An Audience-Centered Analysis of Cues to Which Group of Disputing Scientists is More Credible

Branden Johnson

Funding: U.S. National Science Foundation grant #1455867

Potential competing interests: The author(s) declared that no potential competing interests exist.

Abstract

Americans evaluated 22 cues laypeople might use to decide which group of disagreeing scientists is more likely correct, rating *perceptions* of the cue's reliability in determining the more valid side, its availability in their information sources, and average Americans' ability to use it effectively, plus self-reported cue use. Overall scientists' experience, research quality, and credentials rated highest on these "expressed value" criteria, findings which can complement future "normative value" (e.g., expert judgments of cues' reliability, availability, and usability) and "persuasive value" (whether exposure to a cue changes lay views) research advancing theory and practice regarding lay assessment of intra-science disputes.

Keywords: Scientific Disputes; Credibility; Cues.

Introduction

Audience-centered communication tailors its content and presentation to complement audience expectations and behavior to promote effective communications, including science communication (e.g., Ledford, Willett, & Kreps, 2012; Logan, 2001). Criteria for evaluating communications also include "audience evaluation" (e.g., do recipients deem a message accurate, helpful, understandable, etc.?), besides effectiveness measures, such as whether observers determine that the message is comprehensible (e.g., recipients report its meanings as observers expect) or that it prompts people facing the same objective risk to report the same subjective risk (Weinstein & Sandman, 1993). Assessing science communications requires taking audience views into account, not because they are the only or most important criterion, but because they can influence communication efficacy (e.g., publicly-devalued content may not affect public beliefs or behavior as intended, however well-designed). To rephrase, audience-centered or "expressed value" judgments can complement "normative value" judgments (e.g., experts' views of the communication's qualities) and "persuasive value" tests (whether communication exposure changes lay views or behavior as expected)

Audience views reported here ("expressed value") were elicited to supplement later experimental manipulations

("persuasive value"), and point to potential "normative value" studies, regarding cues that might help laypeople judge the relative credibility of disputing scientists. Before scientists reach consensus (or not), confusion about "the truth" can affect more than public knowledge: public disputes may undermine scientific authority partly stemming from its perceived consensus (Campbell, 1985; Collingridge & Reeve, 1986; Jasanoff & Wynne, 1998; Stilgoe, 2007; Zehr, 2000), with consequent effects on citizens' knowledge and/or uncertainty, trust in topical experts or in science, and policy and research support, among others (Irwin & Wynne, 1996). Yet concealing disputes could have similar or worse consequences (Beatty, 2006; De Melo-Martin & Intemann, 2013; Halfon, 2006; Miller, 2016; Solomon, 2007).

This study is part of a larger project exploring public reactions to and interpretations of scientific disputes, specifically the relative credibility of two groups of scientists, defining "credibility" as which position the lay observer deems relatively more valid. Intra-science disputes can be as critical to science communication as more-studied disputes between scientists and non-scientists (e.g., climate change and vaccine impacts; Van der Linden, Clarke, & Maibach, 2015; Van der Linden, Leiserowitz, Feinberg, & Maibach, 2015). Effects of and cues for mass disputes may differ from those between individual scientists (Thomm & Bromme, 2016).[1] Here lay Americans rated cues for scientific disputes varying in familiarity and salience: the nature of dark matter, recommended daily salt intake, and nanotechnology's risks and benefits. People generally agreed on "best" cues (e.g., those deemed most reliable for judging relative credibility) and "worst" cues.

Background

Credibility Cues

Lay understandings of scientific findings and scientific processes (Irwin & Wynne, 1996; Lysaght & Kerridge, 2012; Yearley, 1994) must be addressed when studying lay credibility cues. Few laypeople—including scientists in other disciplines or subfields (Hardwig, 1985)—understand technical scientific claims (Collins & Evans, 2007). Yet in certain local cases, and among the highly motivated, laypeople can identify limits to some science applications (Epstein, 1996; Irwin & Wynne, 1996; Lysaght & Kerridge, 2012; Wynne, 1989; Yearley, 1994), and *generally* determine who is (more) trustworthy and whether there is a consensus, if needed information is available and used (Anderson, 2011; Goldman, 2001). Lay processes can be studied with participant-observation and other qualitative approaches, experiments, and surveys quantifying self-reports; this study entailed the first survey.

This study limited its probes of lay credibility cues to citizens' judgments on multiple dimensions, not scientific accuracy (Collins & Evans, 2007, p. 31; Fallis & Frické, 2002; Frické, Fallis, Jones, & Luszko, 2005) or other objective measures: how people understand the *experts* rather than the *evidence* scientists provide (Collins & Evans, 2007; Fallis & Frické, 2002). Many field studies of societally-disputed science (Irwin & Wynne, 1996) cover cues regarding *anyone* a layperson deems expert or trustworthy in geographically or socially specific contexts (Collins & Evans, 2007, p. 53). The current concern was cues discriminating among scientists only, cues *ubiquitous* (widely available) even to those unfamiliar with the experts involved (Collins & Evans, 2007). Cues were limited to scientists *en masse*, excluding cues about

individual scientists' performance, such as understandable and acceptable explanations (Anderson, 2011; Wagenknecht, 2015), reactions to challenges (Brewer, 1998; Gelfert, 2011; Matheson, 2005), or honesty and ethical behavior about their own and opponents' claims (Fallis & Frické, 2002). "Civic epistemologies," how citizens evaluate knowledge claims for collective decisions, are institutionalized practices of judging science that differentiate societies (Jasanoff, 2005), which vary across countries on such attributes as trust, accepted bases of expertise, transparency of expert bodies, and ways to assess objectivity. This cross-sectional survey of an American sample allowed only for indirectly testing accepted bases of expertise (e.g., training, experience, or professional practices as more credible cues). Finally, this study excluded well-studied lay evaluations of credibility of specific disputant messages (e.g., Flanagin & Metzger, 2007; Gauchat, O'Brien, & Mirosa, 2017; Johnson, 2003; Johnson & Slovic, 1995, 1998; Van der Linden, Clarke, et al., 2015; Van der Linden, Leiserowitz, et al., 2015).

The communications literature on source credibility only somewhat overlaps with cues probed here. First, credibility of media sources or outlets (e.g., "newspapers" or "ABC News") is more closely related to persuasive value tests about whether cues affect observers, bringing in associated factors (e.g., news attention; Williams, 2012) unrelated to lay judgments of specific relative-validity cues. Second, the literature on credibility of sources cited within a given story has emphasized individuals. Third, cues covered in these literatures relate variously to those here. Some cues covered here are arguably covered partly as well by source-credibility scholars: e.g., the "information quality" cue here shares recency with tweet posting recency which influenced source credibility of a Twitter.com page owner (Westerman, Spence, & Van Der Heide, 2014), while one of the strongest source characteristics (expertise; e.g., Wilson & Sherrell, 1993) is echoed by some cues here (e.g., relative experience or prestige of protagonists' universities), but probably weaker because all disputants are scientists. Some source-credibility cues are excluded here: e.g., manipulated or perceived honesty characterizes individual rather than masses of scientists, and message quality's mutual effects on credibility judgments (e.g., Slater & Rouner, 1996) does not apply as there is no message here to evaluate. Finally, some cues here are irrelevant to assessing traditional source-credibility sources: e.g., individual disputants cannot represent a majority of scientists on one side. In short, cues covered here potentially expand, without contradicting, those in previous source credibility research in communications, marketing, and associated fields.

