American Psychological Association

Journal of Experimental Psychology: Applied

© 2020 American Psychological Association
ISSN: 1076-898X

http://dx.doi.org/10.1037/xap0000279

How Heuristic Credibility Cues Affect Credibility Judgments and Decisions

AQ: 1

AQ: au
AQ: 2

Leo Gugerty and Drew M. Link
Clemson University

We investigated how heuristic credibility cues affected credibility judgments and decisions. Participants saw advice in comments in a simulated online health forum. Each comment was accompanied by credibility cues, including author expertise and peer reputation ratings (by forum members) of comments and authors. In Experiment 1, participants' credibility judgments of comments and authors increased with expertise and increased with the number of reputation ratings for supportive ratings and decreased with number of ratings for disconfirmatory ratings. Also, results suggested that the diagnosticity (informativeness) of credibility cues influenced credibility judgments. Using the same credibility cues and task context, Experiment 2 found that when high-utility choices had low credibility, participants often chose alternatives with lower utility but higher credibility. They did this more often when less utility had to be sacrificed and when more credibility was gained. The influence of credibility and utility information on participants' choices was mediated by their explicit credibility judgments. These findings supported the predictions of a Bayesian belief-updating model and an elaboration of Prospect Theory (Budescu, Kuhn, Kramer, & Johnson, 2002). This research provides novel insights into how cues including valence and relevance influence credibility judgments and how utility and credibility trade off during decision making.

Public Significance Statement

People often need to judge the credibility of information (e.g., news, advice) that is outside their expertise. Two studies showed that people effectively used rules of thumb like "credibility increases with the amount of corroborating information" when judging the credibility of advice on an online health forum and when making decisions based on low-credibility advice. However, study participants may have overweighted advice from forum members who lacked health expertise.

Keywords: ambiguity, Bayesian models, credibility, reputation

During argumentation and decision making, people often consider factual information relevant to claims or outcomes of choices. Most theories assume that the strength of arguments influences people's beliefs about the world (Hahn & Oaksford, 2007) and the utility of outcomes influences their decisions (Speekenbrink & Shanks, 2013). However, strong arguments and useful outcomes are worth little if their factual basis is suspect. Therefore, good argumentation and decision making requires that people also consider the credibility of evidence and of its source before using it. In this project, we conducted two experiments to investigate how people use credibility information when evaluating whether to believe a factual claim (or its source) and when making decisions. We used mathematical models of argumentation
AQ: 3 or belief updating (Hahn & Oaksford, 2007) and decision making (Budescu et al., 2002) to define credibility-related constructs and to guide our predictions.

Research Questions

Perceived Credibility During Argumentation (Belief Updating)

An early and continuing body of credibility research began with persuasive argumentation and belief updating (e.g., Hovland, Janis, & Kelley, 1953; Petty & Cacioppo, 1986) and later investigated credibility in online settings (e.g., Hilligoss & Rieh, 2008). One focus of this research has addressed how people's judgments of the credibility of factual claims are influenced by external cues such as the amount of evidence or cues to source credibility (e.g., expertise or reputation). In addition, researchers have developed models of argumentation and belief updating and empirically evaluated these models in the context of everyday (Hahn & Oaksford, 2007), legal (Lagnado, Fenton, & Neil, 2012) and scientific (Corner & Hahn, 2009) argumentation.

This research framed our first research question: How do credibility cues such as source expertise and reputation as well as the

Leo Gugerty and Drew M. Link, Department of Psychology, Clem-
AQ: 23 son University.

Correspondence concerning this article should be addressed to Leo Gugerty, Department of Psychology, Clemson University, 418 Brackett Hall, Clemson, SC 29634. E-mail: gugerty@clemson.edu

amount and valence of evidence influence peoples' subjective judgments (perceptions) of the credibility of a factual claim? (Valence refers to whether evidence is supportive or disconfirmatory of a claim.) We used a Bayesian model of belief updating, AQ: 4 which was derived from Bayes's theorem by Griffin and Tversky (1992), to predict how these and other cues interact in influencing the perceived credibility of a claim. In this model, people weight the evidence relevant to a claim based on how confident versus uncertain they feel about the evidence, for example, based on how credible they rate an information source. Griffin and Tversky and others (e.g., Keynes, 1921) call this weighting factor the weight of evidence. We empirically evaluated these predictions in Experiment 1. To our knowledge, relatively little research has investigated how peoples' credibility judgments are influenced by valence or by the interaction of valence and amount of evidence.

Credibility During Decision Making

A key reason for assessing the credibility of beliefs—for example, beliefs about the outcomes of choices—is to guide decision making and action (Crisci & Kassanove, 1973; Petty & Cacioppo, 1986; Yaniv & Kleinberger, 2000). Ellsberg (1961) pointed out that prior models of human decision making mistakenly assumed that the informational inputs to decisions (outcomes and their probabilities) are maximally credible. Since then, researchers have investigated how decision makers handle decision information that is uncertain. Researchers in the fields of argumentation and decision making tend to use different terms to describe uncertain information—low credibility in the former and ambiguity in the latter. However, just as belief-updating models assume that people weight evidence based on its uncertainty, decision making models assume that people weight the inputs to a decision based on its uncertainty (e.g., Budescu et al., 2002). Decision-making researchers have investigated a variety of cues to ambiguity, including amount of information and source reputation (Einhorn & Hogarth, 1985) and imprecision (e.g., Budescu et al., 2002). For example, consider choosing between one medical treatment with a precise success rate (20%) and another with an imprecise success rate (18–28%) that has higher expected utility. In decisions like this where credibility and utility conflict, people often sacrifice or trade off some utility to avoid choosing the ambiguous outcome (Budescu et al., 2002). Our second research question focused on these kinds of choices: To what extent will people sacrifice utility to choose a more credible alternative that gives them more confidence about the outcomes? In Experiment 2, participants made multiattribute decisions where credibility and utility information conflicted, like the one above. Budescu et al.'s (2002) model of decision making guided our predictions.

To our knowledge, little empirical research has investigated how credibility or ambiguity may influence both argumentation and decision making. Later, we consider the similarities and differences between the constructs of credibility and ambiguity and discuss how the credibility cues considered in these two frameworks may fit within a single model of belief updating.

Does Perceived Credibility Mediate Choice?

If external credibility cues influence people's perceptions of the credibility of claims (as in our first research question) and guide decisions (as in our second question), then perceived credibility may mediate the effects of these cues on decisions. This was our third research question. For each decision in Experiment 2, participants rated the perceived credibility of the outcomes for the choice they made, using the same credibility measures as in Experiment 1. A mediation analysis compared the direct effects of credibility cues on participants' choices to their indirect effects, that is, as mediated by their credibility judgments.

In the next two sections on traditional and Bayesian views of credibility, we summarize researchers' views of credibility and define the constructs we used to measure participants credibility judgments in both experiments. In the last section of the introduction, we review empirical research on credibility.

Traditional Views of Credibility

Credibility as Belief in a Claim

In the persuasion and Internet-credibility research, researchers have applied the construct of credibility to both information and its source. Credibility is seen as a subjective psychological judgment and is often defined in terms of *believability*—that is, credibility judgments are thought to reflect the degree to which someone believes in or agrees with the claim in a message (Hilligoss & Rieh, 2008; Hovland et al., 1953; Kelman & Hovland, 1953; Petty AQ: 5 & Cacioppo, 1986). If people strongly believe a claim to be true, they may intend to act on it. Thus, researchers have also measured the credibility of a claim using *intention to act* (Chaiken & Ma- AQ: 6 heswaren, 1994; Flanagin & Metzger, 2013).

Credibility of Evidence

The term credibility has another sense. When researchers use terms like source credibility or amount of information, they seem to be referring to attributes of the *evidence* that directly influence people's confidence in the evidence and only indirectly influence degree of belief in a *claim*. Two key attributes that are thought to influence confidence in evidence are the trustworthiness and accuracy of the evidence source. *Trustworthiness* refers to veracity and objectivity (i.e., lack of bias), whereas *accuracy* refers to predictive validity. Schum (1989) has discussed how accuracy, veracity, and objectivity are critical to judging source credibility. O'Keefe (1990) suggested that the expertise of a source (e.g., as cued by credentials) influences judgments of source credibility.

Thus, credibility refers both to how strongly people believe a claim (sense 1) and how much confidence they have in some evidence (sense 2). We measured participant's judgments of credibility, in both of its senses, using survey questions regarding the trustworthiness, accuracy and believability of messages and their sources, as well as intentions to act on the message. As the discussion above shows, these questions assess some of the key characteristics that researchers have used to describe the construct of credibility. Below, we define these four aspects of credibility more explicitly using Bayesian theory.

The dual-process models of Chaiken (1980) and Petty and Cacioppo (1986) have remained influential in understanding how external credibility cues affect people's beliefs. In these models, when people have the time, motivation, *and* domain expertise to make reflective, systematic judgments about the credibility of a message, they focus more on *semantic* cues in the message content. However, if people lack *any* of these three things, they tend to make fast, low-effort credibility judgments based on *heuristic* cues external to the message content. Researchers have identified a number of heuristic cues to credibility, including credentials, reputation, endorsements, imprecision, and amount of corroborating information (Budescu et al., 2002; Chaiken, 1980; Metzger, Flanagin, & Medders, 2010). Our first experiment focused on how certain heuristic cues affected credibility judgments.

A Bayesian View of Credibility

In the following, we frame some of the credibility-related constructs that were discussed above in terms of a Bayesian model of belief updating, which, in our view, enables a clearer, more precise understanding of these constructs and how they are interrelated. This model is based on the Bayesian belief-updating model presented in Hahn and Oaksford (2007) and Corner and Hahn (2009). AQ: 8 Following Anderson's (1990, 1991) idea that psychological models can address three levels of explanation—rational, process, and physiological—Hahn and colleagues describe their model as a rational (also called normative) model. According to Anderson, rational models describe the goals and outputs of a cognitive function, the environmental constraints on the function, and the behavioral model that optimizes (in an evolutionary sense) computing the output that meets the goal. In addition to Hahn and colleagues, many other researchers (Griffiths & Tenenbaum, 2009; Lu, Yuille, Liljeholm, Cheng & Holyoak, 2008; Meder & Mayrhofer, 2017; Oaksford & Chater, 2003) have developed rational, psychological models of argumentation or reasoning based on Bayesian belief updating and tested these models against human behavior.

Credibility as Bayesian Degree of Belief in a Claim

According to Hahn and Oaksford (2007), the goal of the cognitive function of argumentation is to persuade yourself or others whether or not to believe claims about the world. To achieve this goal, people need to compute the degree to which they believe particular claims to be true. In Bayesian terms, the latter construct is called personal degree of belief, personal probability (Edwards, Lindman, & Savage, 1963) or credibility (Kruschke & Vanpaemel, 2015) and is expressed as a probability ranging between two endpoints that denote certainty, 0 (false) and 1 (true). We define the credibility of a claim or hypothesis using the Bayesian construct of degrees of belief.

Prior to obtaining any evidence regarding a hypothesis (H), the Bayesian assumption is that someone's degree of belief—or personal probability—is maximally uncertain, that is, $P(H) = P(\neg H) = .5$. The term *personal* means that degrees of belief may differ across individuals. However, Bayes's theorem describes a normative procedure by which individuals can update their prior degree of belief based on an evidence set (E) that may contain AQ: 9 multiple pieces of evidence,

$$\frac{P(H|E)}{P(\neg H|E)} = \frac{P(E|H)}{P(E|\neg H)}\ \frac{P(H)}{P(\neg H)}. \qquad (1)$$

Thus, the posterior odds that a hypothesis is true versus false given some evidence depends on the relative likelihood of observing the evidence given that the hypothesis is true versus false (the likelihood ratio) and the prior odds that the hypothesis is true versus false. The posterior degree of belief, $P(H|E)$, is easily calculated from the posterior odds. In Bayesian updating, the posterior probability is calculated incrementally. For each new piece of evidence, the posterior degree of belief is updated using Equation 1. Then the posterior degree of belief becomes the prior for updating based on the next piece of evidence. Individuals who have different prior beliefs but update their beliefs in a Bayesian fashion using the same evidence set should arrive at similar posterior beliefs if enough evidence is available. The Bayesian construct of personal degree of belief in a claim corresponds closely to the traditional view of credibility as *believability* (e.g., Hovland et al., 1953), which was one of our measures of perceived credibility.

In line with the idea that people often judge credibility to guide action, Bayesians define the construct of personal degrees of belief in terms of how it affects decisions to act (Edwards et al., 1963). AQ: 10 For example, someone's personal degree of belief that a coin will come up heads is represented by a probability of .5 if the person is indifferent to making a high-stakes bet on heads or tails. Thus, measuring perceived credibility in terms of *intention to act*—another of our credibility measures—also fits within the Bayesian framework.

The *overall* likelihood ratio in Bayes's theorem (Equation 1)—where overall means based on *all* pieces of evidence in a set—requires some unpacking. Theorists from Keynes (1921) to Einhorn and Hogarth (1985) to Griffin and Tversky (1992) to Massey and Wu (2005) to Lagnado et al. (2012) have pointed out that the overall likelihood ratio, and therefore posterior degree of belief, depends on both the *strength of evidence* and the *weight of evidence*. Griffin and Tversky (1992) showed how the overall likelihood ratio can be decomposed into the strength, valence and weight of evidence. Strength of evidence specifies the magnitude of the change in degree of belief based on some evidence, whereas valence refers to the direction of belief change. Griffin & Tversky describe strength in terms of the "extremeness" of the evidence and see it as analogous to effect size.

Credibility or Weight of Evidence

Weight of evidence refers to factors (e.g., the amount or informativeness of evidence) that moderate or weight the effect of strength of evidence in changing degree of belief. Griffin and Tversky describe the weight of evidence in terms of predictive validity and see it as analogous to the statistical concept of precision (e.g., the confidence interval around an effect size). Thus, weight of evidence refers to cues that influence how confident versus uncertain reasoners feel about the evidence—the second sense of credibility discussed above. Cues that influence weight of evidence include the precision, relevance, or amount of evidence and the reputation that the information source has for making accurate and unbiased claims. Two of our questions measuring perceived credibility asked participants to judge the accuracy and trustworthiness (i.e., lack of bias) of the comment.

For example, the results from a single poll could strongly support the hypothesis that candidate X will win a two-person election (65% prefer X) or strongly disconfirm it (35% prefer X). However, this strong evidence—of either supportive or disconfirmatory valence—should not change belief in the election outcome much if the poll's margin of error is large (imprecision), the poll is old (low relevance), or the pollster has a poor reputation owing to an inconsistent track record (low accuracy), bias (low objectivity or veracity), or lack of training (low expertise). Going beyond one piece of evidence, the number of polls (amount of evidence) also influences the weight of evidence. In general, the overall likelihood ratio (in Equation 1) is only high if strength of evidence and *all* the factors comprising weight of evidence are high. It is low if *any* of these things is low. Note that these cues to the weight of evidence include cues emphasized by researchers investigating credibility in the context of belief updating (i.e., source credibility and amount of information) and decision making (i.e., precision).

Griffin and Tversky's Model

Griffin and Tversky (1992) showed that the overall likelihood ratio can be decomposed into four factors—the strength, valence, amount, and diagnosticity of evidence. Their version of Bayes's theorem is a special case that makes assumptions including: there are only two mutually exclusive hypotheses; the evidence takes only two values; and the prior odds ratio is 1. We present their model using an example where the hypothesis (*H*) is that a mineral supplement has a side effect of headache. The evidence set (*E*) is a description in a reputable scientific journal or a little-known website of a set of 10 or 80 scientific studies. Each study has one of two outcomes: *yes*, headache is a side effect or *no*, it is not. For each set of studies, N_{yes} conclude *yes* and N_{no} conclude *no*. For this example, Griffin and Tversky's equation is,

$$\frac{P(H|E)}{P(\neg H|E)} = \left(\frac{P(yes|H)}{P(yes|\neg H)}\right)^{\left(\frac{N_{yes}}{N} - \frac{N_{no}}{N}\right)N}, \qquad (2)$$

where $N = N_{yes} + N_{no}$. Appendix A describes how Equation 2 is derived from Equation 1. The *strength* of evidence is the difference in the proportion of studies concluding *yes* versus *no*. The evidence supports or disconfirms *H* depending on whether the strength of evidence is positive or negative, respectively. Thus, the sign of the strength of evidence represents the *valence* of the evidence. The total number of studies, *N*, represents the *amount* of evidence.

The base of the exponent is the likelihood ratio of a *single piece* of evidence. This likelihood ratio is a commonly used metric for quantifying the construct of *diagnosticity* (Tversky, 1977), which is also called informativeness (Hahn & Oaksford, 2007). Diagnosticity is the capacity for a piece of evidence to change degree of belief in a claim. The pieces of evidence in a set may have different diagnosticities. However, Equation 2 assumes that each piece of evidence has the same diagnosticity. Griffin and Tversky (1992) and Massey and Wu (2005) consider diagnosticity to be part of the weight of evidence. Corner and Hahn (2009) suggest that one factor influencing diagnosticity is source credibility, which they operationalize by a source's reputation for accuracy. In the current example, the scientific journal is assumed to exhibit higher accuracy, and therefore diagnosticity, than the website in interpreting the results of each study and discriminating between whether the mineral caused headache or not.

