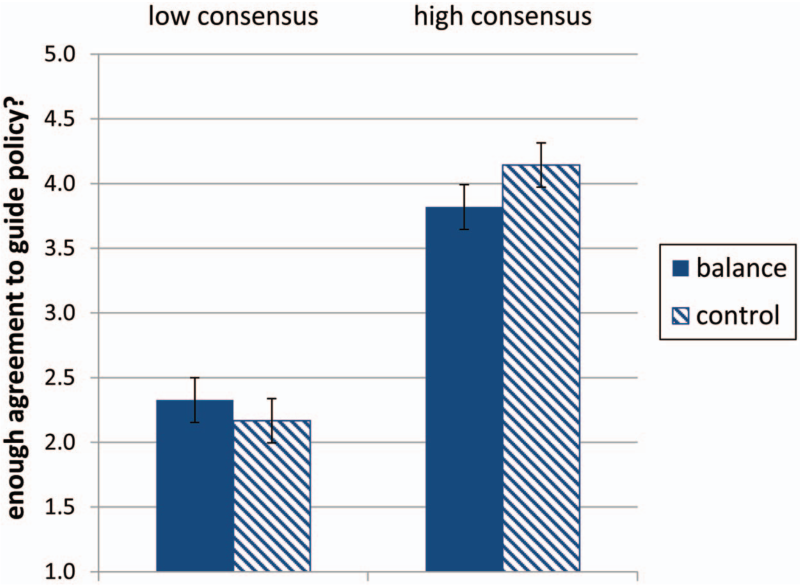


Figure 5. Mean endorsement that there is enough agreement among the experts to guide government policy, by level of expert consensus on the issue and condition (balance vs. control) in Study 2. Error bars indicate 95% confidence intervals. See the online article for the color version of this figure.

with the issue statement presented first) was also examined. In an ANOVA of the combined agreement measurement in the balance condition only (as the order variables did not apply to the control condition), neither variable had significant main effects nor did they interact with issue consensus. The same held in separate analyses of each of the constituents of the combined agreement measure (*agree, 2-experts, 20-experts*). The only analysis that showed any effect of the order variables was that of the *policy* measure, which revealed stronger endorsement of the claim that there was enough agreement among the experts to guide government policy when the first of the two example comments came from the expert who agreed with the statement ($M = 3.25$, 95% CI [3.08, 3.43]) rather than from the one who disagreed ($M = 2.87$, 95% CI [2.68, 3.06]), $F(1, 186) = 8.50$, $p = .004$, partial $\eta^2 = .044$. There was no significant interaction, however, between this order variable and issue consensus on the *policy* ratings.

General Discussion

The present studies showed that balanced comments from disagreeing experts systematically influence perceptions of the distribution of expert opinion. Presentation of balanced conflicting comments from an expert on either side of an issue had the effect of reducing the sharpness with which participants' ratings distinguished between issues with high and low expert consensus. This result held even though "weight of evidence" information was provided, summarizing the proportion of experts whose opinions fall on either side of the issue. The second study indicated that the effect of balanced comments was to reduce perceived agreement among experts specifically on high-consensus issues. This finding is consistent with the concerns of media critics that false balance (i.e., presenting both sides of an issue when in fact one side is overwhelmingly supported by the majority of experts) can distort public opinion by inflating perceptions of disagreement and uncertainty among experts. The first study showed, furthermore, that consistent presentation of balanced comments across multiple issues (in that study, movies) led to global perceptions of greater disagreement among experts and reduced certainty or confidence in the predictive validity of expert opinion in that domain.

The finding that balanced expert comments were found to exert a distorting influence on perceptions of the distribution of expert opinion in the present studies was particularly notable in that it held despite participants being provided with precise "weight of evidence" information regarding the proportion of experts on the panel with opinions on either side of the target issue. Previous research has established that weight-of-evidence information can attenuate the distorting influence of (falsely) balanced coverage of an issue (Clarke et al., 2015; Dixon et al., 2015; Kortenkamp & Basten, 2015). Some of this previous research also suggested that effects of false balance are not entirely remedied by provision of imprecise, qualitative weight-of-evidence information (e.g., that one of the experts "represented the minority viewpoint (one of the few who disagrees)"; Kortenkamp & Basten, 2015). In the present research, the question was whether false balance can still exert a distorting influence even in the presence of precise, numerical weight-of-evidence information, and evidence was found that it can.