Specific cues have been identified both empirically and logically. Field observations (Bubela et al., 2009; Irwin & Wynne, 1996) identified social/institutional affiliations, self-interested information processing (Frické et al., 2005), reference group messages (Frické et al., 2005), and accuracy. Despite lay epistemic dependence on experts—they cannot independently assess experts' technical claims (Hardwig, 1985)—philosophers and scholars of social studies of science cited cues non-experts *might* use: credentials (Goldman, 2001), the scientific majority's position (Goldman, 2001), interests and biases (Goldman, 2001), track records (Goldman, 2001), and length of experience (Collins & Evans, 2007). Ubiquitous cues by which scientists evaluate other scientists' claims could help laypeople: employer, failures, doctoral university's size and prestige, nationality (Collins & Evans, 2007, pp. 50–51, note 10). Factors affecting trust of hazard managers (Earle & Cvetkovich, 1995; Kuklinski, Metlay & Kay, 1982; White & Johnson, 2010) also might judge disputing scientists: employer, salient shared values, reference groups, precautionary performance. Cues volunteered by focus groups in this or an earlier study (Maxim & Mansier, 2014) included their employer, political ideology, proportion of studies per side, research protocols, data recency, and use of established methods. Blogs about the H1N1 vaccine frequently

cited scientific content, interests and bias, blind trust (i.e., no cue use), and “commonsensical social judgments” seemingly dominated by generalized suspicion, but rarely used formal authority indicators (e.g., credentials; Kutrovátz, 2010).

Repetition across diverse literatures suggested a taxonomy would improve this study’s design, and perhaps future theory. Seven classes of credibility cues were defined:

- interests: e.g., employer pressure to earn or retain money or prestige; scientists’ similar motivations
- shared values: e.g., with scientists or reference groups
- credentials: e.g., degrees, awards
- performance: e.g., scientists’ prior accuracy, their experience
- demographics: e.g., nationality, demographic similarity to observer
- vote-counting: i.e., proportion of scientists or studies per position
- research quality: e.g., the field’s uncertainty, use of methods such as control groups or replication

These seven categories appear to represent a minimal set of cue types; e.g., how scientists work (research quality) cannot be subsumed under outcomes (performance), nor scientists’ motivations (interests) conflated with the majority position in a scientific dispute. At least the taxonomy provides a foundation for wider discussion of potential lay cues to scientific disputants’ relative credibility.

Research Questions

This study aimed to identify what laypeople *say* are better or worse cues to the relative credibility of disputing scientists, expressed value data to *complement* experiment-revealed or expert-normative values for the same cues, just as measures of subjective and objective topical knowledge carry different implications for communications (e.g., Rose et al., 2019).

We could ask people which are best and worst for relative-validity judgments on a list of cues, or elicit a complete ranking from best (#1) to worst (#22, in this case). This “explicit” ranking directly contrasts some or all cues, and may integrate judgments across multiple dimensions salient if not all conscious to the respondent, but not necessarily known to the researcher. Rankers’ social desirability or self-presentation motivations might rank high those cues they think researchers favor. The number of cues ranked can burden respondents, potentially undermining results’ validity and reliability.

Alternatively, the researcher could impute “implicit” rankings from mean ratings of one or more scales, such as a cue’s value for determining relative credibility of disputants’ positions. Raters are unlikely to compare all cues before rating, as full explicit ranking does, much less compare cues across scales, nor be aware of rankings being imputed from their collective ratings. Salience likely varies across scales or raters, although rarely assessed by researchers. Yet rating one cue, on one scale at a time, burdens raters less than explicit ranking, increasing mean scores’ reliability despite potentially lower validity.

Many laypeople are only vaguely familiar with the existence of or reasons for scientific disputes, may use rules of thumb in cue evaluations they cannot fully articulate, and might want to present themselves positively (Collins & Evans, 2007). Researchers should take no one approach or answer at face value, given these factors. Recent work on why

laypeople think scientists might disagree (Johnson & Dieckmann, 2018; cf. Dieckmann & Johnson, 2019)distinguished with difficulty people favoring one reason (e.g., incompetence versus topical complexity), even controlling for dispute interest. The authors suggested that this task may be novel enough for respondents that their on-the-fly answers are less coherent collectively than for more familiar questions. If true here, we need both explicit and implicit ranking to answer

RQ1. *What do lay Americans report as better and worse credibility cues for scientific disputes?*

Benefits of multiple answers to RQ1 imply multiple scales should underlie implicit rankings. How people decide the relative credibility of *groups* in a mass scientific dispute was central here, lay judges may distinguish such cues from those to the competence of "scientists," although most cues here apply logically to both.[2] Thus each was subject to a separate rating scale (hereinafter "inter-group reliability" and "scientist reliability," respectively).

These reliability scales were complemented by three others. Self-reports of cue use ("use") may echo reliability ratings but are not independent observations. People's imputed reliability and use rankings might converge due to accuracy (people use only reliable cues), and/or to self-presentation concerns.

A cue's reliability is irrelevant if it is inaccessible, so respondents also rated whether their information sources about scientific disputes include it ("availability"). This subjective measure implies that a reliable available cue trumps a reliable unavailable one. If laypeople *and* scientists agree a cue is reliable but unavailable, this information also could inform science communication. If reliability and use are only moderately correlated, availability may be a moderator: self-reported use of more reliable cues is higher when people see them as available.

The fifth scale concerned how able the rater thought "the average American" would be to find, process and use the cue to judge relative credibility ("ability"). Belief in one's own competence (implied but not directly assessed by the use scale) and in others' competence may not converge (cf. drivers' belief they are above average; Svenson, 1981). If overall these rankings diverge, this might shape science communication, whether to make "good" cues more visible, accessible and/or usable (and "bad" cues less so), or to offset people's unjustified relative optimism about their skill in understanding scientific disputes. Comparing rankings derived from this and other scales (e.g., use) *might* be used to infer third-person effects (Davison, 1983), belief others will be persuaded by the message but not oneself. But no direct comparison of self to others was elicited, no mention was made of persuasion, and *none* of the five scales asked people to rate whether or how much they thought the cue would influence them *or* others (a judgment whose accuracy could not be assessed via self-reports, only by experiments).

Rankings imputed from the two reliability ratings were expected to be most correlated, followed by reliability's correlations with use and availability, with availability perhaps moderating the reliability-use link. Rankings from (average-American) ability ratings were expected to correlate least with other rankings. Explicit reliability rankings were expected to correlate most with implicit reliability rankings.

RQ2. *Do implicit rankings converge across rating scales and with explicit rankings?*

Interest in scientific disputes likely varies widely, potentially affecting cue ratings and rankings. Those highly

interested may not be more familiar with certain cues, but will more likely engage with evaluation, perhaps providing more valid and reliable answers. Dispute topic also might affect cue responses: e.g., arcane subjects such as dark matter seem unlikely to offer, or make salient, cues about economic interests as recommended salt intake. Belief science yields "truth" and mistrust of science might increase perceived relative value of credentials and interests cues as respective signals of scientists' competence or bias, while those familiar with scientific reasoning might emphasize research quality cues. Demographics also might affect cue responses, although no prior data or theory suggest how. Testing these variables' associations with cue ratings and rankings helps us grasp the extent of lay consensus.