In addition to reputation for accuracy, expertise is also considered to be a cue to source credibility, and therefore, diagnosticity (Einhorn & Hogarth, 1985). In scientific meta-analysis, precision is used to represent diagnosticity; so that overall effect size is calculated by weighting the effect size (strength) of each study by its precision (Sutton & Abrams, 2001). From now on, we distinguish the two likelihood ratios discussed in this section by using LR_{ALL} (Equation 1) to refer to the likelihood ratio for *all* the evidence in a set and LR_{ONE} (Equation 2) to refer to the likelihood ratio for *one* piece of evidence. We only use the term diagnosticity for LR_{ONE}.

In summary, Griffin and Tversky's model (Equation 2) shows precisely how to integrate information about the strength, valence, diagnosticity, and amount of evidence to compute posterior degree of belief in a claim (with diagnosticity and amount of evidence comprising weight of evidence). Following Hahn and Oaksford (2007) and Anderson (1991), we interpret Equations 1 and 2 as specifying a psychological model—at the rational level of explanation—of how to compute posterior belief in a claim after examining evidence.

Model predictions. Figure 1 (calculated from Equation 2) shows a number of important relationships between evidence and degree of belief that are inherent in Bayesian belief updating. First, people tend to assume moderate prior belief in a hypothesis, .5, when evidence is lacking. Second, degree of belief approaches its maximum, 1.0, with large amounts of supportive evidence and its minimum, 0 (disbelief), with large amounts of disconfirmatory evidence. If a set of evidence consists of *all* supportive or *all* disconfirmatory evidence (as shown in Figure 1), degree of belief changes *monotonically* from moderate levels toward strong belief or disbelief as amount of information increases.

Third, the rate of change of belief with increasing evidence is proportional to the diagnosticity of the evidence. Minimum diagnosticity (and minimum change in belief) corresponds to the flat, dotted line, where the likelihood ratio for each piece of evidence (LR_{ONE}) is 1. For 100% supportive evidence, diagnosticity—the rate of change in belief with increasing evidence—increases as LR_{ONE} goes from 1 to 1.5 (dashed line) to 3.0 (solid line), that is, a positive correlation. For 100% disconfirmatory evidence, diagnosticity increases as LR_{ONE} goes from 1 to 0.67 (dashed line) to 0.33 (solid line), that is, a negative correlation. Thus, the model predicts an interaction between diagnosticity, valence and amount of information.

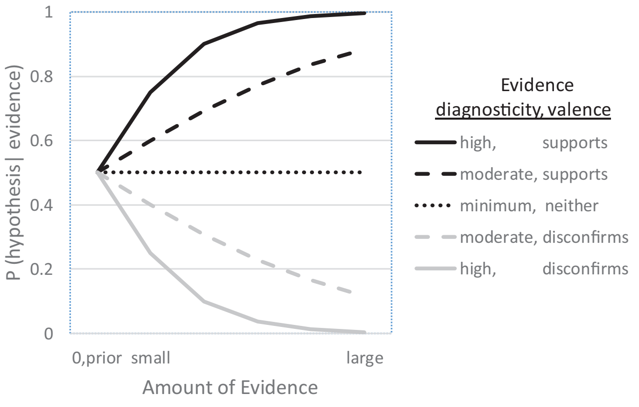


Figure 1. How the valence (supportive vs. disconfirmatory), amount and diagnosticity of evidence affects posterior degree of belief in a claim according to Griffin and Tversky's model, which is based on Bayes's theorem. See the online article for the color version of this figure.

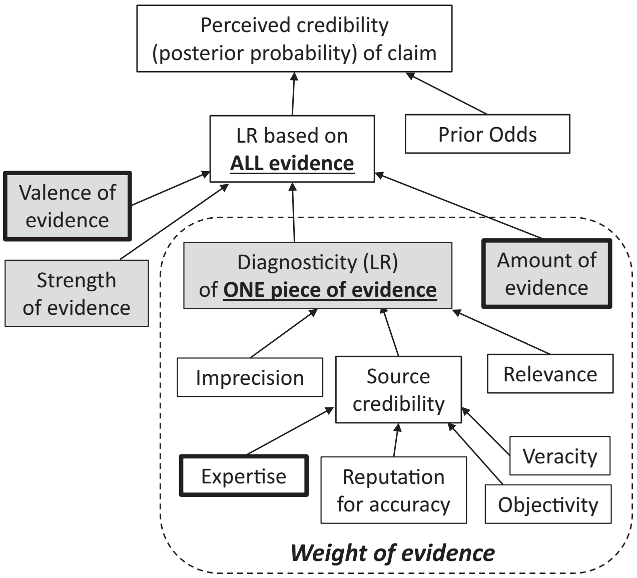


Figure 2. Conceptual model showing influences of credibility cues on perceived credibility. Bolded nodes were manipulated in Study 1. Gray nodes are part of the Bayesian updating model discussed in the text and described in Equation 2. LR means likelihood ratio.

Conceptual versus mathematical models. The diagram in F2 Figure 2 is intended as a tentative, high-level sketch of how many of the credibility-related constructs discussed in the literature might be related. This is a conceptual model that only describes *what* credibility factors influence each other and ultimately degree of belief, but not *how* this happens. The top two levels of the hierarchy represent Bayes's theorem (Equation 1). The third level (gray boxes in figure) shows Griffin and Tversky's decomposition of the overall likelihood ratio ($\mathrm{LR}_{\mathrm{ALL}}$) in Equation 2. The figure also shows other aspects of the weight of evidence that, according to the credibility literature, are thought to influence diagnosticity. This conceptual model is not specific enough to engender testable predictions. In contrast, the Bayesian updating model in Equation 2 makes testable predictions because it describes precisely *how* information from various credibility cues should be integrated in influencing degree of belief.

In this section, we described Bayesian belief-updating models as psychological models that frame traditional credibility-related constructs in a mathematical framework. Hereafter, we use the term *perceived credibility* to refer to a person's degree of belief in a claim after considering evidence and we interpret the credibility of evidence in terms of the cues comprising weight of evidence.

Empirical Research on Heuristic Credibility Cues

In this section, we review empirical research relevant to the credibility models described above and to Experiment 1. Empirical research specifically relevant to Experiment 2 is reviewed in the introduction to that study.

Reputation and Endorsements

Although *reputation* is sometimes interpreted as the name recognition of a source (e.g., Petty, Cacioppo, & Schumann, 1983), it has a more important meaning that goes beyond this. Yaniv and Kleinberger (2000) define reputation as being known for providing accurate (i.e., valid) information. Their study showed how people can learn about a source's reputation in the latter sense from repeated personal experiences of following the source's advice and receiving feedback about its accuracy. This suggests that a good reputation can be based on a history of providing accurate information. They also found that people sought out advice from high-reputation more than low-reputation sources and weighted high-reputation advice more heavily when making decisions with monetary payoffs. When people lack personal experience with an information source, they may rely on *endorsements* or *censures* to judge its accuracy. These can be made through person-to-person communication or through specialized social procedures, for example, a reputation system in an online forum (Fox & Duggan, 2013).

A study by Corner and Hahn (2009), which is discussed further below, found that belief in a claim increased with source reputation. Studies have found that participants trusted information on informational websites and online forums more when it came from a higher-reputation source (Ayeh, 2015; Hilligoss & Rieh, 2008; Metzger et al., 2010; Yi, Stvilia, & Mon, 2012; Zhang & Watts, 2008). Studies of actual online marketplaces that used reputation systems found that sellers with higher reputation had more income than those with lower reputation (Moreno & Terwiesch, 2014; Resnick, Zeckhauser, Swanson, & Lockwood, 2006).

Amount of Information

Corner and Hahn (2009) found that participants believed more strongly in a claim when it was based on more evidence. In Nisbett, Krantz, Jepson, and Kunda (1983), participants sometimes appropriately used sample-size information when using evidence to assess claims. Participants using websites and online forums reported that information was more credible to the extent that it could be corroborated from other sources (Hilligoss & Rieh, 2008; Metzger et al., 2010; Yi et al., 2012).

Expertise

The expertise of an information source is usually seen as reflecting education, experience and knowledge in a domain and is often measured by credentials (Ohanian, 1990). Traditional cre- AQ: 11 dentials include degrees and job experience, whereas online credentialing can consist of things like the number of posts in an online forum (Metzger et al., 2010). Early research on persuasion showed that participants acted on the advice of a credentialed expert (Dr.) much more than a noncredential person (Mr.; Crisci & Kassanove, 1973). Participants rated the credibility of information on a health website higher when the author had a relevant degree and work experience than when these credentials were lacking (Eastin, 2001). Lo and Yao (2019) found that hotel reviewers on a AQ: 12 website with community-generated content were judged as more credible to the extent that they had higher expertise (i.e., more reviews posted).

Model Evaluations

Corner and Hahn (2009) evaluated the Bayesian updating model described above, but only as it applies to supportive evidence.

Their participants saw claims like "anti-inflammatory drug X has no side effects" and rated how convinced they were of the claim after seeing evidence like "50 [or 2] studies of drug X showed no side effects" that was reported in a reputable scientific journal or a little-known website. Because all evidence was supportive, valence was not manipulated. Thus, Corner and Hahn manipulated amount of information and diagnosticity (as cued by reputation for accuracy). Participants became more convinced of claims as the amount of information increased and this increase was greater with high than low diagnosticity. This supported the diagnosticity by amount of information interaction shown in the top half of Figure 1. In Experiment 1, we used both supportive and disconfirmatory evidence, which allowed us to test aspects of the Bayesian model that Corner and Hahn did not.

Experiment 1

In Experiment 1, participants judged the credibility of information displayed in a simulated online health forum with community-generated content. To minimize participants' domain knowledge, the forum focused on pet health and used uncommon pets. Online forums can be found on many topics, including health, politics, and consumer information (Fox & Duggan, 2013). A key problem with unmoderated forums is that information is provided by nonprofessionals who are often unknown to other forum members. In response, developers have created reputation systems, which allow forum members to give evaluations of authors and forum content ranging from positive (endorsements, e.g., 5 stars) to negative (censures, e.g., 1 star). Most reputation mechanisms include some peer feedback method, for example, an average of forum users' ratings (Dellarocas, 2003; Xu, 2013). Reputation systems allow community members to leverage the experience and feedback of a community of peers.

Independent Variables

Participants saw a series of forum posts, with each post containing a question and an answering comment that also contained information about the comment's author. We investigated how variation in three credibility cues—expertise, reputation valence, and amount of reputation information—influenced participants' judgments of the credibility of the comments and their authors. *Expertise* was manipulated using three levels of credentials and domain experience for comment authors. *Reputation valence* (supportive or disconfirmatory) and *amount of reputation evidence* (high or low) were manipulated using star ratings, which signified endorsements or censures in the forum's reputation system. This expertise (3) by valence (2) by amount of evidence (2) design manipulated two aspects of the Bayesian updating model in Equation 2, valence and amount of evidence. In addition, although expertise does not fit directly into this model, expertise may influence diagnosticity (see Figure 2), which is part of the model.

A Generalization Factor

In each post, participants saw separate reputation ratings for the comment and its author. For example, a single post would contain a textual comment along with the information that forum members had given the comment either a high (4 to 5 stars) or a low (1 to 2 stars) average rating on a 5-point scale and that the rating was based on few or many member ratings. The post contained the same kinds of star ratings of the comment's author. Thus, the reputation cues had two referents, comments and authors. Across posts, reputation valence and amount of reputation information were varied independently for the two reputation-cue referents.

Manipulating reputation information for both comments and authors made our credibility displays more representative of the type of credibility information presented in online forums, because some forums allow users to give separate reputation ratings of individual comments and of their authors. The credibility literature we reviewed above supports this practice because it demonstrates that people make credibility judgments about both factual claims (information) and their sources. Varying the referent of the reputation cues allowed us to test whether the effects of reputation valence and amount of information generalized to both comment and author cues, which increased the external validity of the study. We did not expect that credibility judgments would differ for comment versus author cues and did not advance hypotheses regarding this factor.

Here we explain our assumption that the reputation (star) ratings manipulated valence and not diagnosticity. Based on the assumptions underlying the Bayesian updating models, we assumed that participants would interpret high reputation ratings (e.g., 5 stars) as high-diagnosticity evidence supporting the claim and low reputation ratings (e.g., 1 star) as high-diagnosticity evidence disconfirming the claim. Importantly, low reputation ratings were expected to lead participants to disbelieve the claim rather than to see the evidence as uninformative. To see this, consider that average reputation ratings of 1 or 5 (on the 1 to 5 scale) show unanimous agreement—and little uncertainty—among the raters (forum members) that the claim in the comment has a bad or good reputation, respectively. Because reputation influences diagnosticity, these high-diagnosticity ratings (1 or 5) should push degree of belief toward 0 or 1, respectively. In contrast, a reputation rating of 3 reflects uncertainty, as it could be based on high disagreement among raters or high agreement on an uncertain, moderate rating of 3. Thus, a 3 rating is uninformative and should not change degree of belief much, which is the definition of low diagnosticity. This means that diagnosticity was not manipulated in Experiment 1, as participants never saw low-diagnosticity information (3 stars) and only saw high-diagnosticity evidence for or against the claim. Instead, the high versus low reputation ratings manipulated the *valence* of evidence. Participants' credibility judgments in this study supported this assumption; that is, low reputation ratings (1 or 2 stars) led participants to give very low credibility judgments like "very untrustworthy" or "very inaccurate" instead of moderate judgments like "neutral."

Study Design

To summarize the design, our hypotheses concerned how variation in expertise (three levels), reputation valence (two), and amount of reputation information (two) influenced credibility judgments. To test whether the effects of valence and amount of information generalized to both comment and author cues, we varied these two independent variables independently for both comments and their authors. Thus, each participant was presented with forum posts containing 48 unique combinations of credibility cues: expertise (three) by comment reputation (two) by comment amount of information (two) by author reputation (two) by author amount of information (two). To test for the effects of the three independent variables and also test for generalization across the two types of reputation cue, we ran two separate expertise by valence by amount of information ANOVAs, one for comment cues and one for author cues.

Dependent Measures

While viewing each post, participants judged its credibility by responding to four questions, which had two different targets. Three questions targeted the comment (and its claim); that is, participants judged the accuracy and trustworthiness of the comment and their intention to act on it. One question targeted the source of the claim, that is, participants judged the believability of the author. The introduction defined each of these constructs in the context of the Bayesian updating model. Other researchers have used these questions to measure credibility (Ayeh, 2015; Chaiken & Maheswaren, 1994; Flanagin & Metzger, 2013; Hovland & Weiss, 1951; Johnson & Kaye, 2015; Sussman & Siegal, 2003; Zhang & Watts, 2008). We refer to these four questions as assessing perceived or judged credibility. AQ: 13

We assumed that the three questions targeting the comment measured the Bayesian construct of degree of belief in a claim. Because the Bayesian model describes factors affecting belief in a claim, the model predictions seem to apply directly to these three credibility judgments. We assumed that the question targeting the author assessed source credibility. In the General Discussion, we discussed the question of whether our assumptions about the mapping of our questions onto belief in a claim versus source credibility were reasonable.

Hypotheses

Valence and amount of information. We assumed that evidential valence in the Griffin and Tversky model (Equation 2) mapped onto our independent variable of reputation valence and that amount of evidence mapped onto the number of ratings underlying each average reputation rating. We hypothesized that as amount of information increases, credibility judgments will increase if the ratings are supportive and decrease if they are disconfirmatory. This hypothesis is predicted by Griffin and Tversky's model, as discussed in the section on *Model Predictions.* The model predictions are shown graphically in Figure 1. To our knowledge, this interaction has not been tested in the credibility literature. A corollary of this interaction hypothesis is that credibility judgments will be higher for supportive than disconfirmatory valence (main effect).

Expertise. As discussed earlier, expertise is a cue to source credibility (Einhorn & Hogarth, 1985), which is a cue to the diagnosticity of evidence (Corner & Hahn, 2009), which, according to Griffin & Tversky's model, influences posterior degree of belief. Therefore, we expected participants to interpret advice from low- versus high-expertise sources as having low versus high diagnosticity, respectively. This led us to hypothesize that greater expertise would lead to larger credibility judgments.

We followed Corner and Hahn (2009) and others who have used Bayesian argumentation models (Hahn & Oaksford, 2007) by using these models in a qualitative manner to justify our predictions, rather than conducting quantitative model fits and comparing the fits of competing models. Quantitative model comparisons are necessary in future work. However, qualitative applications of mathematical models can be useful in the early stages of applying these models in a particular domain. For example, they can generate predictions that may not have been tested in earlier research, such as our prediction of a valence by amount of information interaction.