Further research is needed to identify the underlying mechanisms associated with the main finding of the present studies, that balanced presentation of conflicting expert comments reduces perceived differences in the distribution of expert opinion between high and low consensus issues. Four possible mechanisms were described in the introduction. Of those, one, the dilution effect, appears a less likely candidate in light of the finding from the Study 1b that two typical—and hence arguably nondiagnostic—expert comments that did not conflict with one another did not reduce discrimination in the same manner as balanced conflicting comments.

The other three potential underlying mechanisms remain as viable explanations. Indeed, it seems likely that the observed effect of false balance on the perceived distribution of expert opinion is multiply determined. Still, it is worth considering how the hypothesized contribution of each possible mechanism might be tested in future research. According to the selection neglect account, comments from a dissenting minority expert are not sufficiently discounted for the fact very few experts share that opinion. False balance would have less of an effect, on this account, under conditions that promote greater correction or adjustment from the initial impression conveyed by comments from the expert minority, as could be tested for example by manipulating various factors previously shown to influence the amount of correction or adjustment from an initial anchor value (e.g., Epley & Gilovich, 2001; Gilbert, Pelham, & Krull, 1988). According to the partition dependence account, exposure to balanced comments from an expert on either side of the issue invokes a twofold partition that biases judgments toward a midpoint between the two positions. This account could be tested by presenting a number of expert comments in proportion to their relative prevalence in the expert population, which would be predicted to have a debiasing influence (e.g., Fox & Clemen, 2005; Ubel, Jepson, & Baron, 2001). According to the conflict sensitivity account, the experience of conflict produced by balanced presentation of two disagreeing expert comments contaminates perceptions of the level of uncertainty among the experts in general on the topic, reducing the extremity of associated perceptions of the distribution of expert opinion. This account could be tested by drawing attention to the actual source of conflict (i.e., that the journalist deliberately selected comments from experts who disagree with one another) in a manner similar to manipulations that have been shown to reduce misattribution in other settings (e.g., Schwarz & Clore, 1983).

It is also worth considering possible moderators of the false balance effect observed in the current studies. Perhaps most salient is the question of whether the results observed here would hold for other issue domains. This research was motivated by concerns raised by media critics over the possible distorting influence of false balance on public understanding of complex issues such as anthropogenic climate change, vaccine safety, or genetically modified organism (GMO) foods. As suggested by these examples, coverage of scientific issues seems particularly susceptible to charges of journalistic false balance. The present studies did not investigate the influence of false balance in scientific domains, but focused rather on cultural (Study 1, movies) and economic (Study 2) issues. These domains were chosen largely because they are quite distinct from one another, providing at least some evidence of generalizability, and because real expert panel data were available from both domains in the format required by the experimental design. While it seems reasonable to expect similar results in any complex domain where the input of experts is sought to aid public understanding, it remains a question for future research whether there are important differences in how false balance influences public perception of expert opinion in different domains. It is possible, for instance, that in more complex or unfamiliar domains people focus more on expert credentials and more readily discount comments from dissenting expert minorities in assessing the distribution of credible expert opinion. Alternatively, in such domains people may be even more sensitive to any appearance of conflict among experts and more ready to conclude on the basis of such apparent disagreement that expert consensus is lacking.

In the present studies, weight-of-evidence information was presented numerically in a summary table while balanced comments from individual experts were presented in the form of a written sentence of two. The magnitude of the false balance effect found in these studies likely depends on how this information was conveyed. Previous research (e.g., Borgida & Nisbett, 1977; De Wit, Das, & Vet, 2008) indicates, for example, that people's judgments are often more influenced by individual comments than by summary statistics. Presenting weight-of-evidence information in a more vivid, compelling format might have increased its impact and accordingly decreased the false balance effect of exposure to the conflicting expert comments. Indeed, Dixon, McKeever, Holton, Clarke, and Eosco (2015) found that presenting numerical weight-of-evidence information was more effective in offsetting the effect of false balance on personal beliefs about the vaccine-autism link when it was accompanied by a photograph of a scientist or group of scientists representing the majority view that vaccines do not cause autism. By contrast, presenting the expert comments in a more concrete or vivid format, for example by showing a video clip of two experts debating one another on the issue, might have increased the distorting influence of false balance.