RQ3. *Do dispute interest, topic, knowledge of or attitudes toward science, or demographics affect cue ratings or rankings?*

Methods

Qualitative Research

Three focus groups (n = 35), varying education and political ideology, preceded this survey. The author defined mass scientific disputes (versus among individuals), and moderated a discussion of "what catches your attention when groups of scientists disagree?" Then several short paragraphs, each describing an actual scientific dispute (e.g., wolves' ecological effects in Yellowstone; molecular biology), were distributed. For each dispute people explained their answers to four questions: the degree of disagreement they observed; their "good or bad" feelings about the dispute; why they thought some scientists took one position versus another; and which position the focus group members thought "more likely to be correct." Any explanations entailing potential cues (e.g., "follow the money" on research funding) were probed, including thoughts related to potential scales. Members' potential actions taken or avoided given the dispute were discussed before reviewing the next scenario. Finally, group members were asked about cue differences across scientific topics, and thoughts prompted by a list of potential cues, not all mentioned earlier.

Focus group members emphasized cues of self-interest, whether scientists had a financial stake in their dispute position; accuracy, whether they had a history of being right; precaution, whether scientists tended to be cautious or reckless in research methods and conclusions; and counting "votes" for a position by the proportion of studies supporting it.

Instrument and Measures

Initial questions on respondents' awareness of and reaction to scientific disputes (Table 1, "Dispute interest") were followed by one dispute scenario (Table 2). People indicated which dispute position they thought more valid, their confidence in this answer (Table 1, "Correct position" and "Confidence"), and likelihood of 16 explanations for why scientists disagree (discussed elsewhere; *reference omitted*).

Then people rated 22 cues (Table 3), randomly ordered per respondent, on five scales each (Table 1, "Scales");

ratings were elicited for all five scales before the next cue. Mean scores for a cue on a scale generated "implicit rankings" reported below; e.g., Cue A would implicitly rank higher than Cue B on availability if its mean score on that scale was higher. After all cues were rated, people ranked them in three steps (Table 1, "Ranking"). First, they indicated which were among the three best and three worst on *reliability* without distinguishing among them, expected to be easier than eliciting all 22 cue ranks. Second, they ranked the three "best" on order (1 best, 2, 3). Third, they ranked the three "worst" (1 worst, 2, 3). Proportions of respondents ranking a cue as best/worst or in the top three produced "explicit rankings": e.g., if 30% of the sample rated Cue A "best" and 25% Cue B, Cue A would rank above Cue B in explicit "best" rankings.

Respondents finished with 10 questions on beliefs about scientific positivism (e.g., "Science provides objective knowledge about the world," 1 *strongly disagree*, 5 *strongly agree*; Rabinovich & Morton, 2012; Steel, List, Lach, & Shindler, 2004), an 11-item scientific reasoning scale (e.g., for understanding of reliability, "A researcher develops a new method for measuring the surface tension of liquids. This method is more consistent than the old method. True or False? The new method must also be more accurate than the old method"; the answer is false; Drummond & Fischhoff, 2017), a 6-item mistrust of scientists scale (e.g., "People trust scientists a lot more than they should," 1 *disagree very strongly*, 7 *agree very strongly*; Hartman, Dieckmann, Stantsy, Sprenger, & DeMarree, 2017), and demographic measures.

Table 1. Selected Measures

Topic	Measure
Dispute interest	Sometimes a large group of scientists disagrees with another large group of scientists about the causes or effects of a natural event or technology. Have you ever heard about such a scientific dispute? (Yes/Don't know/No) [If yes/DK] Have you ever tried to decide which group of scientists was more likely to be correct? (Yes/Don't know/No)
Correct position	Which of the two positions about this topic do you think is more likely to be correct? *Salt*: Position A- People at risk of health problems should cut dietary salt intake by one-half, the rest by one-third Position B- All people should cut dietary salt intake by one-third *Dark Matter*: Position A- Most of the matter in the universe is made of WIMPs Position B- Most of the matter in the universe is made of axions *Nanotechnology*: Position A- Nanotechnology will have many benefits Position B- Nanotechnology could have risks we don't know about
Confidence	How confident do you feel about your choice of one group of scientists as more likely to be correct on this topic than the other group? (1 *I have little or no idea which group is more expert*, 2 *I have some idea which group is more expert—but I am more unsure than sure*, 3 *I have a good idea which group is more expert—and I am more sure than unsure*, 4 *I am pretty sure I know which group is more expert*; adapted from Ref. 25)
Scales	1) How reliable is this information as a signal of whether scientists are competent? (1 *not at all reliable*, 5 *extremely reliable*) 2) How reliable is this information as a signal that one group of scientists is more competent than another group of scientists? (1 *not at all reliable*, 5 *extremely reliable*) 3) How often do your real-life information sources about scientific disputes include this kind of information? (1 *never*, 5 *always*) 4) How able do you think the average American would be to find, understand and use that information to decide which group of scientists was more likely to be correct? (1 *not at all able*, 5 *extremely able*) 5) How often have you used this kind of information to decide which group of scientists involved in a disagreement with other scientists was most likely to be correct? (1 *never*, 5 *always*)
Ranking	Below is a list of cues again, which we would like you to rank in terms of their overall performance—considering the ability of the average American to find, understand and use that information—to help people decide which group of disagreeing scientists is most likely to be correct. In other words, which cues are the best and worst for figuring out which group of scientists are correct? First, select the "Best 3" cues. Second, select the "Worst 3" cues. Now please rank the 3 best cues (the best should be #1) you selected by dragging them up or down the list Now please rank the worst 3 cues (the worst should be #1) you selected by dragging them up or down the list

Table 2. Dispute Scenarios

Dark Matter. About 85% of all matter in the universe is 'dark matter' which scientists know is there due to its gravitational pull on visible matter such as galaxies and radiation, but they cannot see it and do not know what it's made of. Some scientists think dark matter is made of axions, (currently hypothetical) subatomic particles formed in the core of a star when X-rays scatter off protons and electrons in a strong electric field. Other scientists think dark matter is made of the lightest of the neutralinos, a different set of (currently hypothetical) subatomic particles resulting from the decay of squarks and other relatively heavy particles. Until research can answer this question, scientists do not know what makes up most of the matter in the universe.
Dietary Salt. Scientific studies on nutrition agree that Americans eat too much salt, both sprinkled on food and far larger amounts in processed foods (bread, cereal, salad dressing, canned vegetables and soup, ketchup, etc.). But they disagree on what level is safe, especially for people at risk for heart disease and stroke: over half the U.S. population, including those with high blood pressure, older than 51, African Americans, and with diabetes, chronic kidney disease, and congestive heart failure. Some scientists think everyone should reduce intake by one-third, with people at risk cutting salt intake by half. Other scientists think there is little or no health benefit from cutting intake in half or giving different recommendations for at-risk subgroups and others.
Nanotechnology. Scientific studies on nanotechnology, a new field which exploits how materials change when very small (1,000–8,000 times narrower than a human hair), disagree about its impact. Some scientists stress benefits: medicines (nano-particles can breach the brain-blood barrier, carrying medicine to treat Alzheimer's or brain cancers hard or impossible to treat now), self-cleaning windows, packaging extending vegetables' shelf life, pollution cleaners, clothing that repels odors and lasts longer, and so forth. Other scientists stress that the same qualities could impose harm (nano-particles can reach places in the human body larger materials cannot), far more research is done on product development than potential risks, and many benefits claims are so far unproven.