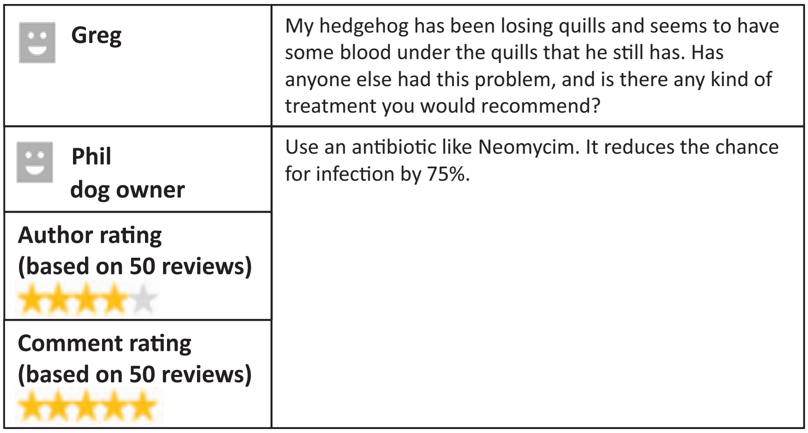
Method

The research in both experiments complied with the American Psychological Association Code of Ethics and was approved by the Institutional Review Board at Clemson University. Informed consent was obtained from each participant.

Participants. We could not find prior experiments to estimate the expected size of the expertise effect or the valence by amount of information interaction. However, Corner and Hahn (2009) found that amount of information accounted for 29% of the variance in credibility judgments, a large effect (Cohen, 1992). Because we predicted that the main effect of amount of information would change credibility in opposite directions for supportive and disconfirmatory valence, we expected that this interaction would have a large effect size as well. Given the repeated-measures study design, power of .8 required 24 and 15 participants when detecting a medium ($f = .25$, $R^2 = .13$) and medium-large effect size ($f = .325$, $R^2 = .18$), respectively, according to G*Power (Faul, Erdfelder, Lang, & Buchner, 2007) and Cohen (1988, 1992). Twenty Clemson University undergraduates (10 female) from 18 to 21 years old ($M = 19.4$) participated. Additionally, as discussed below, the power of the study was increased because analyses were based on cell means based on multiple trials per mean (Krull & MacKinnon, 2001; Morey, 2016). No participants had prior knowledge of illnesses common in household pets.

Design. The 3 × 2 × 2 × 2 × 2 within-subjects design, which is described above, yielded 48 types of posts. In the first block of 48 trials, participants saw a single scenario involving a question about a hedgehog losing quills and an answering comment (see Figure 3). Within a block, the credibility cues changed, but the question and comment did not. The second block used a different pet-health scenario. Trial order was randomized within each block. This was done separately for each participant. On each trial, participants responded to four perceived-credibility questions while viewing one of the 48 cue combinations. F3

Materials. An example of the information displayed on each trial is shown in Figure 3. Disconfirmatory reputation valence was indicated by 1 or 2 stars and supportive valence by 4 or 5. Low amount of information was indicated by 1 to 10 reviews and high information by 41 to 50. High, moderate and low domain expertise were indicated by "veterinarian," a person owning the type of pet in the question, or an unrelated-pet owner, respectively. While each scenario was visible, the participant answered four questions: How trustworthy is the answer? How likely would you be to continue looking for answers? How accurate was this answer? How believable is the person who wrote this answer? Seven-point response scales were used, with the midpoint labeled as "neutral" and the low and high endpoints, respectively, labeled as: very untrustworthy, very trustworthy; very unlikely (to continue), very



Greg	My hedgehog has been losing quills and seems to have some blood under the quills that he still has. Has anyone else had this problem, and is there any kind of treatment you would recommend?
Phil dog owner	Use an antibiotic like Neomycim. It reduces the chance for infection by 75%.
Author rating (based on 50 reviews) ★★★★☆	
Comment rating (based on 50 reviews) ★★★★★	

ONLINE COLOR

AQ: 26

Figure 3. Sample scenario showing a question, an answering comment, and credibility cues. See the online article for the color version of this figure.

likely; very inaccurate, very accurate; very unbelievable, very believable. The likely-to-continue-looking question is the inverse of a common credibility question about intention to use information.

Procedure. The instructions for participants indicated that the study was investigating how people determine what information is reliable when browsing the web and explained the elements in each display. Participants completed 96 trials in about 25 min.

Results and Discussion

The four survey responses for each post assessed perceived credibility. Greater perceived credibility meant that participants: considered the comment to be more trustworthy or accurate, considered the author to be more believable, or considered themselves less likely to continue looking for information. After reversing the likely-to-continue judgments, higher numbers represented higher credibility judgments. We hypothesized main effects of credibility judgments increasing with expertise and valence and an interaction such that credibility judgments would be moderate when there was little information (few reputation ratings), high when there were many supportive ratings, and low when there were many disconfirmatory ratings.

Analysis of composite credibility judgments. Our first analysis tested these hypotheses using a composite credibility variable formed by reversing the likely-to-continue judgments and averaging participants' judgments on the four credibility questions on each trial. Data analysis used repeated measures ANOVA. Because we did not have a factorial design in which valence, amount of information, and reputation-cue referent (comment vs. author) were crossed, we could not test the valence by amount of information interaction within a single, factorial model. Therefore, we ran two factorial ANOVAs, 3 (expertise) × 2 (comment valence) × 2 (comment amount of information) and 3 (expertise) × 2 (author valence) × 2 (author amount of information). For example, in the ANOVA for effects of manipulating *comment* valence and amount of information, each participant contributed 12 condition means. Each of these means were computed by averaging the composite variable over the eight trials created by crossing the two blocks with the four trials where *author* valence and amount of information were manipulated. Because we tested the same hypotheses in two analyses, we reduced the likelihood of inflation of Type I error by using a Bonferroni correction, that is, alpha was set to .025.

Statisticians have pointed out that when the unit of analysis in an ANOVA is an aggregate (e.g., mean) based on a number of individual observations (eight trials, in these analyses), as the number of observations per mean increases, the within-cell error variance decreases and the test becomes more powerful (Barcikowski, 1981; Krull & MacKinnon, 2001; Morey, 2016). This may explain why, despite the Bonferroni correction, some of the effects in the ANOVAs for Experiment 1 were significant although effect sizes were negligible. To be conservative, when significant effects had negligible effect sizes, we based our conclusions regarding whether hypotheses were supported on effect size, not significance.

Regarding effect size, Bakeman (2005) pointed out that repeated responses tend to be positively correlated within subjects and that within-subjects variance increases with this correlation. Therefore, η_p^2 overestimates effect size because it removes within-subject variance from the denominator. Bakeman recommends ameliorating this problem by using Olejnik and Algina's (2003) generalized eta squared (η_G^2), which includes within-subjects variance and other subjects-related sources of variance in the denominator. This leads to smaller effect sizes than η_p^2 and to effect sizes that are comparable across studies that investigate the same factors but employ different designs. Given our completely within-subjects design, we estimated effect sizes using η_G^2, which represents percentage of variance accounted for. Based on Cohen (1988, 1992), we interpreted η_G^2 values of .02 as small, .13 as medium, and .26 as large. AQ: 14

Expertise. The means and significance levels for the expertise effect were identical in the models based on comment versus author cues. In support of the hypothesis, composite credibility judgments increased significantly with expertise, from 3.26 (SE = 0.12) at low expertise, to 3.66 (SE = 0.14) at moderate expertise, and 3.62 (SE = 0.14) at high expertise, $F(2, 38) = 10.4$, $p < .001$. The expertise effect was smaller when comment cues were manipulated, $\eta_G^2 = .10$, than when author cues were manipulated, $\eta_G^2 = .19$ (Table 1 summarizes T1 the effect sizes for these analyses).

Table 1
Effect Sizes (η^2_G) and Significance Decisions (see) for Effects of Credibility Cues on Composite Credibility Judgments for Experiment 1*

	Effects of referent (R) of reputation cues (comment vs. author)			Effects of diagnosticity (D) of reputation cues		
Measure	Comment cues	Author cues	$\Delta\eta^2_G$	High-diag. cues	Low-diag. cues	$\Delta\eta^2_G$
Expertise (E)	.104*	.187*	.083	.085*	.213*	.128
Valence (V)	.784*	.676*	.108 (VR)	.805*	.510*	.295 (VD)
Amount of info. (A)	.062*	.025*		.064*	.015*	
V × A interaction	.209*	.065*	.144 (VAR)	.182*	.066*	.116 (VAD)

Note. Shown on the left are the results for the original analysis where separate ANOVAs were run for comment versus author cues. On the right are the results for the second analysis where separate ANOVAs were run for high- versus low-diagnosticity cues.
* $p < .025$; significance tests were not conducted for the interaction effects under $\Delta\eta^2_G$.

F4 ***Valence and amount of information.*** Figure 4A shows the valence by amount of information interaction for both the comment- and author-cue models. Credibility was judged to be higher with supportive than with disconfirmatory valence for comment cues, $F(1, 19) = 113.8$, $p < .001$, $\eta^2_G = .78$, and author cues, $F(1, 19) = 123.4$, $p < .001$, $\eta^2_G = .67$. These effects were very large. The valence by amount of information interaction was significant for both comment cues, $F(1, 19) = 61.3$, $p < .001$, $\eta^2_G = .21$, and author cues, $F(1, 19) = 18.5$, $p < .001$, $\eta^2_G = .06$. These findings supported our hypotheses. Interestingly, the effect size for the valence main effect and the valence by amount of information interaction were much larger for comment than author cues. This was unexpected.

Concerning the other effects in the factorial model, there was an unhypothesized main effect such that, for both comment and author cues, more information led to significantly higher credibility judgments (comment: $M = 3.65$, $SE = 0.10$; author: $M = 3.57$, $SE = 0.11$) than less information (comment: $M = 3.38$, $SE = 0.14$; author: $M = 3.45$, $SE = 0.13$). For comment cues, $F(1, 19) = 30.1$, $p < .001$, $\eta^2_G = .062$, and for author cues, $F(1, 19) = 8.98$, $p < .001$, $\eta^2_G = .025$. Because these main effects were part of significant (and stronger) interactions, we do not focus on them. There were two other unpredicted effects that were significant and had non-negligible effect sizes, the expertise by valence interaction for comment cues, $\eta^2_G = .027$, and the three-way interaction for author cues, $\eta^2_G = .027$. These are not discussed further because of their

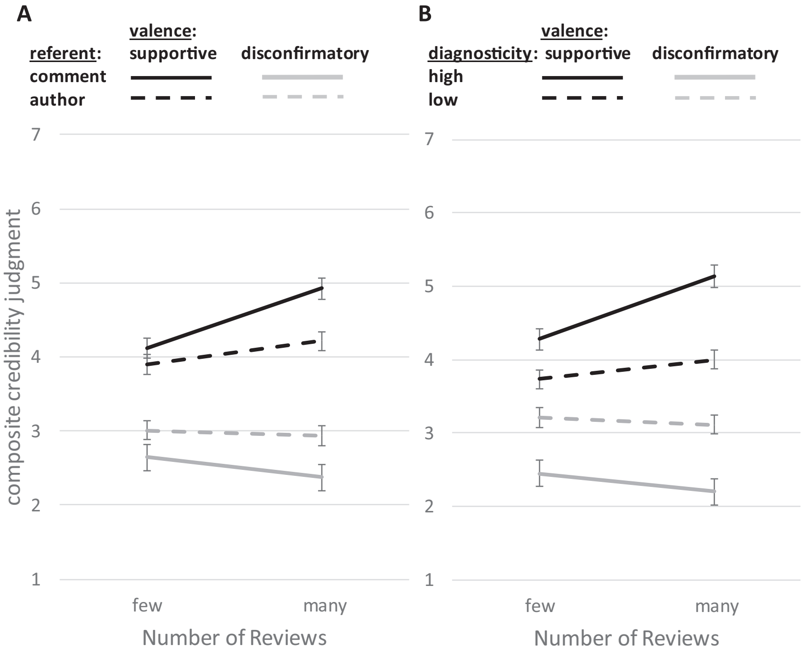


Figure 4. Both graphs show the effect of varying reputation valence (supportive vs. disconfirmatory) and amount of reputation information (few vs. many reviews) on composite credibility judgments. Panel A shows the effects when the composite variable was averaged within comment and author cues. Panel B shows the effects when the composite variable was averaged within high versus low diagnosticity cues. Error bars represent standard errors.

low effect size. All other effects had negligible effect sizes, $\eta_G^2 <$ .02.

In summary, the hypotheses regarding expertise, valence, and the valence by amount of information interaction were all supported in the analysis of both the comment- and the author-cue data. An unexpected finding was that all three of the hypothesized effects differed in size depending on whether the analysis focused on comment or author cues. Table 1 shows that these effect-size difference ranged from .08 to .14. For expertise, the larger effect sizes with author than comment cues make sense when one considers that expertise pertains more directly to the author than the comment.

Analysis of individual credibility variables. Given that we varied reputation-cue referent only to test the generalizability of the effects of reputation valence and amount of information, we were unsure why the effects of valence and amount of information seemed to be stronger with comment than author cues. To explore this question, for both comment and author cues, we conducted separate expertise by valence by amount of information ANOVAs for each of the four credibility questions. These analyses also allowed us to see whether the findings for the composite variable were similar to those for each of the individual credibility variables. Because we conducted eight ANOVAs in this phase of the analysis, we set alpha to .0625.

F5

Here we present an overview of the findings from these ANOVAs. The statistical evidence, which supports the conclusions made here, is provided in Appendix B. The expertise effects for the individual variables and the composite variable were similar. Regarding valence and amount of information, Figure 5 shows that the findings for the composite variable were also evident with the trustworthiness, accuracy and continue variables. That is, the valence main effect for these variables was much larger for comment cues (η_G^2 of .67, .63, 27, respectively) than author cues (.31, .13, .02) and the valence by amount-of-interaction was larger for comment cues (.11, .09, .04) than author cues (.01, .01, .001).

The results were quite different for the believability variable. The difference in effect sizes for valence was *reversed*, with much larger effects with author cues (.60) than comment cues (.25). Also, the valence by amount of information interaction was *not* larger for comment than author cues; instead, these interactions were similar in size (both .04). Thus, the effect sizes for the valence effect and the interaction were dissociated based on reputation-cue referent. Especially for the valence effect, the effects of comment versus author cues found for the composite variable and for trustworthiness, accuracy and continue were reversed for believability.

We suggest a post hoc explanation of this dissociation that involves the *relevance* credibility cue discussed earlier. Perhaps the dissociation occurred because the trustworthiness, accuracy and continue questions pertained to the comment, whereas the believability question pertained to the author. This could have led participants to think that forum-members' reputation ratings for *comments* were more relevant to trustworthiness, accuracy and continue judgments than their author reputation ratings. In the conceptual model in Figure 2, relevance is a cue to diagnosticity. Thus, because of their greater relevance, comment reputation ratings for these three questions should be more diagnostic of posterior degree of belief than author reputation ratings. Conversely, participants may have concluded that forum-members' reputation ratings for *authors* were more relevant to and diagnostic of author credibility judgments (i.e., believability) than comment reputation ratings. More generally, perhaps the perceived relevance of a credibility cue—and therefore its diagnosticity—will be greater when the referent of the cue (e.g., comment or author) matches the target of the credibility judgment and smaller when the cue referent and judgment target are mismatched.

Thus, by varying both the referent of the credibility cue and the target of the credibility questions, we may have *inadvertently* varied relevance and therefore diagnosticity. Under this interpretation, we could assume that participants interpreted the following to be high diagnosticity cues: comment cues for the trustworthiness, accuracy and continue questions and author cues for the believability question, with the reverse assignment indicating low diagnosticity cues. Under this post hoc assumption, a comparison of the data in Figure 5 and the Bayesian model predictions in Figure 1 suggests that our findings for all four credibility questions support the full diagnosticity by valence by amount of information

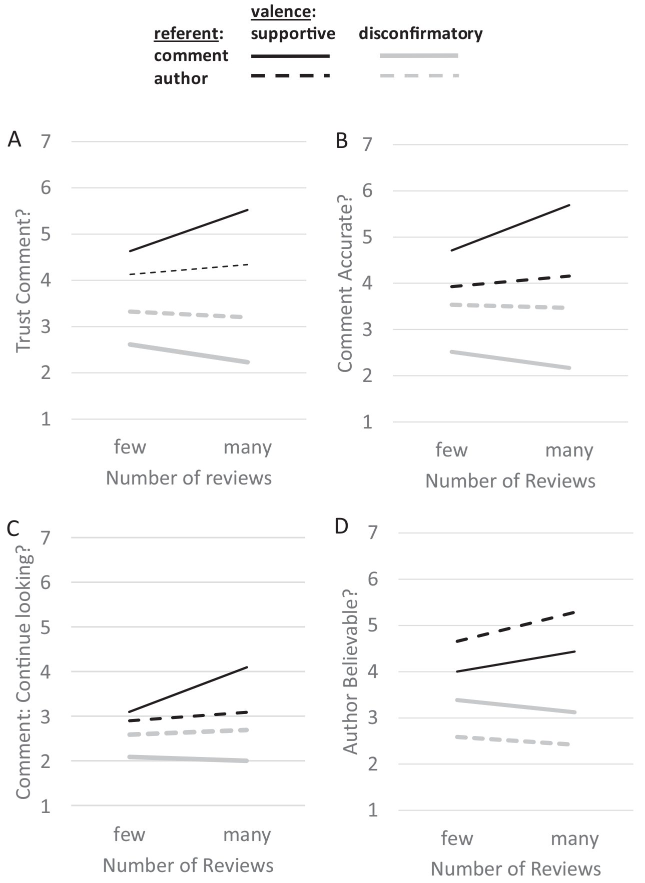


Figure 5. Effect of varying reputation valence (supportive vs. disconfirmatory) and amount of information—for both comment and author ratings—on four perceived-credibility measures. A *y* axis label of 7 represents judgments of "very trustworthy | accurate | believable" or "not continuing to look for information;" a label of 1 represents judgments of "very untrustworthy | inaccurate | unbelievable" or "continuing to look." Solid lines show effect of varying the valence and amount of information of comment ratings. Dashed lines show same effect for varying author ratings.

interaction predicted by the Bayesian model. That is, as amount of information increases, judged credibility approaches extreme values (maximum for supportive evidence and minimum for disconfirmatory evidence) faster with high- than low-diagnosticity information. The data for each question also appear to support the diagnosticity by valence interaction predicted by the model.