Finally, there is the question of how false balance might affect perceptions of individuals who already hold strong opinions on the target issue. The present studies largely avoided this question, either by stripping away identifying information in such a way that participants were forced to rely exclusively on expert opinion (Study 1, with unidentified movies) or by focusing on complex issues that participants were unlikely to feel strongly about and lacked expertise to assess based on their own knowledge (Study 2, with economic issues). In the case of more politically charged issues, such as climate change, people who lack expertise nevertheless have strong preexisting opinions. The question of how such individuals are potentially influenced by false balance is another important topic for future research. Strong preexisting opinions may make it easier to dismiss the comments from an expert with whom one disagrees, particularly if that expert is clearly part of a small dissenting minority. In an age of "echo chambers" in which people can readily select their news sources to reinforce their existing views, journalistic norms of balance may be more important than ever. Psychological research can contribute to our knowledge of when journalistic balance aids, and when it distorts, public understanding of what experts have to tell us.

References

Anderegg, W. R., Prall, J. W., Harold, J., & Schneider, S. H. (2010). Expert credibility in climate change. *Proceedings of the National Academy of Sciences of the United States of America, 107,* 12107–12109. http://dx.doi.org/10.1073/pnas.1003187107

BBC Trust. (2011, July). BBC Trust review of impartiality and accuracy of the BBC's coverage of science. Retrieved from http://downloads.bbc.co.uk/bbctrust/assets/files/pdf/our_work/science_impartiality/science_impartiality.pdf

BBC Trust. (2014, July). *Trust conclusions on the executive report on science impartiality review actions*. Retrieved from http://downloads.bbc.co.uk/bbctrust/assets/files/pdf/our_work/science_impartiality/trust_conclusions.pdf

Bennett, W. L. (1996). An introduction to journalism norms and representations in politics. *Political Communication, 13,* 373–384. http://dx.doi.org/10.1080/10584609.1996.9963126

Borgida, E., & Nisbett, R. E. (1977). The differential impact of abstract vs. concrete information on decisions. *Journal of Applied Social Psychology, 7,* 258–271. http://dx.doi.org/10.1111/j.1559-1816.1977.tb00750.x

Boykoff, M. T. (2013). Public enemy no. 1? Understanding media representations of outlier views on climate change. *American Behavioral Scientist, 57,* 796–817. http://dx.doi.org/10.1177/0002764213476846

Boykoff, M., & Boykoff, J. (2004). Balance as bias: Global warming and the U.S. prestige press. *Global Environmental Change, 14,* 125–136. http://dx.doi.org/10.1016/j.gloenvcha.2003.10.001

Brenner, L. A., Koehler, D. J., & Tversky, A. (1996). On the evaluation of one-sided evidence. *Journal of Behavioral Decision Making, 9,* 59–70. http://dx.doi.org/10.1002/(SICI)1099-0771(199603)9:1<59::AID-BDM216>3.0.CO;2-V

Clarke, C. E., Dixon, G. N., Holton, A., & McKeever, B. W. (2015). Including "evidentiary balance" in news media coverage of vaccine risk. *Health Communication, 30,* 461–472. http://dx.doi.org/10.1080/10410236.2013.867006

Cook, J., Nuccitelli, D., Green, S. A., Richardson, M., Winkler, B., Painting, R., . . . Skuce, A. (2013). Quantifying the consensus on anthropogenic global warming in the scientific literature. *Environmental Research Letters, 8,* 024024. http://dx.doi.org/10.1088/1748-9326/8/2/024024

Corbett, J. B., & Durfee, J. L. (2004). Testing public (un)certainty of science: Media representations of global warming. *Science Communication, 26,* 129–151. http://dx.doi.org/10.1177/1075547004270234

Dearing, J. W. (1995). Newspaper coverage of maverick science: Creating controversy through balancing. *Public Understanding of Science, 4,* 341–361. http://dx.doi.org/10.1088/0963-6625/4/4/002

De Neys, W., Cromheeke, S., & Osman, M. (2011). Biased but in doubt: Conflict and decision confidence. *PLoS ONE, 6,* e15954. http://dx.doi.org/10.1371/journal.pone.0015954