Table 3. Cue Categories and Items

Interests
Employer: What institution (for example, government, business, nonprofit) or specific organization employs the scientists who take that position.
Scientist interests (Grants): Whether the scientists who take that position can obtain research grants as a result.
Scientist interests (Business): Whether the scientists who take that position can be promoted, sell a patent, or start a business as a result.
Scientist interests (Prestige/Influence): Whether the scientists who take that position can earn prestige or influence as a result.
Self-interest: Whether you will benefit from the position that the scientists take (for example, get better or cheaper products, more safety, environmental quality, or convenience).
Shared Values
Salient shared values: The scientists who take that position appear to share values with you that you think are important to making decisions about this topic.
Reference group positions: Whether groups or organizations you trust take the same position, or announce that they trust the scientists who take that position.
Credentials
Type of Degree: Whether the scientists who take that position have advanced degrees (such as Ph.D. or M.D.) in a field closely related to the topic.
Source of Degree: Whether the scientists who take that position have degrees from well-known and respected universities
Awards: Whether the scientists who take that position have received prestigious awards (Nobel Prize, U.S. National Medal of Science, etc.).
Performance
Accuracy: Whether the scientists who take that position have been right or wrong about similar scientific issues in the past.
Precaution: Whether the scientists who take that position tendency in their work to be cautious or enthusiastic, risk-taking or risk-averse, or take extreme versus moderate positions
Experience in the field: How long they have been working on this topic; whether they have spent a lot of time learning practical or theoretical details of the issue.
Demographics
Nationality: Whether the scientists who take that position are Americans or not.
Similarity: Whether the scientists who take that position are similar to you; for example, the same gender or ethnic group or age.
Vote-Counting
Scientists: The proportion of scientists (for example, 50% or 90%) that take this position.
Studies: The proportion of peer-reviewed scientific studies (for example, 50% or 90%) that take this position.
Research Quality
Uncertainty of the field: How much uncertainty you think there is in understanding and making accurate predictions for this topic.
Control groups: Whether the scientists test how outcomes change when a factor is present rather than absent (for example, health outcomes when people randomly get a new drug or a fake one).
Comparison: Whether the scientists who take that position have tested in the same studies the effect of their explanation against the effects of competing explanations.
Study replication: Whether the scientists who take that position have evidence favoring that position from more than one study, conducted by scientists who are independent of each other.
Information quality: Whether the evidence supporting that position is up-to-date, or has been collected with the best available techniques.

Sampling

A sample of Americans from Survey Sampling International's online panel responded October 16–19, 2015 (median completion = 22.7 minutes). Removing 26 responses (22 for completion < 8 minutes; 4 nonsensical answers) left 534 responses. A third each read one dispute scenario (33.5% dark matter, 34.3% salt, 32.2% nanotechnology), topics based upon focus group results and an earlier experiment (*reference omitted*).

Analyses

Analyses included *t* tests, one-way ANOVAs, and multiple linear regressions. Although frequency of rank ties varied, Kendall's tau-b (τ) rank-order correlation was used consistently for its ability to account for ties, yielding weaker, less significant correlations than did Spearman's rho rank-order correlation. For multiple comparisons, the Benjamini & Hochberg (1995; Glickman, Rao, & Schultz, 2014) method was used to retain sufficient power to minimize false positives, whose testing of significance for individual items is more conservative than Bonferroni-type adjustments (which tests a universal null hypothesis against an omnibus alternative). The false discovery rate (FDR) $d = .01$ used here, recommended for empirically- (versus theoretically-) driven conditions (Glickman et al., 2014), meant no more than 1% of significant tests were expected to be false positives. Tests of availability's moderation of the reliability-use relationship used PROCESS 3.0, model 1 (Hayes, 2018), with 5000 bootstrap samples for 95% confidence intervals, and the HC4 heteroskedasticity-consistent standard-error estimator (Hayes & Cai, 2007).

Results

Demographics

Mean age was 38.60 ($SD = 15.54$, median = 34, range 18–82); 74% were women, 74.2% non-Hispanic whites, 26.9% had a high-school-graduate education or less, 35.7% college degrees or more, 26.3% politically conservative, and 28.5% liberals. The sample was more female than the U.S. population (50.8%) and more-educated (29.3% college degrees or more among those 25-plus), but similar in age (median = 37.4) and white ethnicity (72.4%), based on 2010–2014 American Community Survey 5-year estimates (U.S. Census, 2016). No significant demographic differences across dispute scenarios confirmed random allocation.

General Views

Two-thirds (66.3%) of the sample were aware of scientific disputes; 15% did not know. Among these respondents, 64.8% had ever tried to decide which side was correct, 23.1% did not, and 12% did not know ($n = 432$). A Dispute Interest nominal measure distinguished those low (unaware; never tried to decide who was right; 37.8%), ambiguous (did not know if they had tried, or did not know if aware but had tried to decide who was right; 13.7%), and high (aware and tried to decide; 48.5%) in interest. Dispute interest differed insignificantly by scenario.[3]

Knowledge of scientific reasoning was moderate ($M = 5.73$, $SD = 2.39$; median = 6 correct of 11). This differed across scenarios ($F(2,521) = 3.21$, $p = .041$)—readers about nanotechnology understood scientific reasoning less ($M = 5.37$, $SD = 2.33$) than did dark matter readers ($M = 6.02$, $SD = 2.43$, $p = .034$)—and by dispute interest ($F(2,521) = 19.61$, $p < .0005$): high-DI respondents showed more understanding ($M = 6.38$, $SD = 2.45$) than did low-DI ($M = 5.09$, $SD = 2.17$,

$p < .0005$) or ambiguous-DI groups ($M = 5.21$, $SD = 2.12$, $p < .0005$). Mistrust of scientists was moderate ($M = 24.67$, $SD = 7.90$, median = 25, range 6-42; $\alpha = .89$), insignificant at $p < .05$ across topic or interest. Scientific positivism items (two non-positivist items reversed) formed an unreliable index ($\alpha = .54$) without omitting those two items ($\alpha = .74$; $M = 27.23$, $SD = 5.25$, median = 27, range 8–40). Positivism differed insignificantly across scenarios or interest.

Cue Ratings

Task Difficulty and Straightlining

Survey researchers screen for response biases, such as acquiescence bias, agreeing or disagreeing with every question; extreme responding, such as answering only with 1s or 5s on 1-5 Likert scales; or straight-lining, answering all scales the same (e.g., here awarding all 2s to the similarity cue, all 4s to control groups). Chronic straightlining can reflect respondent inattention or indifference to question content, or inability to make requested distinctions, prompting that person's removal from analyses. Yet just as zero straightlining need not indicate fully valid and reliable responses, someone providing different straightlined ratings across cues (see similarity versus control-group example above) may still rate cues' relative value as best she can. Asking people to rate 22 cues on five scales each might amplify straightlining, due to the task's novelty (discussed earlier) and cognitive fatigue from making 110 judgments, although randomizing cue order here should have minimized fatigue variance across cues. Most implicit rankings (excluding average-American-ability, unless respondents identified as "average") were expected to correlate (Research Questions), which identical ratings across criteria would only amplify.

Most respondents avoided straightlining per cue (68.5% nationality to 77.0% comparison), but only 25.7% ($n = 137$) avoided it for all cues ("Zeroes"). Two-thirds (64.8%; $n = 346$) straightlined for zero to five cues of the 22 ("Minimals"), with identical scores rarely awarded to different cues under multiple-cue straightlining. Rankings for the most important scales —inter-group reliability and ability—correlated highly between these groups (τs = .82 and .84, respectively). Rankings including people straightlining up to seven cues still correlated with Zeroes' rankings at $\tau = .82$ for inter-group reliability, but only $\tau = .68$ for ability, expanding the sample modestly (71%, $n = 379$). Some analyses here (regression analyses, contrasts in dispute interest) benefited from larger sample sizes, for which the Minimal group seemed sufficient. Online Supporting Information (SI) for this paper reports both Zero and Minimal results, but the main text focuses on the latter.