There are theoretical arguments for considering the relevance of evidence as a cue influencing posterior degree of belief in a claim. Pearl (1988) noted that in Bayesian terms, evidence is *relevant* to a claim if adding the evidence to prior knowledge changes the probability that the claim is true. Hahn and Oaksford (2007) found evidence suggesting that many logical fallacies are better interpreted as weak inductive arguments than as invalid deductions. This interpretation was based on a Bayesian argumentation model like the one used in the current study; furthermore, it depended on using the relevance of evidence as a cue affecting degree of belief in a claim.

Composite analysis based on diagnosticity. According to the above interpretation, in the previous composite analyses (where we tested the effects of manipulating valence and amount of information on the composite credibility variable separately for comment and author cues) comment cues consisted of three high-diagnosticity cues and one low-diagnosticity cue, whereas author cues consisted of three low-diagnosticity cues and one high-diagnosticity cue. To explore the implications of our post hoc interpretation regarding diagnosticity, we conducted two more ANOVAs in which we replaced the comment versus author cue distinction with the high- versus low-diagnosticity cue distinction. Whereas in the previous analysis we created the composite credibility variable by averaging over the four credibility variables within comment cues and within author cues, in the new analysis we created the composite variable by averaging over these variables within high-diagnosticity and within low-diagnosticity cues. As before, we set alpha to .025 and ran two ANOVAs: 3 (expertise) × 2 (high-diagnosticity valence) × 2 (high-diagnosticity amount of information); and 3 (expertise) × 2 (low-diagnosticity valence) × 2 (low-diagnosticity amount of information). We tested the same hypotheses as in the ANOVAs based on comment versus author cues. We do not consider these new ANOVAs to be independent hypothesis tests but rather tests conducted in the context of a different extraneous variable.

In both of the ANOVAs using the composite variable averaged within diagnosticity levels, the hypothesized effects—expertise, valence, valence by amount of information—were all significant (see Table 2) and had nonnegligible effect sizes (see Table 1). (Of the eight unhypothesized effects across the two ANOVAs, all had negligible effect sizes except one, the amount of information effect in the high-diagnosticity ANOVA.) Figure 4 shows how cue valence and amount of information influenced overall credibility for the diagnosticity (4B) and comment versus author cue (4A) analyses. The main difference between the two graphs seems to be in the effect size of the valence main effect. This effect was smaller ($\eta_G^2 = .510$) for low-diagnosticity cues than for author cues ($\eta_G^2 = .676$), although still very large.

Importantly, comparing Figure 4B with Figure 1 suggests that the findings of the diagnosticity analysis are consistent with predictions of the Bayesian model that went beyond our hypotheses. That is, Figure 4B shows the diagnosticity by valence interaction and the full diagnosticity by valence by amount of information interaction predicted by the model. Because the data for the high versus low-diagnosticity lines in Figure 4B came from separate analyses, we could not test these interactions in an ANOVA. However, the difference in the size of the valence effect for high- versus low-diagnosticity cues is an estimate of the size of the diagnosticity by valence interaction. This difference, $\Delta\eta_G^2 = .29$ (see Table 1), showed a large effect size.

Table 2
Results of Key Significance Tests for Separate ANOVAs for High- Versus Low-Diagnosticity Cues

	High-diagnosticity cues			Low-diagnosticity cues		
Measure	*df*	*F*	*p*	*df*	*F*	*p*
Expertise (E)	2,38	10.4	.003	2,38	10.4	.003
Valence (V)	1,19	118.5	.001	1,19	68.2	.001
Amount of info. (A)	1,19	29.5	.001	1,19	7.71	.012
V × A interaction	1,19	53.4	.001	1,19	23.3	.001

Note. Degrees of freedom for *F* tests are 1,19.

Similarly, the difference in the size of the valence by amount of information interaction for high- versus low-diagnosticity cues is an estimate of the size of the diagnosticity by valence by amount of information interaction. This difference, $\Delta\eta_G^2 = .12$, represented a small to medium effect size. Thus, the diagnosticity analysis provides tentative support for these predictions of the Bayesian model. This support is tentative because the categorization of dependent variables into high versus low diagnosticity was post hoc and because we could not directly test the significance of the critical interactions.

In this section, we have been using the term diagnosticity to mean "diagnosticity as cued by relevance." However, because diagnosticity is also cued by expertise (see the conceptual model, Figure 2), perhaps we should expect that expertise would interact with valence and amount of information in the same way that relevance-based diagnosticity did. Although expertise showed its hypothesized main effects on perceived credibility (with η_G^2 ranging from .085 to .213) it showed mostly negligible ($\eta_G^2 < .02$) interactions with valence and amount of information, with a few close-to-negligible interactions ($\eta_G^2 < .03$). In contrast, the interactions of relevance (diagnosticity) with valence and valence by amount of information had much larger effect sizes .29 and .12. We suggest that relevance but not expertise interacted with valence and amount of interaction because relevance, valence and amount of information all were defined in terms of the reputation cues (the star ratings), whereas expertise was a separate credibility cue. This issue is discussed further in the General Discussion.

Summary. Analyses using a composite perceived-credibility variable supported three hypotheses, which were based on the Bayesian updating model. Perceived credibility increased with expertise and reputation valence. It also increased with amount of information when cue valence was supportive and decreased with more information when valence was disconfirmatory. This occurred regardless of whether the composite variable was averaged within comment and author cues or within diagnosticity levels. The diagnosticity analysis gave tentative support for further model predictions that we did not hypothesize.

Experiment 2

Experiment 1 focused on how external credibility cues affected people's subjective perceptions of the credibility of a claim or its source (research question 1). Experiment 2 focused on how external credibility cues and perceptions of credibility influenced decision making. We investigated how conflict between credibility and utility influenced people's choices (question 2) and whether perceived credibility mediated the effects of credibility cues on choice (question 3).

Handling Credibility–Utility Conflict

Regarding the first question, many decision-making researchers have provided evidence that people sometimes show *ambiguity aversion,* that is, not choosing a high-utility decision alternative with outcomes (or probabilities of outcomes) that are ambiguous and instead choosing a lower-utility alternative that is less ambiguous (Budescu et al., 2002; Curley, Yates, & Abrams, 1986; Kunreuther, Hogarth, & Meszaros, 1993; Koch & Schunk, 2013; Maffioletti & Santori, 2005; Ritov & Baron, 1990). A number of these researchers have noted that there are two kinds of uncertainty in decision making. The first kind—called risk—is uncertainty about whether a decision outcome will occur. Risk is quantified by probabilities. The second kind—sometimes called second-order uncertainty—is uncertainty about the "reliability, credibility or adequacy" (Ellsberg, 1961, p. 659) of the probability of an outcome (Curley & Yates, 1985; Einhorn & Hogarth, 1985; Frisch & Baron, 1988) or the outcome itself (Budescu et al., 2002). These researchers conceptualize ambiguity as second-order uncertainty. Some researchers prefer the term vagueness to ambiguity (Budescu et al., 2002; Heath & Tversky, 1991). Beyond the general notion of ambiguity as second-order uncertainty about outcomes and their probabilities, researchers have proposed a variety of cues that influence ambiguity, including precision (Budescu et al., 2002), source credibility and amount of information (Einhorn & Hogarth, 1985), awareness of missing information (Ritov & Baron, 1990), and instructions to feel confident in versus uncertain about evidence (Hogarth & Kunreuther, 1989).

AQ: 15

AQ: 16

Research on ambiguity in decision making and credibility in argumentation is often pursued as if these were unrelated topics. However, an important question concerns whether people are using similar cognitive processes when they avoid choosing a decision alternative because its outcomes are uncertain (i.e., ambiguous) and when they believe a factual claim less strongly because the relevant evidence is uncertain (i.e., has low credibility). We suggest that the answer to this question may be *yes* and that further research on this question is warranted. Some researchers seem to agree with this conclusion. Tentori, Crupi, and Osherson (2010) suggested that uncertainty about outcomes of a decision alternative may be similar to uncertainty about the information provided by a source during argumentation. More specifically, Einhorn and Hogarth (1985) and Frisch and Baron (1988) claimed that what decision-making researchers call ambiguity is similar or identical to what argumentation researchers call the weight of evidence (which is a factor influencing belief in a claim in Griffin and Tversky's (1992) belief-updating model). The conjecture here is that down-weighting and avoiding decision alternatives with ambiguous outcomes is cognitively similar to assigning a low weight to low-credibility evidence and thus a low degree of belief to claims based on that evidence.

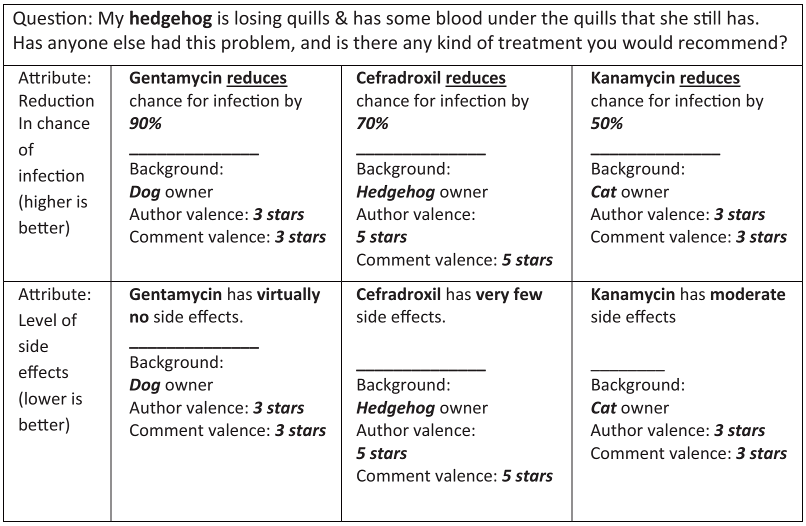
Recognizing that reasoners may down-weight uncertain information in both decision making and argumentation allowed us to operationalize ambiguity about decision inputs using different cues in Experiment 2 than in many prior studies, which have often used imprecision. We varied the ambiguity of outcomes in a decision task using reputation cues (i.e., amount and valence of information) and expertise in an online health forum, as in Experiment 1. It seems plausible that people consider the reputation and expertise of the sources of information they base their decisions on, especially when they gather information about decision outcomes on their own. For example, parents deciding whether to have their child get the measles-mumps-rubella vaccination may want to evaluate the credibility of a website claiming that this vaccine causes autism. Also, people deciding how vigorously to implement social distancing during the COVID19 pandemic might be more strongly influenced by estimates about the health risks of the disease that come from more credible sources. Manipulating ambiguity in a decision-making task in terms of credibility cues like reputation and expertise is a novel contribution of this study.

Much of the ambiguity-aversion research has used abstract tasks involving monetary gambles with two decision alternatives (e.g., Koch & Schunk, 2013). A smaller number of studies have used more realistic decisions, but still focused on simple decisions (one attribute and two alternatives). For example, when Kunreuther et al. (1993) gave actuaries insurance pricing scenarios, the actuaries set higher selling prices for insurance against imprecise than precise losses, even though the expected value of both losses was the same. In interviews, some actuaries reported that they set prices by calculating the expected value of the potential loss and then adjusting it upward if the information about the loss was ambiguous. In Experiment 2, we used more complex decisions than in many ambiguity-aversion studies—multiattribute with three alternatives. We manipulated two factors that we predicted would influence the degree to which participants would show ambiguity (vagueness) aversion, the degree of credibility–utility conflict and credibility gain.

Manipulating credibility–utility conflict. Experiment 2 used the domain of pet health, like Experiment 1. Participants made multiattribute noncompensatory decisions where one decision alternative had the best utility, one was second best, and one was worst on every outcome. For each decision, all the outcomes for one alternative had high credibility while all the outcomes for the other two alternatives had low credibility. We varied the degree of conflict between credibility and utility by varying whether the single high-credibility alternative had the best, moderate, or worst utility. Thus, this variable specified the *utility premium* (or loss) participants had to incur to select the high-credibility alternative. The highest premium occurred when the high-credibility alternative had the worst utility.

We did not vary credibility cues factorially, as in Experiment 1, because we wanted a clear contrast between one decision alternative with high credibility across all credibility cues for both outcomes versus two other alternatives with lower credibility on all credibility cues and outcomes. For example, in the decision in Figure 6, both outcomes for the high-credibility alternative (Cefradoxil) have a relevant-pet owner (high expertise) and high (5-star) author and comment reputation ratings based on many raters; while both outcomes for the other two, low-credibility alternatives have a nonrelevant pet owner (low expertise) and

F6



Question: My **hedgehog** is losing quills & has some blood under the quills that she still has. Has anyone else had this problem, and is there any kind of treatment you would recommend?			
Attribute: Reduction In chance of infection (higher is better)	**Gentamycin reduces** chance for infection by ***90%*** Background: ***Dog*** owner Author valence: ***3 stars*** Comment valence: ***3 stars***	**Cefradroxil reduces** chance for infection by ***70%*** Background: ***Hedgehog*** owner Author valence: ***5 stars*** Comment valence: ***5 stars***	**Kanamycin reduces** chance for infection by ***50%*** Background: ***Cat*** owner Author valence: ***3 stars*** Comment valence: ***3 stars***
Attribute: Level of side effects (lower is better)	**Gentamycin** has **virtually no** side effects. Background: ***Dog*** owner Author valence: ***3 stars*** Comment valence: ***3 stars***	**Cefradroxil** has **very few** side effects. Background: ***Hedgehog*** owner Author valence: ***5 stars*** Comment valence: ***5 stars***	**Kanamycin** has **moderate** side effects Background: ***Cat*** owner Author valence: ***3 stars*** Comment valence: ***3 stars***

Figure 6. Example decision in Experiment 2. The alternative with high credibility has moderate overall utility. There is a small credibility gain of 2 stars (i.e., 5–3) between high- and low-credibility alternatives. Although not shown here, stimulus displays indicated that all reputation ratings were based on high amount of information (many raters).

moderate (3-star) author and comment ratings based on many raters. Unlike in Experiment 1, reputation ratings in Experiment 2 could take any value between 5 stars (maximum credibility) and 1 star (minimum). Amount of information was always high. According to the Bayesian updating model presented earlier, the high, moderate and low reputation ratings in Experiment 2 were expected to lead participants to assign high, moderate, and low degrees of belief, respectively, to the outcome information. Thus, varying the reputation ratings manipulated perceived credibility of the outcomes.

Manipulating credibility gain. If people reject a high-utility decision alternative with low credibility in favor of a lower-utility alternative with higher credibility, it seemed likely that they would only do this when the gain in credibility offsets the loss in utility. This idea led us to a second independent variable: the *credibility gain* participants received when they avoided a low-credibility alternative and instead chose a high-credibility alternative. This manipulation was made using the author and comment reputation valence ratings, for example, 5 versus 3 stars for a small gain and 5 versus 2 stars for a large gain. Credibility gain was not manipulated using expertise because the levels of expertise did not have enough precision; this reduced the strength of the credibility-gain manipulation.

Experiment 2 used a factorial design: three utility premiums (degrees of credibility–utility conflict) by two credibility gains. For each decision, participants chose one alternative, reported their confidence in their choice, and answered the same perceived-credibility questions as in Experiment 1.

Models and Hypotheses

Ellsberg (1961) proposed a decision-making model that incorporated ambiguity. We focused on Budescu et al.'s (2002) model, which combined elements of Ellsberg's model and Prospect Theory (Kahneman & Tversky, 1979). Budescu's model elaborated Prospect Theory based on Ellsberg's model so that it could handle low- as well as high-credibility outcomes. In addition, Budescu's model allowed both outcomes (and their associated utilities) and the probability of outcomes to be ambiguous. Budescu et al.'s model assumes that decision makers behave like the actuaries in Kunreuther et al. (1993). When they have unambiguous, high-credibility information about a decision outcome and its probability, they evaluate the outcome and probability as in Prospect Theory. When they have ambiguous information about either the outcome or its probability, they tend to make worst-case assumptions about the outcome or probability, especially when faced with losses. This leads to ambiguity aversion. The model represents high versus low ambiguity by expressing outcome and probability information in terms of ranges versus point estimates, respectively.