De Wit, J. B., Das, E., & Vet, R. (2008). What works best: Objective statistics or a personal testimonial? An assessment of the persuasive effects of different types of message evidence on risk perception. *Health Psychology, 27,* 110–115. http://dx.doi.org/10.1037/0278-6133.27.1.110

Dixon, G., & Clarke, C. (2013). Heightening uncertainty around certain science: Media coverage, false balance, and the autism–vaccine controversy. *Science Communication, 35,* 358–382. http://dx.doi.org/10.1177/1075547012458290

Dixon, G. N., McKeever, B. W., Holton, A. E., Clarke, C., & Eosco, G. (2015). The power of a picture: Overcoming scientific misinformation by communicating weight-of-evidence information with visual exemplars. *Journal of Communication, 65,* 639–659. http://dx.doi.org/10.1111/jcom.12159

Doran, P. T., & Zimmerman, M. K. (2009). Examining the scientific consensus on climate change. *Eos, Transactions, American Geophysical Union, 90,* 22–23. http://dx.doi.org/10.1029/2009EO030002

Dunwoody, S. (2005). Weight-of-evidence reporting: What is it? Why use it? Nieman Reports. Retrieved from http://niemanreports.org/articles/weight-of-evidence-reporting-what-is-it-why-use-it/

Entman, R. W. (1989). *Democracy without citizens: Media and the decay of American politics*. New York, NY: Oxford University Press.

Epley, N., & Gilovich, T. (2001). Putting adjustment back in the anchoring and adjustment heuristic: Differential processing of self-generated and experimenter-provided anchors. *Psychological Science, 12,* 391–396. http://dx.doi.org/10.1111/1467-9280.00372

Fox, C. R., & Clemen, R. T. (2005). Subjective probability assessment in decision analysis: Partition dependence and bias toward the ignorance prior. *Management Science, 51,* 1417–1432. http://dx.doi.org/10.1287/mnsc.1050.0409

Fox, C. R., & Rottenstreich, Y. (2003). Partition priming in judgment under uncertainty. *Psychological Science, 14,* 195–200. http://dx.doi.org/10.1111/1467-9280.02431

Gallup. (2014). *One in four in U.S. are solidly skeptical of global warming*. Retrieved from http://www.gallup.com/poll/168620/one-four-solidly-skeptical-global-warming.aspx

Gilbert, D. T., Pelham, B. W., & Krull, D. S. (1988). On cognitive busyness: When person perceivers meet persons perceived. *Journal of Personality and Social Psychology, 54,* 733–740. http://dx.doi.org/10.1037/0022-3514.54.5.733

Greenberg, M., Robbins, D., & Theel, S. (2013). Media sowed doubt in coverage of UN climate report. *Media Matters for America*. Retrieved from http://mediamatters.org/research/2013/10/10/study-media-sowed-doubt-in-coverage-of-un-clima/196387

Griffin, D., & Tversky, A. (1992). The weighing of evidence and the determinants of confidence. *Cognitive Psychology, 24,* 411–435. http://dx.doi.org/10.1016/0010-0285(92)90013-R

Hamill, R., Wilson, T. D., & Nisbett, R. E. (1980). Insensitivity to sample bias: Generalizing from atypical cases. *Journal of Personality and Social Psychology, 39,* 578–589. http://dx.doi.org/10.1037/0022-3514.39.4.578

Jensen, J. D., & Hurley, R. J. (2012). Conflicting stories about public scientific controversies: Effects of news convergence and divergence on scientists' credibility. *Public Understanding of Science, 21,* 689–704. http://dx.doi.org/10.1177/0963662510387759

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review, 80,* 237–251. http://dx.doi.org/10.1037/h0034747

Koehler, J. J., & Mercer, M. (2009). Selection neglect in mutual fund advertisements. *Management Science, 55,* 1107–1121. http://dx.doi.org/10.1287/mnsc.1090.1013

Kortenkamp, K. V., & Basten, B. (2015). Environmental science in the media effects of opposing viewpoints on risk and uncertainty perceptions. *Science Communication, 37,* 287–313. http://dx.doi.org/10.1177/1075547015574016