Independent *t* tests with listwise deletion found no differences significant at $p < .05$ between Zeroes ($n = 127$, given missing data on comparison variables) and those doing *any* straightlining ($n = 370$) in demographics (gender, age, education, non-Hispanic white ethnicity, political ideology), dispute topic, dispute interest, or scientific-reasoning knowledge. The only significant differences at $p < .05$ included science mistrust (Zero $M = 25.81$, $SD = 7.84$; Any $M = 24.11$, $SD = 7.86$, $p = .036$) and scientific positivism (Zero $M = 28.10$, $SD = 5.01$; Any $M = 26.95$, $SD = 5.29$, $p = .033$), neither meeting the FDR criterion for statistical significance. Repeated with the Minimal group ($n = 319$; Other straightliners, $n = 178$), four of ten variables exhibited differences significant at $p < .05$: scientific-reasoning knowledge (Minimal $M = 5.96$, $SD = 2.42$; Other $M = 5.42$, $SD = 2.32$, $p = .016$), dispute interest (Minimal $M = 2.18$, $SD = 0.93$; Other $M = 1.98$, $SD = 0.89$, $p = .020$), education (Minimal $M = 4.91$, $SD = 1.43$; Other $M = 4.61$, $SD = 1.45$, $p = .023$), and white

ethnicity (Minimal $M = 0.71$, $SD = 0.45$; Other $M = 0.80$, $SD = 0.40$, $p = .024$). None met the FDR criterion![4] The only difference between these sub-samples and other respondents concerned straightlining's extent, so analyses focus on people apparently both motivated and capable of distinguishing among unfamiliar cues and criteria.

By focusing on those with minimal straightlining—a larger subsample than the zero-straightlining group yielding potentially more reliable estimates, and likely reflecting inability to determine how to rate a given cue on the five scales except that it rates more or less than another cue—later analyses excluded respondents most likely to be indifferent or inattentive. Zero straightlining yielded little difference in results reported here for minimal straightlining.[5]

Rating Differences and Availability as Moderator

One-way ANOVAs revealed ratings differed insignificantly ($p < .05$) by topic. On dispute interest (DI) Minimals exhibited statistically significant ANOVAs (36 of 110; SI 1), but differences in only three cue-scale ratings met the FDR criterion. For the similarity cue the high-DI group ($n = 172$) differed from the low-DI group ($n = 110$) on scientist reliability ($F(2,312) = 12.65$, $p < .0005$; $M = 2.38$, $SD = 1.29$; $M = 3.06$, $SD = 1.00$, $p < .0005$) and inter-group reliability ($F(2,312) = 17.17$, $p < .0005$; $M = 2.30$, $SD = 1.22$; $M = 3.05$, $SD = 1.11$, $p < .0005$). On both reliability scales low-DI respondents scored similarity higher. Reported use of the replication cue was higher for high- versus low-DI respondents ($F(2,312) = 13.37$, $p < .0005$; $M = 3.58$, $SD = 0.96$; $M = 2.91$, $SD = 1.22$, $p < .0005$). Later analyses contrasted high- and lower-DI rankings only.

Availability's potential moderation of the relationship of inter-group reliability and use was poor: only one cue (degree type) exhibited an interaction significant at $p < .05$, and three at $p < .10$ (similarity, business, shared values).

Implicit Rankings

Implicit rankings derived from mean scores on the same scale across cues. Rank-order correlations of Minimals' implicit rankings appear in Table 4. Each cell reports full-sample results, and those for high- and low-dispute interest, in that order; bracketed correlations are between high- and low-DI rankings. Reliability rankings were highly correlated, particularly for high DIs. Also as expected, reliability correlated highly with self-reported use, with an even greater gap between high-DI and low-DI than for reliability; implicit rankings for scientist reliability and use were identical. Minimals reported reliable cues as available without much divergence by dispute interest. Reliability correlated moderately with judged ability, with mixed and modest differences on whether high DIs were more likely to think the average American could "find, understand and use" reliable cues. Finally, implicit-ranking correlations between high- and low-DI respondents for each scale were roughly $\tau = .5$.

Table 4. Kendall's tau-b Implicit Rank-Order Correlations Across Scales, Overall and by Dispute Interest

	Criteria				
	1	2	3	4	5
1. How reliable is this information as a signal of whether scientists are competent?	[.49***]				
2. How reliable is this information as a signal that one group of scientists is more competent than another group of scientists?	.95*** .87*** .59***	[.48**]			
3. How often do your real-life information sources about scientific disputes include this kind of information?	.70*** .79*** .68***	.69*** .81*** .67***	[.57***]		
4. How often have you used this kind of information to decide which group of scientists involved in a disagreement with other scientists was most likely to be correct?	.80*** 1.00*** 1.00***	.82*** .87*** .59***	.78*** .79*** .68***	[.49***]	
5. How able do you think the average American would be to find, understand and use that information to decide which group of scientists was more likely to be correct?	.56*** .48** .59***	.55*** .48** .48**	.73*** .64*** .57***	.61*** .48** .59***	[.50***]

*Minimal straightlining results (n = 343-346 full, 183-185 high DI, 124-127 low DI). Cell numbers, vertically, report two-tailed Kendall's tau rank-order correlations (τ) for the full sub-sample, and high and low DI groups. Bracketed correlations in diagonal report rank-order correlations for high- and low-DI groups for Minimal (bottom) groups. Rankings n = 22. † p < .10 * p < .05 ** p < .01 *** p ≤ .001*

These correlation patterns illustrate mostly convergent implicit rankings, plus the value of multiple rankings given occasional differences by dispute interest. These results address RQ2 about convergence among implicit rankings. They confirm expectations the two reliability scales would yield similar rankings, although somewhat less for low-DI respondents, so subsequent ranking analyses include only inter-group reliability. By also confirming correlations—particularly among high DIs—between inter-group reliability and use claims, further analysis of use data would be redundant. Thus to answer RQ1 on which cues were implicitly "best" and "worst," we emphasize inter-group reliability, availability, and ability rankings.

Full implicit ranking data on these scales (SI 2 Zeroes, SI 3 Minimals) are summarized in Table 5 as the number of times, among the full Minimal sub-sample, high- and low-DI respectively, a given cue ranked among the best or worst three cues on a scale.[6] The best cues fell into research quality (excluding field uncertainty) and credentials categories, plus experience in the performance category for low-DI respondents. High-DIs were more likely to rank research quality cues among the best on these three scales, but low-DIs often agreed.

"Worst" cues were interests, demographics, and field uncertainty; counting scientists' "votes" (which dispute position has a majority) was seen as unavailable by low-DI respondents.

Table 5. Implicit Rankings of Cues by Inter-Group Reliability, Availability, and Ability (summary; SI 2-3 for full results)

	Inter-Group Reliability		Availability		Ability	
	Best 3	Worst 3	Best 3	Worst 3	Best 3	Worst 3
Interests						
Employer						
Grants				√ H		√H
Business		√√		√L		√L
Prestige		√L				√L
Self-interest						
Shared values						
Salient values						
Reference group						
Credentials						
Type of degree	√ L		√√		√√	
Source of degree					√ H	
Awards					√√	
Performance						
Accuracy						
Precaution						
Experience	√L		√ L			
Demographics						
Nationality		√√		√H		
Similarity		√ H				√√
Vote-counting						
Scientists				√ L		
Studies						
Research quality						
Uncertainty of field				√ L		√ H
Control group	√ H		√ H			
Comparison	√H					
Replication	√ H		√H			
Information quality	√L		√√		√L	

Derived from SM2 (Minimal Group) on whether overall sub-sample, high dispute-interest (DI), and/or low DI groups ranked cue in the best three or worst three for the column's scale. Number of check marks indicates how many groupings so ranked the cue. If less than all (3 checks), letter indicates whether high- (H) or low- (L) DI respondents ranked it so (e.g., √ L indicates full sub-sample and low DI group gave this ranking, while high DI group did not).