Because Budescu et al. (2002) prefer the term vagueness to ambiguity, we use the former term in describing their model. In the model, the utility of one outcome for one alternative (i.e., one cell in a decision matrix) is,

$$U_{cell} = V[w_x x_{worst} + (1 - w_x)x_{best}] \times f[w_p p_{worst} + (1 - w_p)p_{best}], \tag{3}$$

where x = an outcome, p = a probability of an outcome, *best* = best case, *worst* = worst case, w_x and w_p = vagueness weighting factors for outcomes and probabilities, respectively, and $V[\bullet]$ and $f[\bullet]$ are the Prospect Theory value and decision-weight functions, respectively. We discuss Equation 3 for the medication effectiveness attribute, which was presented to participants using the relative risk reduction statistic (see Figure 6). Note that this statistic is not a probability, so Equation 3 treats it as a utility and uses the $V[\bullet]$ function and the w_x parameter. For example, suppose people read in an online health forum that the effectiveness of a medication is 70% and they imagine that the actual effectiveness can range from 50% to 90%. Thus, x_{worst} is 50% and x_{best} is 90%. If they judge that the source of this information has very high credibility, then w_x is .5 and they think of effectiveness as being at the center of the range (70%). If they judge the source as having very low credibility, then w_x is 1 and they think of a worst-case effectiveness of 50%. (The w_p parameter for probabilities works the same way.) Below, we describe *qualitatively* how this model leads to vagueness aversion. In Appendix C, we give other examples and work them out *quantitatively*.

Utility-premium hypothesis. To see how the model works for decisions like Figure 6, we focus on the effectiveness attribute (the model makes the same predictions for side effects). Gentamycin has the best objective utility because it is most effective (90%). However, it also has low credibility. Cefradoxil has lower utility (70%) but high credibility. If people ignore credibility, they would choose Gentamycin. If they consider credibility as described in Equation 3, they might reason as follows. "Gentamycin is supposedly 90% effective but the person telling me this has low credibility, so it might be less effective than this person claims. In the worst case, it might be only 60% effective, which is less than Cefradoxil. The person recommending Cefradoxil is very credible, so I trust this person's statement that it is 70% effective. So, I'll choose Cefradoxil." People who make this choice are sacrificing (trading off) some utility to choose a less vague (more credible) alternative, that is, they are showing vagueness aversion.

However, the imagined worst-case utility for Gentamycin (60% in the example above) might still be greater than the extremely low objective utility of Kanamycin (50% effective). This should lead people to choose the low-credibility Gentamycin if Kanamycin is the high-credibility alternative, which does *not* show vagueness aversion. The pattern shown in these two model predictions leads to a hypothesis about when people will trade off utility for credibility and when they will not, i.e., when choosing the high-credibility alternative requires sacrificing more utility, this sacrifice will be made less often. We called this the *utility premium hypothesis.* It predicts that people will choose the high-credibility alternative most often when this alternative has the highest utility, less often when it has moderate utility, and least often when it has the lowest utility.

Credibility-gain hypothesis. In the example immediately above (where vagueness aversion does not occur), the credibility gain between the high- and lower-credibility alternatives was small (5 vs. 3 stars), which meant that choosing a high-credibility but low-utility alternative did not gain much credibility. Now we reconsider this example with a larger credibility gain (5 vs. 2 stars), that is, Gentamycin has high utility (90% effective) but very low credibility (2 stars) and Kanamycin has very low utility (50%) but high credibility (5 stars). According to Budescu's model, Gentamycin's very low credibility might cause participants to think as follows. "I don't trust the person recommending Gentamycin at all. So, the 90% effectiveness they claim might be as low as 45%, which is less than the very trustworthy 50% effectiveness rating for Kanamycin. So, I'll chose Kanamycin." Thus, the model chooses the worst-utility but high-credibility alternative (Kanamycin) over the best-utility but low-credibility alternative (Gentamycin) when this choice yields a *large* credibility gain, that is, it exhibits vagueness aversion.

In contrast, in the example just above this one, when choosing the worst-utility Kanamycin yields a *smaller* credibility gain, the model chooses the best-utility Gentamycin despite its low credibility, thus failing to show vagueness aversion. The pattern of model predictions for these two examples leads to the *credibility-gain hypothesis*—that people will choose the high-credibility but low-utility alternative more often when this allows them to gain more credibility. More specifically, this hypothesis states that people will choose the low-utility but high-credibility alternative instead of the high-utility but low-credibility alternative more often when the difference in credibility between these two alternatives is large (e.g., 5 vs. 2 star reputation, respectively) than when this difference is smaller (e.g., 5 vs. 3 stars).

Interaction hypothesis. When the high-credibility alternative had the best utility, people were expected to choose it very frequently without considering other alternatives. In this case, the credibility gain from choosing the high-credibility alternative should not affect choice. This reasoning led us to hypothesize an interaction between the credibility-gain and utility-premium factors, such that the probability of choosing the high-credibility alternative should decline as the utility premium increases from low to high, but this decline should be greater (steeper) with a small than a large credibility gain.

Perceived Credibility as a Mediator

Budescu et al.'s model assumes that decision makers make judgments about the credibility of decision outcomes, which then influence their choices. For each decision in Experiment 2, we had participants judge the credibility of the outcomes for their chosen alternative. These data allowed us to investigate whether participants' explicit credibility judgments statistically mediated the effects of the credibility and utility manipulations on choices. In the data analysis without considering mediation, the main dependent variable used to evaluate the effects of the credibility and utility manipulations was the probability of participants' choosing the high-credibility alternative. In the mediation analysis, we compared the direct effects of the manipulations on this dependent variable with the indirect effects, as mediated by perceived credibility (see Figure 7). F7

The idea that credibility judgments influence choice fits with empirical research suggesting that these judgments are often not ends in themselves, but also serve to guide behavior (Petty &

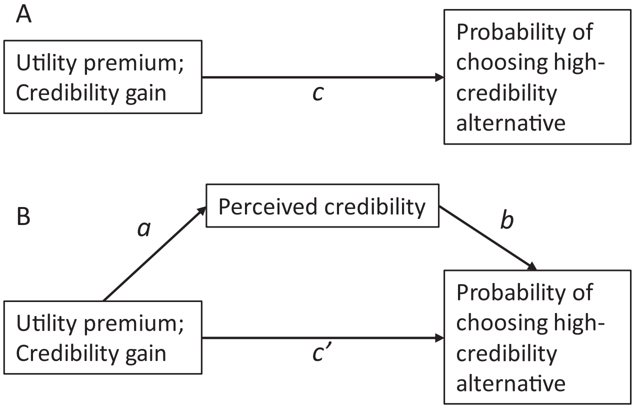

A
Utility premium; Credibility gain
c
Probability of choosing high-credibility alternative
B
a
Perceived credibility
b
Utility premium; Credibility gain
c'
Probability of choosing high-credibility alternative

Figure 7. Conceptual models showing unmediated (A) and mediated (B) effects of utility premium and credibility gain on probability of choosing the high-credibility alternative.

Cacioppo, 1986; Yaniv & Kleinberger, 2000). For example, Crisci and Kassanove (1973) found that mothers who were advised by a psychologist to buy a book on parenting were more likely to buy the book if the psychologist's title was Doctor than Mister.

Method

Participants. Because we were not able to find similar comparison studies, we allowed for a smaller effect size than in the first study. Given the repeated measures study design, detecting a small-medium effect size (f = 0.20) for the utility-premium and credibility-gain effects required 52 participants when power was .8, according to G*Power. Fifty-five Clemson University undergraduates (28 female) participated. Age ranged from 18 to 23 (M = 19.5). None of the participants had advanced knowledge of illnesses common in household pets.

Design. The experiment used a 3 (utility premium) × 2 (large vs. small credibility gain), within-subjects factorial design. High credibility was paired three times with the smallest utility premium (best utility) alternative, six times with a moderate premium, and five times with the highest premium (worst utility). We used this unbalanced design because we expected participants' responses to be very reliable when high credibility was paired with the best utility. Of the 14 problems, five involved a small credibility gain and nine involved a large gain. These nine included two, four, and three problems in the conditions when high credibility was paired with best, moderate, and worst utility, respectively. Each participant saw the same, randomized order of problems.

Materials. Each decision problem contained a question and answering comments in a matrix format. The information was arranged similarly to Figure 6, with the question at the top, three columns containing decision alternatives, two rows containing evaluative attributes, and six cells (comments) providing information about the outcomes of alternatives. In each problem, one alternative dominated (i.e., highest utility on both outcomes), one had the second-best utility on each outcome, and one had the worst utility. In each problem, both outcomes of one alternative had higher credibility and both outcomes of the other two alternatives had lower credibility.

Utility was manipulated in terms of two attributes: degree of reduction in the chance of infection and frequency of side effects. Credibility was manipulated using reputation ratings and expertise, but not amount of information, which was high (reputation ratings from 46 to 50 people) for all outcomes. On each problem, the single high-credibility alternative had author and comment reputation ratings of 4 or 5 stars and the comment was authored by an owner of a pet relevant to the question. The two low-credibility alternatives had author and comment reputation ratings of 1, 2 or 3 stars and a nonrelevant pet owner. Credibility gain was manipulated solely using the comment and author reputation ratings. A large credibility gain compared 5 stars for high- versus 2 stars for low-credibility alternatives or 4 versus 1. A small gain compared 5 versus 3 or 4 versus 2 stars. Location in the matrix (left, center or right alternative) was not confounded with location of the high-utility or the high-credibility alternative.

For each decision, the question and decision matrix were visible until participants completed all responses. First, participants chose one of the alternatives by clicking on the column header. Then, they judged the credibility of the outcome information for their chosen alternative by answering the same four perceived-credibility questions as in Experiment 1. Finally, they answered the following question (using a seven-point scale): How confident are you in your choice?

Procedure. After receiving brief training on the decision-making task, the participant completed the demographics questionnaire (including pet-experience questions) and the decision problems in about 30 min.

Results and Discussion

Repeated measures ANOVAs were used to analyze the choice, confidence, and credibility data. η^2_G was used for effect size.

Effects of credibility manipulations on choices. We hypothesized that participants would (a) choose the high-credibility alternative less often as its utility decreased, thus requiring participants to pay a higher utility premium for the choice, and (2) choose the high-credibility alternative more often as its credibility differed more from the two low-credibility alternatives, thus allowing a greater credibility gain. However, the decline in choosing the high-credibility alternative as utility premium increased was expected to be steeper for a small than a large credibility gain. The data in Figure 8 are relevant to these hypotheses. We intended to test the hypothesis with utility premium (three levels) and credibility gain (two levels) as independent variables and whether the participant chose the high credibility alternative as the dependent variable. However, participants always chose the high-credibility alternative when it had the best utility, so two conditions of this design had no variability and this model could not be run. Instead, we ran a 2 × 2 model using the levels where utility premium was moderate or high. As hypothesized, the high-credibility alternative was chosen less frequently as its utility decreased, $F(1, 54) = 21.3$, $p < .001$, $\eta^2_G = .08$, and more frequently with a large credibility gain than a small one, $F(1, 54) = 40.4$, $p < .001$, $\eta^2_G = .13$. F8

The interaction was negligible in size and not significant, $F(1, 54) = 1.95$, $p = .17$, $\eta^2_G = .004$. However, when all six conditions of the 3 × 2 design are considered, the data in Figure 8 seem to

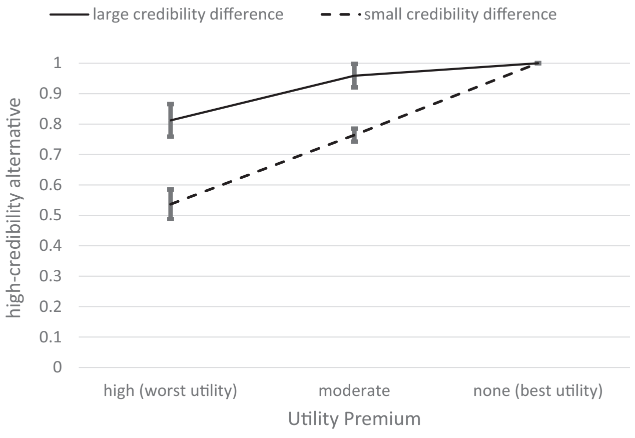


Figure 8. Proportion of participants choosing the high-credibility alternative when it had the worst, moderate, and best utility (i.e., when utility premium was high, moderate or none, respectively) and when the difference between high and low credibiltiy alternatives (credibility gain) was large and small. Error bars represent 1 standard error.

support the hypothesized interaction between utility premium and credibility gain. When the credibility gain was large, the frequency of choosing the high-credibility alternative decreased only slightly from its maximum as utility premium increases; but when the credibility gain was small, the frequency of choosing the high-credibility alternative decreased markedly. Statistical support for this interaction comes from the finding that the proportion of choosing the high-credibility alternative was significantly greater with a large than a small credibility gain when the utility premium was moderate, $F(1, 54) = 26.2$, $p < .001$, $\eta^2_G = .15$, but identical for both credibility-gain levels when the utility premium was 0 (best utility).

Thus, the utility premium and credibility gain hypotheses were supported with small to medium effect sizes, although support for the interaction hypothesis was weaker. On the decisions where participants could show ambiguity aversion (because the high-credibility alternative had suboptimal utility), they avoided the best-utility alternative in favor of the high-credibility alternative 80.7% of the time. Although participants' overall frequency of ambiguity aversion was high, these findings identified two factors that influenced this frequency. The findings supporting the utility-premium hypothesis show that participants considered how much utility they had to sacrifice to choose the higher-credibility alternative. The evidence for the credibility-gain hypothesis shows that participants considered how much credibility they would gain by choosing a higher-credibility alternative. Thus, these findings show that when participants demonstrated ambiguity aversion, they traded off utility for credibility in a fine-grained, quantitative fashion.

Effects of credibility cues on confidence in choices. In addition to actual choices, researchers sometimes use participants' confidence in choices as an alternative measure of their decision preferences (Sieck & Yates, 1997). As shown in Figure 9, confidence decreased strongly as the utility premium increased, $F(2, 108) = 59.2$, $p < .001$, $\eta^2_G = .23$, and was slightly higher with a large than a small credibility gain, $F(2, 108) = 14.2$, $p < .001$, $\eta^2_G = .02$. The rate of decrease in confidence as utility premium increased was not much different with a small than a large credibility gain, $F(2, 108) = 7.5$, $p < .01$, $\eta^2_G = .01$. Based on these effect sizes, the confidence data aligned closely with the choice data for the utility premium main effect. However, the effect sizes for the credibility-gain main effect and the interaction were quite small.

Mediation by perceived credibility. Here we consider whether perceived credibility mediated the effects of the credibility and utility manipulations on the probability of choosing the high credibility alternative (called "high-credibility choice" in the following), as shown in Figure 7B. After making each choice, participants judged the credibility of the alternative they chose using the same questions as in Experiment 1 (trustworthiness, continue, believability, and accuracy). We created a composite, perceived-credibility variable by reversing the agreement judgments for the continue question and then averaging the judgments for the questions, so that 1 represented lowest and 7 highest credibility. The data in Figure 9 (which represents causal link *a* in Figure 7B) show that utility premium and credibility gain affected perceived credibility similarly to how these manipulations affected high-credibility choice (see Figure 8).

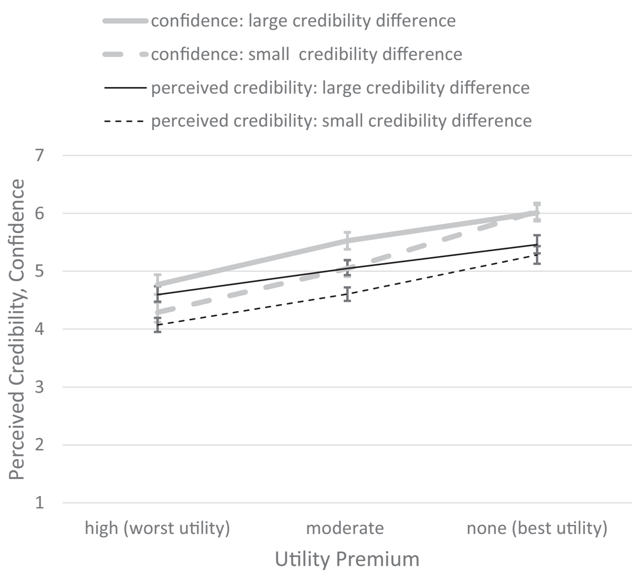


Figure 9. Confidence in choice (gray lines) and composite perceived credibility (black lines) when the high-credibiltiy alternative had the worst, moderate, and best utility (i.e., when utility premium was high, moderate or none, respectively) and when the difference between high and low credibiltiy alternatives (credibility gain) was large and small. Error bars represent 1 standard error.

To conduct the mediation analyses, we needed an analysis method where each participant's responses on a single decision constituted one case, because perceived-credibility varied on a problem by problem basis within participants. Therefore, we used multilevel modeling (Snijders & Bosker, 2012). As in our previous analysis of the effects of the manipulations on choices, we used a 2 (utility premium) × 2 (credibility gain) design. We could not use the two cells of the full 3 × 2 design where utility premium was lowest, because all participants chose the high credibility alternative in these cells. Because high-credibility choice was dichotomous, we calculated effect sizes using the R^2_{dicho} measure in Snijders and Bosker (2012).