Leiserowitz, A., Feinberg, G., Rosenthal, S., Maibach, E., & Roser-Renouf, C. (2014). *Climate change in the American mind, April 2014. Yale project on climate change communication*. Retrieved from http://environment.yale.edu/climate-communication/files/Climate-Change-American-Mind-April-2014.pdf

Nisbett, R. E., Zukier, H., & Lemley, R. E. (1981). The dilution effect: Nondiagnostic information weakens the implications of diagnostic information. *Cognitive Psychology, 13,* 248–277. http://dx.doi.org/10.1016/0010-0285(81)90010-4

National Public Radio. (2012, May). *NPR ethics handbook* Retrieved from http://ethics.npr.org/wp-content/uploads/2012/05/NPR-Ethics-Handbook-5.2.2012-Final-Edition.pdf

Schwarz, N., & Clore, G. L. (1983). Mood, misattribution, and judgments of well-being: Informative and directive functions of affective states. *Journal of Personality and Social Psychology, 45,* 513–523. http://dx.doi.org/10.1037/0022-3514.45.3.513

Stocking, S. H. (1999). How journalists deal with scientific uncertainty. In S. M. Friedman, S. Dunwoody, & C. L. Rogers (Eds.), *Communicating uncertainty: Media coverage of new and controversial science* (pp. 23–41). Mahwah, NJ: Erlbaum.

Troutman, C. M., & Shanteau, J. (1977). Inferences based on nondiagnostic information. *Organizational Behavior and Human Performance, 19,* 43–55. http://dx.doi.org/10.1016/0030-5073(77)90053-8

Ubel, P. A., Jepson, C., & Baron, J. (2001). The inclusion of patient testimonials in decision aids: Effects on treatment choices. *Medical Decision Making, 21,* 60–68.

Weber, E. U., & Stern, P. C. (2011). Public understanding of climate change in the United States. *American Psychologist, 66,* 315–328. http://dx.doi.org/10.1037/a0023253

Wilson, K. M. (1995). Mass media as sources of global warming knowledge. *Mass Communications Review, 22,* 75–89.

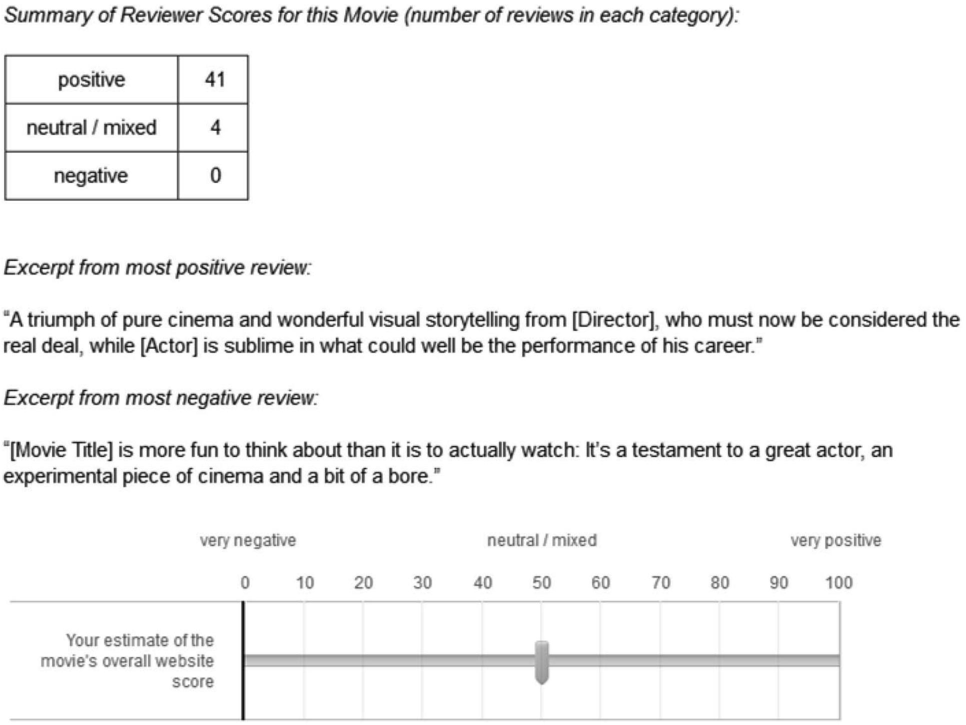
Wilson, K. M. (2000). Drought, debate, and uncertainty: Measuring reporters' knowledge and ignorance about climate change. *Public Understanding of Science, 9,* 1–13. http://dx.doi.org/10.1088/0963-6625/9/1/301

(*Appendix follows*)

Appendix A

Example Screenshot from Balance Condition of Study 1a

Positive and negative review excerpts were not presented in the control condition, which otherwise followed the same format.