On best implicit ratings high- and low-DI respondents agreed for no reliability, two availability, and two ability ratings, or four of 15 top-ranked cues across the three scales; on worst implicit ratings they agreed on two reliability and one ability ratings, or three of 14 bottom-ranked cues. However, they generally agreed on categories of cues, if not specific ones,

better or worse on these criteria.

Explicit Rankings

Some 38 people ranked more or fewer (5 people) than three “best” cues (161 provided no rankings at all), and 39 more or fewer (3) than three “worst” cues, all excluded from analyses. Two different rankings—based on the proportion who ranked a cue absolutely best or worst, or among the three best or three worst—were derived, with rank-order correlations in Table 6 for the full Minimals sub-sample, and high and low DI groups.

Explicit “best” rankings correlated well as expected, and more strongly for explicit “worst” rankings. Both “worst” ranking approaches converged across dispute interest, slightly less for the “best 3” approach. High and low DIs converged little for “best #1” cues. Expected but overall weaker negative correlations between “best” and “worst” rankings indicated people were not simply reversing “best” rankings to rank “worst” cues, particularly among low DIs.[7]

Table 6. Kendall’s tau-b Rank-Order Correlations Across Minimals’ Explicit Rankings, Overall and by Dispute Interest

	1	2	3	4
1. Best #1	[.43*]			
2. Best 3	.76*** .77*** .61***	[.42**]		
3. Worst #1	-.45** -.57*** -.06	-.47** -.53*** -.06	[.66***]	
4. Worst 3	-.38* -.57*** -.06	-.37* -.54*** -.06	.81*** .77*** 1.00***	[.67***]

*n = 343-346 full, 183-185 high DI, 124-127 low DI). Cell numbers, vertically, report two-tailed Kendall’s tau rank-order correlations (τ) for the full sub-sample, and high and low DI groups. Bracketed correlations in diagonal report rank-order correlations for high- and low-DI groups for Minimal groups. Rankings n = 22. † p < .10 * p < .05 ** p < .01 *** p ≤ .001*

Explicit rankings by Minimals (Table 7; full details, Zeroes SI 4, Minimals SI 5) thus revealed slightly broader answers to RQ1 about best and worst cues than among implicit rankings, if with more agreement across dispute interest. Overall best cues were experience and degree type, with control groups and replication also favored by high dispute-interest respondents, and awards and reference groups by low DIs. The worst overall were self-interest of the lay observer and a scientist’s business opportunities, with demographic similarity (and to a lesser degree prestige and nationality) mentioned by high-DIs and employer by low-DIs. Field uncertainty ranked among both best and worst cues (fifth best *and* seventh worst), as did field uncertainty (eighth best *and* fourth worst).

Table 7. Explicit Rankings of Best and Worst Cues

	Best		Worst	
	#1	Top 3	#1	Top 3
Interests				
Employer		√L	√L	√√L
Grants			√	
Business			√√	√√
Prestige			√√	√L
Self-interest			√√	√√
Shared values				
Salient values				√H
Reference group		√L		
Credentials				
Type of degree	√√	√√		
Source of degree				
Awards	√L			
Performance				
Accuracy	√√H	√√H		
Precaution				
Experience	√√	√√		
Demographics				
Nationality			√√H	√√H
Similarity			√H	√√H
Vote-counting				
Scientists				
Studies				
Research quality				
Uncertainty of field	√L	√√L	√L	√L
Control group	√H	√√H		
Comparison				
Replication	√√H	√H		
Information quality	√√			

Derived from SM5 ((Minimal Group): Number of checks indicates whether overall sub-sample, high dispute-interest (DI), and/or low DI groups ranked that cue among the best three or worst three for reliability. If less than all (3 checks), associated letter indicates whether high- (H) or low- (L) DI respondents ranked it so (e.g., √√ L indicates full sample and low DI group gave this ranking, while high DI group did not).

Rank-order correlations of explicit versus implicit rankings (Table 8) further answer RQ2 about convergence. Of 24 correlations involving dispute interest sub-groups, all but four (ability, concerning #1 best and worst rankings) were stronger for high- versus low-DIs, but all cases exhibited identical signs. The #1 versus top-3 methods did not obviously differ. Explicit best rankings exhibited positive, thus convergent correlations with all implicit rankings. Scales most correlated with explicit best rankings were inter-group reliability and availability, indicating people largely followed instructions specifying inter-group reliability as the explicit-ranking criterion (Table 1). The more people saw cues as reliable for determining which group of scientists is more credible, and available in their own information sources, the more likely they were to explicitly rank those cues best. Given positively correlated implicit and explicit best rankings, and negatively correlated explicit best and worst rankings, negative correlations of implicit best and explicit worst rankings were expected, and again the greater negative correlation with implicit inter-group reliability rankings underlines that people followed explicit-ranking instructions.

Table 8. Kendall's tau-b Rank-Order Correlations Between Implicit Rankings and Explicit Rankings, Overall and by Dispute Interest

	Explicit Rankings			
	Best		**Worst**	
Implicit Rankings	**#1**	**Top 3**	**#1**	**Top 3**
Inter-group reliability	.55*** .65*** .34*	.53*** .62*** .28†	-.79*** -.76*** -.54***	-.81*** -.82*** -.54***
Availability	.50*** .58*** .43**	.50*** .55*** .36*	-.62*** -.66*** -.36*	-.56*** -.61*** -.36*
Ability	.47** .36* .35*	.43** .35* .41**	-.49** -.38* -.30†	-.43** -.33* -.30†

*Numbers in each cell, vertically, report two-tailed Kendall's tau-b rank-order correlations (τ) for the full sample, and respondents with high and low dispute interest. n = 22. † p < .10 * p < .05 ** p < .01 *** p < .001 or better*

Factors in Inter-Group Reliability Ratings

Multiple linear regression analyses using Minimal subjects probed factors affecting inter-group reliability ratings for selected "best" and "worst" cues in both implicit and explicit rankings. Non-reliability dimensions were omitted for brevity, with cues selected to avoid multiple examples from the same category. "Best" examples included degree type, experience, and replication; "worst" examples included business opportunities and nationality.

Each reliability rating was regressed on demographics (gender, age, education [dummy coded, reference category = < "some college"; this level of U.S. education or higher is most associated with exposure to scientific thinking], ethnicity, political ideology), dispute-interest dummies (reference category = no), topic dummies (reference category = dietary salt),

scientific reasoning knowledge, science mistrust, and scientific positivism.[8] Table 9 shows that dominant effects came from scientific reasoning knowledge and scientific positivism beliefs: greater knowledge meant less positive reliability ratings, while belief in science's ability to reveal truth did the reverse, across these five cues. Better replication ratings also stemmed from interest in scientific disputes, viewing the dietary salt versus the nanotechnology scenario, and being older. Better ratings for scientists' business opportunities also were associated with non-Hispanic white ethnicity, and for nationality also from conservative political ideology. Among these examples, more variance in ratings was explained for "worst" than "best" cues.