In a model where only the utility premium and credibility gain manipulations predicted high-credibility choice (i.e., model *c* in Figure 7A), these manipulations accounted for 38% of the variance in high-credibility choice. The diagram in Figure 7B illustrates how to test whether perceived credibility mediated the effects of the manipulations. Link c' represents the *direct* influence of the manipulations on high-credibility choice, whereas links *a* and *b* represent their *indirect* effect. A model where the manipulations *and* perceived credibility predicted high-credibility choice (model b & c' in Figure 7B) accounted for 57% of the variance in high-credibility choice. Furthermore, the *unique* or *direct* contribution of the credibility manipulations (link *c'* only) accounted for only 13% of the variance in high-credibility choice. This marked decrease in the size of the direct effect of the credibility manipulations when perceived credibility is added to the model (from 38% in model *c* to 13% for link c' in model b & c') suggests that perceived credibility mediated the effect of the manipulations.

To quantify the mediation effects more precisely, we used the ratio of the *indirect*, mediated effect of the credibility manipulations on high-credibility choice (slope $a \times$ slope b, Figure 7B) to the *total* effect in a model without mediators (slope c, Figure 7A; Preacher & Kelley, 2011), where slope refers to the unstandardized regression coefficient for a single model term. There were three terms in each model: utility-premium effect, credibility-gain effect, interaction. For the utility-premium manipulation, the indirect AQ: 17 effect was 37% of the total effect and a Sobel (1982) test showed significant mediation, $s = 4.58$, $p < .001$. For credibility gain, the mediation ratio was 54% and $s = 4.46$, $p < .001$. Perceived credibility did not significantly mediate the interaction effect, but this effect was small in the unmediated model.

General Discussion

Effects on Perceived Credibility

Our first research question considered how heuristic credibility cues affected participants' perceptions of the credibility of claims and their sources. Our initial analysis of the Experiment 1 data used a composite perceived-credibility variable averaged over the four credibility questions. The findings from two separate ANOVAs—with composite credibility averaged within comment and within author cues—supported the hypotheses that perceived credibility would increase with expertise and reputation valence and that credibility would increase with amount of information when valence was supportive and decrease with more information when valence was disconfirmatory. These findings supported the predictions of the Bayesian belief-updating model and extended Corner and Hahn's (2009) version of this model to encompass disconfirmatory evidence.

Unexpectedly, the valence effect and the valence by amount of information interaction were noticeably larger for comment than author cues. Further analyses, which looked at the effects of reputation cues on each credibility variable separately, suggested a post hoc explanation of this unexpected finding. Perhaps participants saw reputation cues as more relevant to—and diagnostic of—a credibility judgment when the referent of the cue (comment or author) matched the target of the credibility judgment. Based on this post hoc interpretation, analysis using a new composite credibility variable—averaged over the credibility questions within diagnosticity levels instead of comment versus author cues—again supported our hypotheses regarding effects of reputation cues. This analysis also provided tentative support for two unhypothesized interactions involving diagnosticity that were predicted by the Bayesian updating model, including the full diagnosticity by valence by amount of information interaction.

Credibility–Utility Conflict

Our second research question dealt with how people handle conflict between credibility and utility during decision making. Experiment 2 showed that participants often (on 81% of decisions) dealt with this conflict by choosing a credible but lower-utility alternative over a best-utility alternative with low credibility, thus demonstrating ambiguity or vagueness aversion. The data on choices and confidence in choices suggested that participants chose the high credibility alternative less frequently as this alternative required sacrificing more utility (utility premium effect). They also sacrificed utility to gain credibility more frequently when doing so gained them a large increase in credibility than when it gained little credibility (credibility gain effect). These findings supported our hypotheses, which were based on Budescu et al.'s (2002) decision making model. They also suggest that when participants demonstrated ambiguity aversion, they traded off utility for credibility in a fine-grained, quantitative fashion.

Perceived Credibility Mediating Decisions

Our third research question focused on whether perceived credibility mediated the effects of credibility manipulations on decisions. The mediation analysis in Experiment 2 showed that one third to one half of the effect of utility premium and credibility gain on high-credibility choice was mediated by participants' self-reported perceptions of credibility. Because the mediation analysis used participants' explicit credibility judgments, these findings support the conclusion that people's explicit judgments of the credibility of outcomes influence their choices. Thus, in the current studies, participants' subjective credibility judgments influenced both their beliefs (Experiment 1) and their choices (Experiment 2).

Contribution

Novel empirical findings. To our knowledge, our finding of a valence by amount of information interaction in Experiment 1 has not been reported in prior credibility research, as prior research has tended to focus on supportive evidence (e.g., Corner & Hahn, 2009). Other findings that may be novel include the evidence suggesting—in a post hoc manner—that the main effect of valence on credibility judgments was stronger when diagnosticity (as cued by relevance) was high than low, as well as the evidence supporting the three-way interaction of diagnosticity, reputation valence and amount of reputation information.

Studying both argumentation and decision making. Research on persuasion and Internet credibility has tended to focus on argumentation (belief-updating) tasks and use credibility cues such as source credibility, reputation and amount of corroborating information (e.g., Corner & Hahn, 2009; Flanagin & Metzger, 2013). Research on ambiguity (vagueness) aversion tends to focus on decision-making tasks and use credibility cues such as the precision of evidence (e.g., Budescu et al., 2002; Koch & Schunk, 2013). By studying credibility using both an argumentation and a decision-making task, we were able to make the theoretical point that the different credibility cues used in these research communities may influence a common factor in the Bayesian updating model, weight of evidence (Einhorn & Hogarth, 1985). In addition, we showed that credibility manipulations generally used in argumentation but not decision-making research—reputation and expertise cues within online reputation systems—influenced behavior in a decision-making task. Finally, the mediation findings in Experiment 2 suggest that decision making models need to incorporate representations of people's subjective credibility judgments. In Budescu et al.'s (2002) model, credibility was represented by the vagueness weights for utilities and probabilities.

An elaborated belief-updating model. One advantage of the current research is that it demonstrates how psychological models expressed mathematically—a Bayesian belief-updating model of and an elaboration of Prospect Theory—can help explain how people use credibility information to evaluate degree of belief in a claim and make decisions. Although we did not develop new mathematical models in this project, we have pointed out important relationships between two belief-updating models that deserve to be highlighted. In presenting their model based on Bayes's theorem (Equation 1), Hahn and colleagues (Corner & Hahn, 2009; Hahn & Oaksford, 2007) did not decompose the overall Bayesian likelihood ratio (based on all evidence) into strength, valence, and weight of evidence, as Griffin and Tversky (1992) did (Equation 2). This decomposition is important for modeling credibility because the weight of evidence includes important credibility cues like the diagnosticity and amount of evidence. Also, as demonstrated in Experiment 1, Griffin and Tversky's model allows predicting how the credibility cues comprising the weight of evidence interact with other information that influences degree of belief, for example, valence. In summary, we were able to describe how many of the heuristic cues investigated in empirical research on credibility—including source credibility, precision, relevance and amount of information—mapped onto a well-accepted framework for modeling how people use evidence to evaluate claims.

Bayesian models are commonly used in cognitive psychology to explain how people reason with evidence in a variety of ways, including judging the strength of causal relationships (Lu et al., 2008), diagnostic reasoning (Waldmann, Cheng, Hagmayer, & Blaisdell, 2008), and gathering evidence to test hypotheses (Oaksford & Chater, 2003). To our knowledge, relatively few researchers—exceptions include Corner and Hahn (2009) and Lagnado et al. (2012)—have empirically tested how Bayesian models can be applied in investigating people's credibility judgments. Using the Bayesian model allowed us to predict and identify the effects of credibility cues—including the valence, relevance, and diagnosticity of evidence—that are not often considered in credibility research.

Potential Limitations

Bayesian belief-updating model. The diagnosticity by valence and the diagnosticity by valence by amount of information interaction in Experiment 1 were based on the post hoc interpretation that diagnosticity depended on the relevance of reputation cues (comment vs. author) to specific credibility questions. In addition, because the experiment was not designed to test these interactions, we could not conduct significance tests and relied only on effect size to document them. For both of these reasons, this finding must be considered tentative and should be replicated in a prospective study.

A related issue is that, although we used the updating model to explain these interactions involving diagnosticity as cued by relevance, we used the same model to predict only main effects of another diagnosticity cue, expertise. This seems inconsistent. Our explanation for this was that valence, amount of information, and relevance were all instantiated in the reputation cues (star ratings), whereas expertise was a separate cue. This explanation suggests a principled reason for predicting when credibility cues will interact, namely, that the updating model makes separate predictions for separate credibility cues. Under this interpretation, the reputation cues varied in terms of diagnosticity, valence and amount of information, while expertise varied only in terms of diagnosticity. Similar to our study, Corner and Hahn (2009) found an interaction between diagnosticity (cued by reputation) and amount of information because a *single* credibility cue, an article, varied in terms of reputation (science vs. nonscience publication) and amount of information (number of studies in the article).

Another concern is that, in the mathematical and conceptual versions of the updating model, two principle credibility-related constructs are degree of belief in claims and the weight of evidence, which includes source credibility. In measuring perceived credibility, we mapped questions about *accuracy*, *trustworthiness*, and *intention to act* onto the construct of degree of belief in a claim (comment) and mapped a question about *believability* onto the construct of source (author) credibility. However, other assumptions seem plausible. A reviewer suggested that believability maps more directly onto the construct of belief in claims. Also, because accuracy and trustworthiness are aspects of source credibility in the conceptual model, they seem to represent characteristics of sources. In response to these arguments, we suggest that querying participants about the accuracy of a claim or the believability of a source—as we did—fits the meaning of these terms in natural language. Peoples' mental models of the concept of credibility are probably much less specific than researchers' models. Thus, the mappings we used may have "made sense" to participants. In Experiment 1, the valence by diagnosticity interaction predicted by the Bayesian model accounted for 35% to 51% of the variance in accuracy, trustworthiness and believability judgments. Because diagnosticity was defined in this instance by whether the credibility cue (comment or author) matched the target of the question, these very large effect sizes suggest that participants saw our mappings of questions onto claims versus sources as reasonable. However, further research on this issue is needed.

Heuristic versus semantic cues. In both studies, we focused on the influence of *heuristic* cues by having participants evaluate the credibility of messages on an unfamiliar topic, thus ignoring the influence of *semantic*, content cues. One rationale for this focus is that people often lack one of the three conditions considered to be necessary for evaluating credibility based on the content of messages—time, motivation, and expertise (Petty & Cacioppo, 1986). In addition, Sloman and Fernbach (2017) argue that most people's general knowledge is quite shallow outside of their particular areas of expertise. These considerations suggest that people need to rely on heuristic cues regularly and for many topics. This may be why education aimed at improving credibility judgments often focuses on heuristic cues. A good example of this is a pamphlet about checking the credibility of news articles (News Literacy Project, 2018), where most of the credibility cues listed are heuristic cues, including some cues studied here: corroborating information across multiple publishers and checking source expertise and reputation.

Social biases. Because participants in Experiment 1 saw information displays with only heuristic credibility cues varying, they may have felt social pressure to give a certain credibility judgment when a certain credibility cue was presented because it accorded with perceived social norms or was expected by the experimenter, that is, demand characteristics. However, to say these findings were attributable to social demands seems not much different from saying that participants' understanding of the meaning of the credibility cues agreed with the consensual meaning of these cues. Also, some of our findings involved complex interactions among valence, amount of information and diagnosticity. To say that these interactions were the result of demand characteristics seems implausible. Finally, the key problem that participants faced in Experiment 1 was arriving at a *single* credibility judgment on each trial by integrating the information from multiple credibility cues. Often, there was conflict among the cues (some indicated high credibility, some low). It is not clear to us how social demands could have guided participants in integrating conflicting cues.

Order effects. Experiment 2 used the same, randomized order of decisions for each participant. Thus, the independent variables could have been confounded with response fatigue.

Applications

Weighting online credibility cues. Eysenbach (2008) pointed out an important difference between traditional and online media that may help in understanding some of our findings. Online forums with community-generated content and a reputation system dispense with the gatekeepers (e.g., editors, expert moderators) who improve credibility in traditional media and some informational websites. Instead, these forums rely on content and feedback generated by peers whose expertise is more varied. One problem with peer ratings as a credibility cue is that, even though peers often have less expertise than traditional gatekeepers, consumers may not down-weight the evidence from peers and in some cases may consider online peers as *more* credible than experts (Glenton, Nilsen, & Carlsen, 2006). For example, Xu (2013) found that when undergraduates read articles on a news aggregation site with a reputation mechanism, their credibility ratings for articles were more strongly influenced by the articles' reputation (number of peer endorsements) than by the credibility and expertise of the article's source.

In Experiment 1, the expertise cue accounted for 8% to 21% of the variance in the four credibility judgments (see right side of Table 1). Peer-reputation cues included reputation valence, the number of peer ratings and diagnosticity as cued by relevance. The valence by amount of information interaction accounted for 7% to 18% of the variance in the credibility judgments. The valence by diagnosticity interaction showed a large effect size of 29%. Thus, our participants gave peer-reputation cues equal or greater weight than source expertise cues, even though no information about the expertise of the peer raters was available and the highest credential was veterinarian. This suggests overreliance on online peer reputation cues.

Heuristic credibility cues online. We have argued that people regularly rely on nonsemantic, heuristic cues, which were investigated in the current studies, to evaluate credibility. As people's access to information and *misinformation* continues to increase—via social media and informational sites on the Internet and the proliferation of TV and radio channels—it seems that they will continue to need to use heuristic cues to evaluate credibility. Understanding how people use and misuse heuristic credibility cues may help in educating people to be better information consumers in an increasingly complex media environment.

Relevance of evidence. In our post hoc explanation of some of the Experiment 1 results, we assumed that the relevance of information is a cue to its diagnosticity and ultimately to degree of belief in claims. We noted earlier that Pearl (1988) used a Bayesian belief model to suggest that evidence is relevant to a claim if adding the evidence to prior knowledge changes the likelihood of the claim. Interestingly, this Bayesian definition of relevance has already been applied in U.S. law, which defines relevant evidence in trials very similarly to Pearl (Federal Rules of Evidence, 2011).

Credibility in decision making. Experiment 2 showed that when credibility and utility information were salient and displayed in a well-organized (matrix) fashion, participants' choices responded appropriately to tradeoffs between credibility and utility. This suggests that decision making may be improved by making decision information, including credibility, more salient and better organized. Decision information is sometimes presented to decision makers in this manner. For example, Mullan et al. (2009) presented information for diabetes treatment decisions to patients in a matrix format showing treatments and evaluative attributes. They also used ranges (e.g., A1C decreases 1% to 2%) instead of point estimates, thus indicating the credibility of decision outcomes in terms of precision. The cone of uncertainty around a hurricane's path is a salient indicator of the credibility of predicted locations (National Hurricane Center, 2019).

In line with our findings, there is evidence that people sometimes use credibility information during everyday decision making. People who perceive information about cancer to be less credible are less likely to engage in preventive health behaviors regarding cancer (Han, Moser, & Klein, 2007). Policymakers sometimes consider the credibility of information during policy decisions (Dieckmann, Robert Mauro, & Slovic, 2010). On the other hand, more research is needed on credibility use in realistic decision contexts, as decision makers sometimes use credibility information inappropriately. In a study by Woloshin et al. (2000), most of the women thought that information credibility was not a problem when making mammography screening decisions and were unaware that mammography information may lack credibility. Weather forecasters often overestimate the credibility of storm-path predictions despite the salient cone-of-uncertainty cue (Herdener, Wickens, Clegg, & Smith, 2016). In Eysenbach and Köhler (2002), many participants reported that they relied on reputational cues to credibility while searching for health information but overlooked these cues when actually searching.

Conclusion

People often use heuristic credibility cues to judge the credibility of information (e.g., news, advice) because they lack domain expertise. Experiment 1 showed that, when judging the credibility of advice on an online health forum with a peer reputation system, participants used heuristic cues like the amount and relevance of information as predicted by a Bayesian belief-updating model. However, participants may have overweighted advice from peers who lacked health expertise. Experiment 2 showed that participants' choices were guided by external credibility cues and utility information in accordance with a decision-making model that was elaborated to include credibility information.