Summary of Reviewer Scores for this Movie (number of reviews in each category):

positive	41
neutral / mixed	4
negative	0

Excerpt from most positive review:

"A triumph of pure cinema and wonderful visual storytelling from [Director], who must now be considered the real deal, while [Actor] is sublime in what could well be the performance of his career."

Excerpt from most negative review:

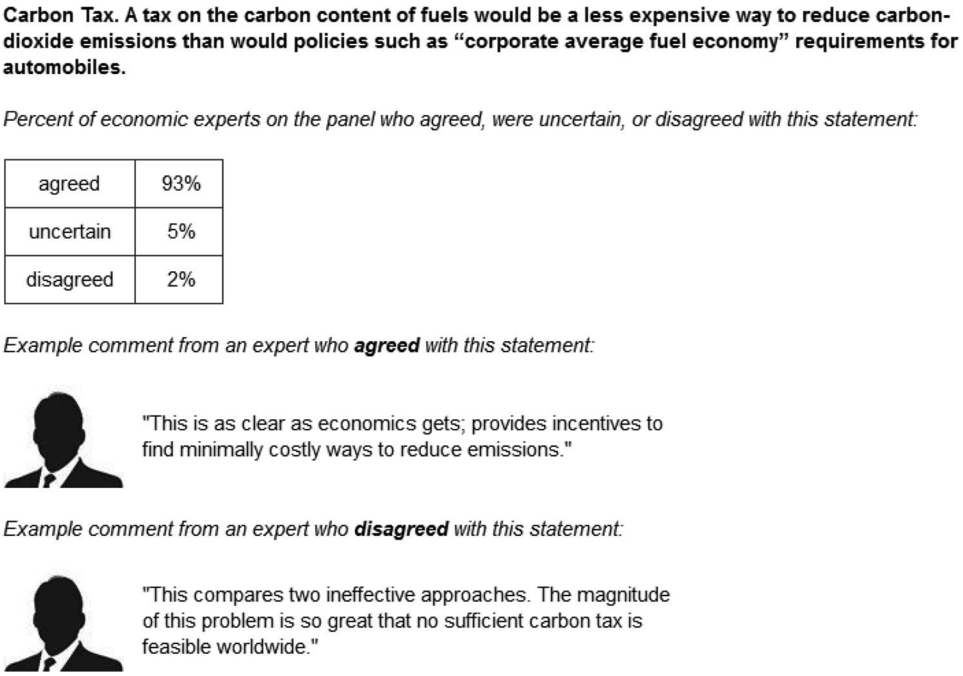
"[Movie Title] is more fun to think about than it is to actually watch: It's a testament to a great actor, an experimental piece of cinema and a bit of a bore."

(*Appendices continue*)

Appendix B

Example Screenshot from Balance Condition of Study 2

Example comments from experts were not presented in the control condition, which otherwise followed the same format. Order of the two comments, as well as whether those comments appeared before or after the summary table, was counterbalanced across participants.



Carbon Tax. A tax on the carbon content of fuels would be a less expensive way to reduce carbon-dioxide emissions than would policies such as "corporate average fuel economy" requirements for automobiles.

Percent of economic experts on the panel who agreed, were uncertain, or disagreed with this statement:

agreed	93%
uncertain	5%
disagreed	2%

Example comment from an expert who ***agreed*** *with this statement:*

"This is as clear as economics gets; provides incentives to find minimally costly ways to reduce emissions."

Example comment from an expert who ***disagreed*** *with this statement:*

"This compares two ineffective approaches. The magnitude of this problem is so great that no sufficient carbon tax is feasible worldwide."

Received June 11, 2015
Revision received November 12, 2015
Accepted November 13, 2015 ■