Table 9. Multiple Linear Regression Analyses of Selected "Best" and "Worst" Cues' Inter-group Reliability

	Best Cues			Worst Cues	
	Type of degree	Experience	Replication	Business	Nationality
Female	.06 .15 1.09	.06 .13 1.06	.07 .15 1.25	.00 .01 .06	-.03 -.08 -.53
Age	-.02 -.00 -.37	.09 .01 1.46	.11† .01† 1.85	-.07 -.01 -1.22	-.07 -.01 -1.13
Education (some college-plus)	.02 .04 .31	-.04 -.08 -.65	.04 .08 .67	.05 .12 .79	.01 .02 .13
White	.04 .10 .69	.01 .02 .17	.07 .15 1.21	.11† .30† 1.86	-.05 -.13 -.79
Liberal	.05 .05 .92	.03 .02 .46	.02 .02 .43	-.02 -.03 -.42	-.16** -.18** -2.88
Dispute Interest—Yes	.01 .01 .08	.06 .11 .90	.18** .35** 2.92	-.02 -.05 -.32	-.03 -.07 -.46
Dispute Interest—Maybe	-.01 -.04 -.19	-.00 -.01 -.05	.05 .17 .76	.02 .08 .65	.02 .09 .37
Dark Matter	-.03 -.06 -.38	.01 .01 .11	.03 .05 .40	.01 .03 .21	-.03 -.08 -.45
Nanotechnology	-.00 -.01 -.03	-.08 -.16 -1.18	-.12† -.24† -1.80	.01 .04 .22	.00 .01 .06
Scientific reasoning	-.18** -.08** -2.75	-.17* -.07* -2.48	.11† .04† 1.74	-.33*** -.17*** -5.40	-.29*** -.15*** -4.62
Science mistrust	-.01 -.00 -.12	.00 .00 .02	.05 .01 .79	.00 .00 .04	.08 .01 2.42
Scientific positivism	.18** .04** 2.95	.18** .04** 3.05	.19** .04** 3.28	.15** .04* 2.57	.19*** .05*** 3.42
F, *p*	2.15*** (12,288)	2.01* (12,286)	3.42*** (12,292)	4.66*** (12,282)	6.88*** (12,268)
R^2	.08	.08	.12	.17	.24

Each variable cell reports in order standardized and unstandardized correlation coefficients, and t values. † *p < .10* * *p < .05* ** *p < .01* *** *p < .001 or better*

Discussion

Major Findings and Implications

Ideally, cues laypeople use to determine which group of disputing scientists is more credible would be "objectively reliable at separating relative scientific truth from falsehood, available to laypeople, and effectively usable by them; their unambiguous presence (e.g., one side of the dispute clearly has more experience or did better-quality research) makes attentive laypeople decide that this dispute position is more valid; and would be seen by laypeople as reliable, available, and something they can use effectively.[9]

Rather than "normative value" and "persuasive value" attributes, this paper focused on the last, "expressed value" attribute: what do laypeople think is the relative value of potential relative-credibility cues? Account had to be taken of potentially widespread lay unawareness of scientific disagreements, indifference to their outcomes, unfamiliarity with some or all cues presented, or unfamiliarity with suggested evaluative scales. Throughout analysis two controls were used: dispute interest, with the highly interested assumed to be less unaware, indifferent, or unfamiliar, and the degree of straightlining answers (here, the same rating across all five scales for a cue), between chronic straightlining—an indicator of inattention or indifference, to be omitted—and modest straightlining, which might reflect inability to distinguish scale dimensions for a few cues in a novel task. Only a quarter of respondents did no straightlining, a sample size reducing analyses' reliability without guaranteeing greater validity. So analyses here mainly focused on a group straightlining for no more than five of the 22 cues, half of whom straightlined only once or not at all. Reporting results for this Minimals group provided a larger sample, with potentially more reliable results.

These Americans reported the most reliable relative-validity cues for disputing groups of scientists included research quality (e.g., replication), experience and degree type, they used these cues, and (usually) their sources of dispute information contained them. Expressed valuations of cues average Americans can use differed slightly, primarily adding credentials (degree source, awards). The worst cues were deemed interests (e.g., whether scientists or their employers could gain money or prestige) and demographics (whether disputants were American or otherwise similar to the lay observer). These results answered RQ1 about "best" and "worst" cues regarding expressed value. These results do not assess cues' normative or persuasive values, nor do they cover all possible cues (e.g., omitting local cues). But their reliability within these limits is enhanced by findings, addressing RQ2-3, of surprisingly few differences in cue rankings, implicit or explicit. We found somewhat stronger effects for high- versus low-dispute interest respondents (e.g., in correlations across rankings), implying more coherence in the former's responses (but not necessarily more accuracy: as noted earlier, social desirability motivations that might increase inter-rank correlations might be stronger among high-DI respondents, but also might be more accurate if, for example, these cues are both available and used by them). Dimensions reflected in the five rating scales tended to correlate as expected (e.g., perceived ability of the average American least associated with the other criteria), approaches to collating explicit rankings did not yield markedly different results, and explicit rankings largely correlated with inter-group reliability implicit rankings, expected if people complied with instructions. Main variations in multiple regression analyses of inter-group reliability ratings for five selected "best" and "worst" cues were between people high and low in knowledge of scientific reasoning, and those high and low in beliefs in scientific positivism, but those relationships did not vary across these cues.

This reassuring convergence of rankings still leaves questions about expressed value. For example, people rated research quality cues highly, among technical information purportedly least accessible to lay understanding (one

"normative value" criterion; Collins & Evans, 2007; Flanagin & Metzger, 2007; Goldman, 2001; Hardwig, 1985). Laypeople may indeed fail to read research literature and accurately determine the relative degree of study replication between disputing scientists. That does not preclude their using an accurate report of replication in their information sources, *if* those sources can access and are willing to report accurate accounts. Meanwhile, people rated interest cues—often deemed vital in qualitative research ("follow the money") and reasons laypeople think scientific disputes occur (Dieckmann et al., 2017; Dieckmann & Johnson, 2019; Johnson & Dieckmann, 2018; Kajanne & Pirttilä-Backman, 1999; Maxim & Mansier, 2014; Sprecker, 2002)—among the least reliable. The source of this discrepancy merits study: e.g., did this larger quantitative sample tend to believe both sides would have interests in its outcome, thus reducing such cues' discriminative value?

The non-representative sample used here, with its above-average education, may have positively affected observed straightlining and results. If chronic (high-frequency) straightlining reflected topical indifference, that would likely increase in an average-education sample, reducing the power of a given sample size to identify clear implicit rankings across cues. If the more moderate straightlining on which reported results are based was due to inability to distinguish among the rating criteria for a given cue in this sample, using a representative sample would only yield greater convergence across implicit rankings. Probing reasons for poor within-cue discrimination may be important: e.g., does difficulty in definitively recalling whether a cue is available in their information sources generalize to rating its reliability or average-American utility? why? But within-cue straightlining does not affect relative rankings of different cues, as the high education of this sample might have: e.g., more emphasis than a representative sample might offer on the perceived reliability of research quality cues, whether due to greater knowledge or greater motivation to offer socially desirable responses.

The utility and parsimony of the proposed cue taxonomy will be validated only with logical analysis and empirical application by others, but in this expressed-value effort generally people treated cues within a category similarly, a partial vindication. For example, all but the field uncertainty cue among research quality cues ranked high, while interest and demographic cues ranked low; credentials' rankings varied more depending upon the ranker, while only experience among performance cues consistently ranked high.