References

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review, 98,* 409–429. http://dx.doi.org/10.1037/0033-295X.98.3.409

Ayeh, J. (2015). Travellers' acceptance of consumer-generated media: An integrated model of technology acceptance and source credibility theories. *Computers in Human Behavior, 48,* 173–180. http://dx.doi.org/10.1016/j.chb.2014.12.049

Bakeman, R. (2005). Recommended effect size statistics for repeated measures designs. *Behavior Research Methods, 37,* 379–384. http://dx.doi.org/10.3758/BF03192707

Barcikowski, R. (1981). Statistical power with group mean as the unit of analysis. *Journal of Educational Statistics, 6,* 267–285. http://dx.doi.org/10.3102/10769986006003267

Budescu, D., Kuhn, K., Kramer, K., & Johnson, T. (2002). Modeling certainty equivalents for imprecise gambles. *Organizational Behavior and Human Decision Processes, 88,* 748–768. http://dx.doi.org/10.1016/S0749-5978(02)00014-6

Chaiken, S. (1980). Heuristic versus systematic information processing and the use of source versus message cues in persuasion. *Journal of Personality and Social Psychology, 39,* 752–766. http://dx.doi.org/10.1037/0022-3514.39.5.752

Chaiken, S., & Maheswaran, D. (1994). Heuristic processing can bias systematic processing: Effects of source credibility, argument ambiguity, and task importance on attitude judgment. *Journal of Personality and Social Psychology, 66,* 460–473. http://dx.doi.org/10.1037/0022-3514.66.3.460

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Mahwah, NJ: Erlbaum.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112,* 155–159. http://dx.doi.org/10.1037/0033-2909.112.1.155

Corner, A., & Hahn, U. (2009). Evaluating science arguments: Evidence, uncertainty, and argument strength. *Journal of Experimental Psychology: Applied, 15,* 199–212. http://dx.doi.org/10.1037/a0016533

Crisci, R., & Kassanove, H. (1973). Effect of perceived expertise, strength of advice, and environmental setting on parental compliance. *The Journal of Social Psychology, 89,* 245–250. http://dx.doi.org/10.1080/00224545.1973.9922597

Curley, S., & Yates, J. F. (1985). The center and range of the probability interval as factors affecting ambiguity preferences. *Organizational Behavior and Human Decision Processes, 36,* 273–287. http://dx.doi.org/10.1016/0749-5978(85)90016-0

Curley, S., Yates, J. F., & Abrams, R. (1986). Psychological sources of ambiguity avoidance. *Organizational Behavior and Human Decision Processes, 38,* 230–256. http://dx.doi.org/10.1016/0749-5978(86)90018-X

Dellarocas, C. (2003). The digitization of word-of-mouth: Promise and challenges of online reputation systems. *Management Science, 49,* 1407–1424. http://dx.doi.org/10.1287/mnsc.49.10.1407.17308

Dieckmann, N., Robert Mauro, R., & Slovic, P. (2010). The effects of presenting imprecise probabilities in intelligence forecasts. *Risk Analysis, 30,* 987-. http://dx.doi.org/10.1111/j.1539-6924.2010.01384.x

Eastin, M. S. (2001). Credibility assessments of online health information: The effects of source expertise and knowledge of content. *Journal of Computer-Mediated Communication, 6,* JCMC643. http://dx.doi.org/10.1111/j.1083-6101.2001.tb00126.x

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review, 70,* 193–242. http://dx.doi.org/10.1037/h0044139

Einhorn, H. J., & Hogarth, R. M. (1985). Ambiguity and uncertainty in probabilistic inference. *Psychological Review, 92,* 433–461. http://dx.doi.org/10.1037/0033-295X.92.4.433

Ellsberg, D. (1961). Risk, ambiguity and the Savage axioms. *The Quarterly Journal of Economics, 75,* 643–669. http://dx.doi.org/10.2307/1884324

Eysenbach, G. (2008). Credibility of health information and digital media: New perspectives and implications for youth. In M. Metzger & A. Flanagin (Eds.), *Digital media, youth, and credibility: The John D. and Catherine T. MacArthur Foundation series on digital media and learning* (pp. 123–154). Cambridge, MA: The MIT Press. AQ: 25

Eysenbach, G., & Köhler, C. (2002). How do consumers search for and appraise health information on the world wide web? Qualitative study using focus groups, usability tests, and in-depth interviews. *British Medical Journal, 324,* 573–577. http://dx.doi.org/10.1136/bmj.324.7337.573

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,* 175–191. http://dx.doi.org/10.3758/BF03193146

Federal Rules of Evidence. (2011). Rule 401. Test for relevant evidence. Pub L. 93–595, §1, Jan. 2, 1975, 88 Stat. 1931; Apr. 26, 2011, eff Dec, 1, 2011.

Flanagin, A., & Metzger, M. (2013). Trusting expert- versus user-generated ratings online: The role of information volume, valence, and consumer characteristics. *Computers in Human Behavior, 29,* 1626–1634. http://dx.doi.org/10.1016/j.chb.2013.02.001

Fox, S., & Duggan, M. (2013). *Health online 2013*. Washington, DC: Pew Research Center's Internet & American Life Project.

Frisch, D., & Baron, J. (1988). Ambiguity and rationality. *Journal of Behavioral Decision Making, 1,* 149–157. http://dx.doi.org/10.1002/bdm.3960010303

Glenton, C., Nilsen, E. S., & Carlsen, B. (2006). Lay perceptions of evidence-based information—A qualitative evaluation of a website for back pain sufferers. *BMC Health Services Research, 6,* 34. http://dx.doi.org/10.1186/1472-6963-6-34

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

Griffiths, T. L., & Tenenbaum, J. B. (2009). Theory-based causal induction. *Psychological Review, 116,* 661–716. http://dx.doi.org/10.1037/a0017201

Hahn, U., & Oaksford, M. (2007). The rationality of informal argumentation: A Bayesian approach to reasoning fallacies. *Psychological Review, 114,* 704–732. http://dx.doi.org/10.1037/0033-295X.114.3.704

Han, P. K., Moser, R. P., & Klein, W. M. (2007). Perceived ambiguity about cancer prevention recommendations: Associations with cancer-related perceptions and behaviours in a U.S. population survey. *Health Expectations, 10,* 321–336. http://dx.doi.org/10.1111/j.1369-7625.2007.00456.x

Heath, C., & Tversky, A. (1991). Preference and belief: Ambiguity and competence in choice under uncertainty. *Journal of Risk and Uncertainty, 4,* 5–28. http://dx.doi.org/10.1007/BF00057884

Herdener, N., Wickens, C. D., Clegg, B. A., & Smith, C. A. P. (2016). Uncertain understanding of uncertainty in spatial prediction. *Human Factors, 58,* 899–914. http://dx.doi.org/10.1177/0018720816645259

Hilligoss, B., & Rieh, S. Y. (2008). Developing a unifying framework of credibility assessment: Construct, heuristics, and interaction in context. *Information Processing & Management, 44,* 1467–1484. http://dx.doi.org/10.1016/j.ipm.2007.10.001

Hogarth, R., & Kunreuther, H. (1989). Risk, ambiguity and insurance. *Journal of Risk and Uncertainty, 2,* 5–35. http://dx.doi.org/10.1007/BF00055709

Hovland, C. I., Janis, I. L., & Kelley, H. H. (1953). *Communication and persuasion: Psychological studies of opinion change*. New Haven, CT: Yale University Press.

Hovland, C. I., & Weiss, W. (1951). The influence of source credibility on communication effectiveness. *Public Opinion Quarterly, 15,* 635–650. http://dx.doi.org/10.1086/266350

Johnson, T., & Kaye, B. (2015). Reasons to believe: Influence of credibility on motivations to use social networks. *Computers in Human Behavior, 50,* 544–555. http://dx.doi.org/10.1016/j.chb.2015.04.002

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47,* 263–292. http://dx.doi.org/10.2307/1914185

Kelman, H. C., & Hovland, C. I. (1953). "Reinstatement" of the communicator in delayed measurement of opinion change. *The Journal of Abnormal and Social Psychology, 48,* 327–335. http://dx.doi.org/10.1037/h0061861

Keynes, J. M. (1921). *A treatise on probability*. London, U.K.: MacMillan.

Koch, C., & Schunk, D. (2013). Limiting liability? – Risk and ambiguity attitudes under real losses. *Schmalenbach Business Review, 65,* 54–75. http://dx.doi.org/10.1007/BF03396850

Krull, J. L., & MacKinnon, D. P. (2001). Multilevel modeling of individual and group level mediated effects. *Multivariate Behavioral Research, 36,* 249–277. http://dx.doi.org/10.1207/S15327906MBR3602_06

Kruschke, J., & Vanpaemel, W. (2015). Bayesian estimation in hierarchical models. In J. R. Busemeyer, Z. Wang, J. T. Townsend, & A. Eidels (Eds.), *The Oxford handbook of computational and mathematical psychology* (pp. 279–299). New York, NY: Oxford University Press. AQ: 20

Kunreuther, H., Hogarth, R., & Meszaros, J. (1993). Insurer ambiguity and market failure. *Journal of Risk and Uncertainty, 7,* 71–87. http://dx.doi.org/10.1007/BF01065315

Lagnado, D. A., Fenton, N., & Neil, M. (2012). Legal idioms: A framework for evidential reasoning. *Argument & Computation, 4,* 46–63. http://dx.doi.org/10.1080/19462166.2012.682656

Lo, A., & Yao, S. (2019). What makes hotel online reviews credible?: An investigation of the roles of reviewer expertise, review rating consistency and review valence. *International Journal of Contemporary Hospitality Management, 31,* 41–60. http://dx.doi.org/10.1108/IJCHM-10-2017-0671

Lu, H., Yuille, A. L., Liljeholm, M., Cheng, P. W., & Holyoak, K. J. (2008). Bayesian generic priors for causal learning. *Psychological Review, 115,* 955–984. http://dx.doi.org/10.1037/a0013256

Maffioletti, A., & Santori, M. (2005). Do trade union leaders violate subjective expected utility? Some insights from experimental data. *Theory and Decision, 59,* 207–253. http://dx.doi.org/10.1007/s11238-005-8633-3

Massey, C., & Wu, G. (2005). Detecting regime shifts: The causes of under- and overreaction. *Management Science, 51,* 932–947. http://dx.doi.org/10.1287/mnsc.1050.0386

Meder, B., & Mayrhofer, R. (2017). Diagnostic reasoning. In M. Waldmann (Ed.), *Oxford handbook of causal reasoning* (pp. 1–43). Oxford, UK: Oxford University Press. http://dx.doi.org/10.1093/oxfordhb/9780199399550.013.23

Metzger, M. J., Flanagin, A. J., & Medders, R. B. (2010). Social and heuristic approaches to credibility evaluation online. *Journal of Communication, 60,* 413–439. http://dx.doi.org/10.1111/j.1460-2466.2010.01488.x

Moreno, A., & Terwiesch, C. (2014). Doing business with strangers: Reputations in online service marketplaces. *Information Systems Research, 25,* 865–886. http://dx.doi.org/10.1287/isre.2014.0549

Morey, R. (2016, January 7). Averaging can produce misleading standardized effect sizes [web log post]. Retrieved March 16, 2019, from http://bayesfactor.blogspot.com/2016/01/averaging-can-produce-misleading.html

Mullan, R. J., Montori, V. M., Shah, N. D., Christianson, T. J., Bryant, S. C., Guyatt, G. H., . . . Smith, S. A. (2009). The diabetes mellitus medication choice decision aid: A randomized trial. *Archives of Internal Medicine, 169,* 1560–1568. http://dx.doi.org/10.1001/archinternmed.2009.293

National Hurricane Center. (2019). *Definition of the NHC track forecast cone*. Retrieved May 9, 2019, from https://www.nhc.noaa.gov/aboutcone.shtml

News Literacy Project. (2018). Ten questions for fake news detection. Retrieved from www.hcdsb.org/Students/Library/Documents/Checkology%2010%20questions%20for%20fake%20news%20detection.pdf

Nisbett, R., Krantz, D., Jepson, C., & Kunda, Z. (1983). The use of statistical heuristics in everyday inductive reasoning. *Psychological Review, 90,* 339–363. http://dx.doi.org/10.1037/0033-295X.90.4.339

Oaksford, M., & Chater, N. (2003). Optimal data selection: Revision, review, and reevaluation. *Psychonomic Bulletin & Review, 10,* 289–318. http://dx.doi.org/10.3758/BF03196492

Ohanian, R. (1990). Construction and validation of a scale to measure celebrity endorsers' perceived expertise, trustworthiness, and attractiveness. *Journal of Advertising, 19,* 39–52. http://dx.doi.org/10.1080/00913367.1990.10673191

O'Keefe, D. (1990). *Persuasion: Theory & research*. Thousand Oaks, CA: SAGE.

Olejnik, S., & Algina, J. (2003). Generalized eta and omega squared statistics: Measures of effect size for some common research designs. *Psychological Methods, 8,* 434–447. http://dx.doi.org/10.1037/1082-989X.8.4.434

Pearl, J. (1988). *Probabilistic inference in intelligent systems: Networks of plausible inference*. San Francisco, CA: Morgan Kaufman.

Petty, R. E., & Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. In *Communication and persuasion* (pp. 1–24). New York, NY: Springer.

Petty, R., Cacioppo, J., & Schumann, D. (1983). Central and peripheral routes to advertising effectiveness: The moderating role of involvement. *Journal of Consumer Research, 10,* 135–146. http://dx.doi.org/10.1086/208954

Preacher, K. J., & Kelley, K. (2011). Effect size measures for mediation models: Quantitative strategies for communicating indirect effects. *Psychological Methods, 16,* 93–115. http://dx.doi.org/10.1037/a0022658

Resnick, P., Zeckhauser, R., Swanson, J., & Lockwood, K. (2006). The value of reputation on eBay: A controlled experiment. *Experimental Economics, 9,* 79–101. http://dx.doi.org/10.1007/s10683-006-4309-2

Ritov, I., & Baron, J. (1990). Reluctance to vaccinate: Omission bias and ambiguity. *Journal of Behavioral Decision Making, 3,* 263–277. http://dx.doi.org/10.1002/bdm.3960030404

Schum, D. A. (1989). Knowledge, probability and credibility. *Journal of Behavioral Decision Making, 2,* 39–62. http://dx.doi.org/10.1002/bdm.3960020104

Sieck, W., & Yates, J. F. (1997). Exposition effects on decision making: Choice and confidence in choice. *Organizational Behavior and Human Decision Processes, 70,* 207–219. http://dx.doi.org/10.1006/obhd.1997.2706

Sloman, S., & Fernbach, P. (2017). *The knowledge illusion: Why we never think alone*. New York, NY: Penguin Random House.

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). London, UK: Sage Publishers.

Sobel, M. E. (1982). Asymptotic confidence intervals for indirect effects in structural equation models. *Sociological Methodology, 13,* 290–312. http://dx.doi.org/10.2307/270723

Speekenbrink, M., & Shanks, D. R. (2013). Decision making. In D. Reisberg (Ed.), *The Oxford handbook of cognitive psychology* (pp. 682–703). Oxford, UK: Oxford University Press.

Sussman, S., & Siegal, W. (2003). Informational influence in organizations: An integrated approach to knowledge adoption. *Information Systems Research, 14,* 47–65. http://dx.doi.org/10.1287/isre.14.1.47.14767

Sutton, A. J., & Abrams, K. R. (2001). Bayesian methods in meta-analysis and evidence synthesis. *Statistical Methods in Medical Research, 10,* 277–303. http://dx.doi.org/10.1177/096228020101000404

Tentori, K., Crupi, V., & Osherson, D. (2010). Second-order probability affects hypothesis confirmation. *Psychonomic Bulletin & Review, 17,* 129–134. http://dx.doi.org/10.3758/PBR.17.1.129

Tversky, A. (1977). Features of similarity. *Psychological Review, 84,* 327–352. http://dx.doi.org/10.1037/0033-295X.84.4.327

Waldmann, M. R., Cheng, P. W., Hagmayer, Y., & Blaisdell, A. P. (2008). Causal learning in rats and humans: A minimal rational model. In N. Chater & M. Oaksford (Eds.), *The probabilistic mind. Prospects for Bayesian cognitive science* (pp. 453–484). New York, NY: Oxford University Press. http://dx.doi.org/10.1093/acprof:oso/9780199216093.003.0020

Woloshin, S., Schwartz, L. M., Byram, S. J., Sox, H. C., Fischhoff, B., & Welch, H. G. (2000). Women's understanding of the mammography screening debate. *Archives of Internal Medicine, 160,* 1434–1440. http://dx.doi.org/10.1001/archinte.160.10.1434

Xu, C. (2013). Social recommendation, source credibility, and recency: Effects of news cues in a social bookmarking website. *Journalism & Mass Communication Quarterly, 90,* 757–775. http://dx.doi.org/10.1177/1077699013503158

Yaniv, I., & Kleinberger, E. (2000). Advice taking in decision making: Egocentric discounting and reputation formation. *Organizational Behavior and Human Decision Processes, 83,* 260–281. http://dx.doi.org/10.1006/obhd.2000.2909

Yi, Y. J., Stvilia, B., & Mon, L. (2012). Cultural influences on seeking quality health information: An exploratory study of the Korean community. *Library & Information Science Research, 34,* 45–51. http://dx.doi.org/10.1016/j.lisr.2011.06.001

Zhang, W., & Watts, S. (2008). Capitalizing on content: Information adoption in two online communities. *Journal of the American Society for Information Science, 9,* 73–94.

Appendix A

Decomposing the Likelihood Ratio Into Strength and Weight of Evidence

Equation 2 in the text is based on Griffin and Tversky's (1992, p. 415) equation, which described how the strength and weight of evidence contribute to the overall Bayesian likelihood ratio (based on a set of multiple pieces of evidence; $\mathrm{LR}_{\mathrm{ALL}}$) and, ultimately, to the posterior probability of a hypothesis given this evidence. Equation 2 is a special case of Bayes Theorem that applies to human judgments about situations like the following one (which was used in the text and is adapted from a problem used by Corner & Hahn, 2009).

An article published in the journal Science [*or* the website excitingnews.com] described a number of studies testing whether mineral supplement X causes headache as a side effect. Fifty studies found that supplement X did cause headache, whereas 30 studies found that it did not. Based on this information, how likely is it that supplement X causes headache?

This example includes a number of constraints that must be met for Equation 2 to apply. First, there are two mutually exclusive hypotheses; the supplement causes headache, *H*, or not, ¬*H*. Second, the evidence consists of a set, *E*, of independent pieces of evidence, with each piece of evidence involving two mutually exclusive outcomes. In the example, *E* consists of *N* studies, each of which has the outcome *yes* (the supplement caused headache) or *no* (the supplement did not cause headache). Third, the diagnosticity of each piece of evidence is the same. Fourth, the likelihoods of the two outcomes are symmetric, i.e., $P(yes|H) = P(no|\neg H)$ and $P(yes|\neg H) = P(no|H)$. Fifth, it is assumed that the reasoner has little information about the prior probabilities of the two alternative hypotheses (uninformative priors), i.e., $P(H) = P(\neg H) = 0.5$, so the prior odds ratio is 1. The two publications in the problem represent a manipulation of diagnosticity using the cue of reputation for accuracy. It is assumed that a scientific journal is more accurate than the website at determining whether each study found evidence of headache or not.

Here we show how Griffin and Tversky's special-case version of Bayes Theorem is derived from Bayes Theorem (Equation 1 in the text). Given the fifth constraint, Bayes Theorem reduces to,

$$\frac{P(H|E)}{P(\neg H|E)} = \frac{P(E|H)}{P(E|\neg H)}, \tag{A1}$$

where the term on the right-hand side is the overall Bayesian likelihood ratio ($\mathrm{LR}_{\mathrm{ALL}}$). In the updating process, upon receiving each new piece of evidence (each study), Bayes Theorem is used to calculate the posterior odds, which then become the prior odds for the next update. Assume that $e_1, \ldots, e_N$ are the outcomes of the *N* studies in evidence set *E*, i.e., a series of *yes's* and *no's*. Based on the Bayesian updating process, Equation A1 becomes,

$$\frac{P(H|e_1, \ldots, e_N)}{P(\neg H|e_1, \ldots, e_N)} = \left(\frac{P(yes|H)}{P(yes|\neg H)}\right)^{N_{yes}}\left(\frac{P(no|H)}{P(no|\neg H)}\right)^{N_{no}}, \tag{A2}$$

where N_{yes} indicates the number of studies with on outcome of headache; N_{no} the number with an outcome of no headache; the left-hand likelihood ratio on the right-hand side is the diagnosticity of a *yes* outcome; and the right-hand likelihood ratio is the diagnosticity of a *no* outcome. Because the hypotheses are symmetric, the diagnosticity of a *no* outcome is the inverse of the diagnosticity of a *yes* outcome. Therefore,

$$\frac{P(H|E)}{P(\neg H|E)} = \left(\frac{P(yes|H)}{P(yes|\neg H)}\right)^{N_{yes}}\left(\frac{P(yes|\neg H)}{P(yes|H)}\right)^{N_{no}} \tag{A3}$$

$$\frac{P(H|E)}{P(\neg H|E)} = \left(\frac{P(yes|H)}{P(yes|\neg H)}\right)^{N_{yes}}\left(\frac{P(yes|H)}{P(yes|\neg H)}\right)^{-N_{no}} \tag{A4}$$

$$\frac{P(H|E)}{P(\neg H|E)} = \left(\frac{P(yes|H)}{P(yes|\neg H)}\right)^{N_{yes}-N_{no}} \tag{A5}$$

$$\frac{P(H|E)}{P(\neg H|E)} = \left(\frac{P(yes|H)}{P(yes|\neg H)}\right)^{\left(\frac{N_{yes}}{N} - \frac{N_{no}}{N}\right)N} \tag{A6}$$

Equation A6 is the same as Griffin and Tversky's equation (Equation 2), so the derivation is complete.

(*Appendices continue*)

Appendix B

Additional Analyses for Experiment 1

Effects of Valence and Amount of Information on Each Credibility Variable

Analyses of the Experiment 1 data using the composite credibility variable, which was an average of judgments on the four credibility questions, showed the hypothesized valence main effect and valence by amount of information interaction (Figure 4A). Unexpectedly, these two effects were much stronger when comment cues were manipulated than author cues. Following the analysis of the composite variable, we discussed analyses that investigated whether these effects found with the composite variable were shown by each of the individual variables. The graphs in Figure 5 in the text suggested that the effects found with the composite variable were also found with the trustworthiness, accuracy and continue variables. However, there was a different pattern for the believability variable. The valence and valence by amount of information effects were still found, but these effects were stronger for author cues than comment cues. Here we present the statistical support for these conclusions from the analyses of the individual credibility variables.

The analyses of the composite variable involved two ANOVAs—expertise by comment valence by comment amount of information, and expertise by author valence by author amount of information. To expand these analyses to the four individual credibility variables requires eight ANOVAs. Therefore, alpha was set to .00625. Table B1 shows the effect sizes and statistical significance decisions for each analysis as well as the difference in effect size between comment and author cues.

TB1
AQ: 22

Valence main effect. The table confirms that the size of the valence main effect was larger with comment than author cues for trustworthiness, accuracy and continue, but this effect size was much larger with author than comment cues for believability. Seven of the eight valence effects were significant; only the effect for the continue variable with author cues was not (see Table B2).

TB2

Valence × Amount of Information interaction. Table B1 also shows that the valence by amount of information interaction was larger with comment than author cues for trustworthiness, accuracy and continue. The interactions for these three variables were significant in the comment-cue analysis, where effect sizes were larger, but not for author cues. For the believability variable, both of the valence by amount of information interactions were significant, but their effect sizes were small and similar in size.

In summary, statistical analyses of the individual credibility variables support the pattern discussed above and in the text; i.e., with trustworthiness, accuracy and continue showing a similar pattern of effect size differences being larger with comment than author cues and (at least for the valence effect) believability showing the opposite pattern. The implications of this finding are discussed in the text.

Effect of Expertise on Each Credibility Variable

The expertise effect was tested in the same ANOVAs as the valence and amount of information effects. However, the expertise effect was identical in the comment and author ANOVAs, because the expertise means were identical in both datasets. Thus, to expand the expertise composite analysis to the four individual credibility variables required four ANOVAs. Therefore, alpha was set to .0125. Figure B1 shows the expertise effects for the individual variables. Table B2 shows that the expertise effect was ~~significant for all four variables~~.
only significant for trustworthiness and accuracy.

FB1

Table B1
Effect Sizes ($\Delta\eta^2_G$) and Significance Decisions for Individual Credibility Variables

	Valence main effect			Valence × Amount of Information interaction		
Variable	η^2_G for comment cues	η^2_G for author cues	$\Delta\eta^2_G$ (comment – author)	η^2_G for comment cues	η^2_G for author cues	$\Delta\eta^2_G$ (comment – author)
Trustworthiness	.674*	.309*	.365	.106*	.013	.093
Accuracy	.635*	.129*	.506	.087*	.011	.076
Continue	.274*	.024	.251	.039*	.001	.038
Believability	.250*	.597*	−.347	.042*	.037*	.005

Note. Statistical significance is indicated only for ANOVA effects, not for $\Delta\eta^2_G$ values.
* $p < .00625$.

(*Appendices continue*)

Table B2
Results of Significance Tests for Individual Credibility Variables

	Expertise *df* 2,38		Valence *df* 1, 19				Valence × Amount of Info *df* 1, 19			
	Both reputation-cue referents		Comment cues		Author cues		Comment cues		Author cues	
Variable	*F*	*p*	*F*	*p*	*F*	*p*	*F*	*p*	*F*	*p*
Trustworthiness	16.4	.001	77.7	.001	34.6	.001	60.0	.001	6.82	.017
Accuracy	9.15	.004	58.1	.001	34.2	.001	40.0	.001	8.70	.008
Continue	4.39	.036	17.8	.001	1.57	.230	14.1	.001	1.03	.320
Believability	7.05	.013	70.8	.001	81.4	.001	40.7	.001	32.9	.001

Note. Degrees of freedom for *F* tests are 1,19.

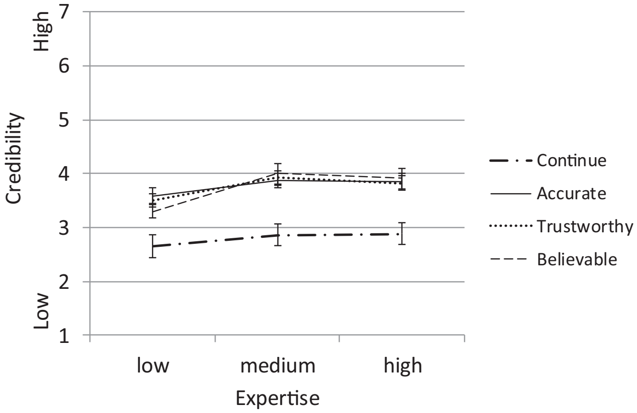


Figure B1. Effects of the comment author's domain expertise on four perceived-credibility dependent variables.

Appendix C

Budescu Model Calculations

Here we show using quantitative examples how Budescu et al.'s (2002) model leads to the utility-premium and credibility-gain hypotheses of Experiment 2. For clarity, we present the model formula (Equation 3 in text) here. It calculates the utility of one outcome for one alternative,

$$U_{cell} = V[w_x x_{worst} + (1 - w_x)x_{best}] \times f[w_p p_{worst} + (1 - w_p)p_{best}], \quad (C1)$$

where x = an outcome, p = a probability of an outcome, *best* = best case, *worst* = worst case, w_x and w_p = vagueness weighting factors for outcomes and probabilities, respectively, and $V[\bullet]$ and $f[\bullet]$ are the Prospect Theory value and decision-weight functions, respectively. For example, if the probability of side effects for a medication (a loss) ranges from 8 to 22%, p_{worst} is .22 and p_{best} is .08. If w_p is .5, the probability is represented by the center of the range. As w_p increases toward 1, the worst-case probability is more preferred, which leads to ambiguity or vagueness aversion. (The w_x parameter for weighting outcome vagueness works the same way.) Below, we describe how the Budescu et al. model handles three example decisions, which are similar to those presented in the text, where the decision was driven by the effectiveness attribute. Here, we focus on the side effects attribute and, because all outcomes are the same for this attribute (some side effects), on the model's probability term. These examples demonstrate that the Experiment 2 hypotheses follow from the model.

Utility Premium Hypothesis

The *utility-premium hypothesis* was that participants would choose the high-credibility alternative almost always when it has the best utility and less frequently as its utility decreased (and its utility premium increased correspondingly). Equation C1 was designed to model decisions where credible and vague probabilities can be specified numerically. Because probabilities were not specified numerically for the side effects attribute, we made assumptions about how our verbal problem information would be translated into numerical terms for these examples. Although making

(*Appendices continue*)

Table C1
Probabilities of the Three Outcomes for Side Effects Before and After Adjusting for Vague (Low-Credibility) Probabilities According to Budescu et al.'s (2002) Model

		Credibility gain					
		Small			Large		
Utility premium	Medication chance of side effects (utility)	Gent. none (best)	Cefra. few (med.)	Kana. mod. (worst)	Gent. none (best)	Cefra. few (med.)	Kana. mod. (worst)
Low (medium utility has high credibility)	Credibility level	med.	high	med.	low	high	low
	If all probabilities high credibility	.02	.07	.15	.02	.07	.15
	Adjustment for lower credibility	.07	.0	.07	.14	.0	.14
	Adjusted probabilities	.09	.07 choice	.22	.16	.07 choice	.29
High (worst utility has high credibility)	Credibility level	med.	med.	high	low	low	high
	If all probabilities high credibility	.02	.07	.15	.02	.07	.15
	Adjustment for low credibility	.07	.07	.0	.14	.14	.0
	Adjusted probabilities	.09 choice	.14	.15	.16	.21	.15 choice

Note. The three probabilities are shown for four decisions, based on crossing two factors, utility premium (whether the medium or worst utility has high credibility) by credibility gain. Choice refers to the alternative chosen by the model. none = virtually no; few = very few; mod. = moderate; med. = medium utility; Gent. = Gentamycin; Cefra. = Cefradoxil; Kana. = Kanamycin.

these assumptions is difficult, the point of this exercise is not to make precise quantitative predictions regarding the decision examples, but rather to show that the model supports our qualitative hypotheses when plausible assumptions are made. The model represents high-credibility probabilities by point probabilities and low-credibility probabilities as ranges centered on a midpoint. For all the examples presented here, we assumed that the decision maker first translates the verbal expressions of side effect frequency, *virtually no*, *very few* and *moderate*, into point probabilities of .02, .07 and .15, respectively; and second, uses the probability within a range that leads to the worst-case utility (i.e., w_p is 1) when the credibility of the side effect information is moderate or low. Because all outcomes in these decisions are losses, higher side effect probabilities lead to worse utilities than lower ones; therefore, the worst-case side-effect probabilities are greater than the midpoint of the probability range.

Example 1. Figure 6 shows a decision where the single high-credibility alternative has medium utility and the credibility gain was small, which meant that high credibility had a reputation rating of 5 stars and two lower-credibility options had a rating of 3 stars. This decision is shown in the upper-left corner of Table C1. The top line of the example shows the point probabilities that we assumed were inferred from the verbal descriptions of side effect frequency.

TC1

We assumed that the decision maker represents the moderately credible probabilities (3-star reputation) in Figure 6 by a range ±.07 units around the point probability. The second line of the example shows the worst case value within this range (+.07) that is applied to the two alternatives with lower credibility. The third line shows the adjusted probabilities after the model adds the imagined worst-case probabilities to the point probabilities inferred from the problem. That is, the model calculates $f[p_i]$ for *virtually no*, *very few* and *moderate* chance of side effects as p_i = .09, .07, and .22, respectively. Thus, the *very-few* alternative would be chosen because it has the lowest adjusted probability of side effects. This choice does not change after application of the decision-weight function (f) when reasonable parameters for this function are assumed (i.e., determined by fitting data in Budescu et al., 2002). Thus, the model chooses the highly-credible but medium-utility alternative (Cefradoxil) over the lower-credibility best-utility alternative (Gentamycin), which involves sacrificing utility for credibility.

Example 2. Consider changing the decision in Figure 6 so that the *worst-utility* alternative (Kanamycin, *moderate* frequency of side effects) is the high-credibility alternative (see Table C1, lower left). Then, Equation C1 calculates $f[p_i]$ for *virtually no*, *very few* and chance of side effects as p_i = .09, .14, and .15, respectively. In this case, the best-utility alternative (Gentamycin, *virtually no* side effects) has the lowest adjusted probability and is chosen (even after application of the decision-weight function). In this example, the model rejects the high-credibility alternative and chooses the best-utility alternative despite its low credibility. This happens because too much utility must be sacrificed to go by credibility when the credible alternative has very low utility. Taken together, Examples 1 and 2 suggest that participants are more likely to reject the maximum-utility alternative to gain credibility when the high-credibility alternative has medium as opposed to worst utility, which describes the utility-premium hypothesis.

Credibility Gain Hypothesis

The credibility gain hypothesis stated that participants will choose the high-credibility alternative more often when this allows them to gain more credibility, i.e., more often with a large than a small credibility gain. In Experiment 2, all problems had one decision alternative with high credibility (e.g., a 5 star reputation rating). For some problems the two alternatives with lower credibility had medium reputation ratings (e.g., 3 stars), whereas for

(*Appendices continue*)

others the lower-credibility alternatives had low reputation ratings (e.g., 2 stars). The former problems had a small and the latter a large credibility gain. In Budescu et al.'s model, the credibility of a probability is inversely related to the width of the range of plausible probability values. So, high-, medium-, and low-credibility probabilities might be represented by ranges of ±0 (point estimates), ±.07 (as in Example 1 and 2) and ±.14, respectively.

Example 3. We now consider the prediction of the model when the worst-utility alternative (Kanamycin) has high credibility (i.e., large utility premium) but the low-credibility alternatives have reputations of 2 stars (i.e., a large credibility gain). This is shown in Table C1, lower-right. The 2-star credibility is represented in the model by a probability range of ±.14. With this change, Equation C1 calculates *f*[p_i] *for virtually no*, *very few*, and *moderate* chance of side effects as p_i = .16, .21, and .15, respectively. Here, the worst-utility outcome (*moderate* chance of side effects) is chosen (even after applying the decision-weight function). In this decision, the model predicts that a large amount of utility (the difference between best and worst utility) will be sacrificed to gain a large amount of credibility. Thus, a large utility loss is accepted in Example 3, where the credibility gain is large, but rejected in Example 2 where the credibility gain is small (Table C1, lower left). Taken together, Examples 2 and 3 suggest that Budescu's model leads to the credibility-gain hypothesis.

Received June 21, 2018
Revision received March 29, 2020
Accepted April 6, 2020 ■