Assuming these results replicate, we need complementary normative and persuasive value studies (e.g., which cues more influence lay decisions about disputants' relative credibility, and which should?). Intriguingly, experimental tests involving eight of these cues yielded odds ratios for persuasive value—how much more often people select Position B as the more valid position when the cue favors Position B over Position A—across three disputes (dark matter, sea level rise, marijuana risks and benefits) whose rank correlation with their rankings in this expressed-value study was a remarkable .71 (*reference omitted*). This does not validate overall survey rankings, but justifies further probing of expressed, normative and persuasive values of cues.

Science communication practitioners may consider some current findings, besides assessing whether cues appear in information sources. They might try to influence which cues appear in mass and social media, despite mass media's constrained ability and willingness to report relevant information in science stories, such as uncertainty or the balance of scientific opinion ("vote-counting") (Friedman, Dunwoody, & Rogers, 1999; Singer & Endreny, 1993; Wakefield & Elliott, 2003). Meanwhile, recent guides for lay assessment of scientific studies stress research quality cues without explaining their strength or other cues' drawbacks, or acknowledging that scientists also use other cues which might help laypeople if

available (Alberts & McNutt, 2013; Collins & Weinel, 2011; Harvard School of Public Health, 2016; Sutherland, Spiegelhalter, & Burgman, 2013). These somewhat mutually-reinforcing factors in journalistic and science education practice may limit opportunities to change availability or lay use of normatively-valued cues, but need not prohibit action. Emphasizing cues endorsed or rejected by high-DI citizens would be the most efficient approach, as they are the ones most likely to attend to cues in their information sources.

Study Limitations

This study emphasized expressed valuation, not normative or persuasive valuation; ubiquitous versus local cues; two-sided disputes between masses of scientists; and Americans' views (see civic epistemologies; Jasanoff, 2005). This national sample's relatively high education limits generalization to the U.S. population, and may have evoked clearer but somewhat different cue rankings than a representative sample (e.g., emphasizing research quality cues rather than credentials).

Conclusions

Scholars of expertise, and of citizens' access to information allowing them to distinguish better from worse science, disagree on which cues are better and suitable evidence (e.g., reasonable arguments, versus lay access to tacit knowledge or social evidence; Alberts & McNutt, 2013). Yet scholars agree on both difficulties of lay discrimination and that in some circumstances laypeople can (somewhat) assess expertise. Building upon earlier conceptual discussions, and limited qualitative or documentary data on ubiquitous versus local (Irwin & Wynne, 1996) cues, this study examined which ubiquitous cues people *say* are reliable, available, usable and used, as a complement to ethnographic and experimental studies of cues people *actually* use, plus "normative value" and "persuasive value" studies noted above. The aggregate convergence on research quality, experience and degree types as "best" cues—with relatively few differences based on dispute interest and evaluative dimension (reliability for relative validity versus the average American's ability to use the cue)—provides a foundation for further scholarly effort in improving the cue taxonomy proposed here, replicating these findings and extending them, and advancing science communication about how laypeople can assess intra-scientific disputes. Collectively these studies' aim should be to help scholars understand strengths and limitations of lay decisions about science using such cues, and potentially improve both science communication and policy-making.

Footnotes

[1] Alternatives not covered here occur in who disagrees (e.g., individuals; non-scientist experts; three-plus parties) on what (e.g., a phenomenon's magnitude or frequency; data quality or methods).

[2] Vote-counting—which side does the majority support?—exemplifies a cue applying only to disputing groups of

scientists, not to disputing individual scientists.

[3] Education differed significantly by dispute interest ($F(2,531) = 9.59$, $p < .0005$): high-DI (dispute interest) respondents averaged some college ($M = 5.07$, $SD = 1.32$), versus vocational education for low-DI ($M = 4.63$, $SD = 1.53$, $p = .004$) and ambiguous-DI respondents ($M = 4.35$, $SD = 1.47$, $p = .001$).

[4] No statistically significant contrast met the looser FDR criterion of $d = .05$.

[5] Briefly, the Zero group exhibited no FDR-significant differences in ratings by dispute interest; Zeroes and Minimals diverged slightly on reliability-availability correlations, as the former exhibited moderately strong correlations, with high-DIs reporting reliable cues as available much more often; Zeroes exhibited slightly stronger correlations than Minimals for three scales but much weaker ones for availability and ability.

[6] Supporting Information highlights the top five, but Table 5 focuses on the top three, to parallel the explicit rankings in Table 6.

[7] People were asked separately for their three best and worst cues to minimize burden; complete explicit rankings might have yielded stronger inverse "best"-"worst" correlations.

[8] Logistic regression analyses merit $\geq$ 50 cases per predictor, impossible even with the full sample, precluding testing factors in explicit rankings.

[9] "Objectively" is a loaded word used as a proxy for debates on social construction of scientific practice beyond this paper's scope (Beatty, 2006; Irwin & Wynne, 1996; Stilgoe, 2007).

References

- Alberts, B., & McNutt, M. (2013). Science demystified. *Science, 342,* 289.
- Anderson, E. (2011). Democracy, public policy, and lay assessments of scientific testimony.*Episteme, 8*(2), 144–164.
- Beatty, J. (2006). Masking disagreement among experts. *Episteme, 3*(1), 52–67.
- Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B, 57,*289–300.
- Brewer, S. (1998). Scientific expert testimony and intellectual due process. *Yale Law Journal, 107,* 1535–1681.
- Bubela, T., Nisbet, M. C., Borchelt, R., Brunger, F., Critchley, C., Einsiedel, E., . . . & Caulfield T. (2009). Science communication reconsidered. *Nature Biotechnology, 27,* 514–518.
- Campbell, B. (1985). Uncertainty as symbolic action in disputes among experts. *Social Studies of Science, 15,* 429–453.
- Collingridge, D., & Reeve, C. (1986). *Science speaks to power: The role of experts in policy-making* London, UK: Frances Pinter.
- Collins, H., & Evans, R. (2007). *Rethinking expertise*. Chicago, IL: University of Chicago Press.
- Collins, H., & Weinel, M. (2011). Transmuted expertise: How technical non-experts can assess experts and expertise. *Argumentation, 25,* 401–413.

- Davison, W. P. (1983). The third-person effect in communication. *Public Opinion Quarterly, 47,* 1–15.
- De Melo-Martin, I., & Intemann, K. (2013). Scientific dissent and public policy: Is targeting dissent a reasonable way to protect sound policy decisions? *EMBO Reports, 14*(3), 231–235.
- Dieckmann, N. F., & Johnson, B. B. (2019). Why do scientists disagree? Explaining and improving measures of the perceived causes of scientific disputes. *PLOS ONE, 14*(2), e0211269. doi:10.1371/journal.pone.0211269
- Dieckmann, N. F., Johnson, B. B., Gregory, R., Mayorga, M., Han, P., & Slovic, P. (2017). Public perceptions of expert disagreement: Expert incompetence or a complex and random world? *Public Understanding of Science, 26,* 325–338.
- Drummond, R., & Fischhoff, B. (2017). Development and validation of the scientific reasoning scale. *Journal of Behavioral Decision Making, 30*(1), 26–38.
- Earle, T. C., & Cvetkovich, G. T. (1995). *Social trust: Toward a cosmopolitan society*. Westport, CT: Praeger.
- Epstein, S. (1996). *Impure science: AIDS, activism and the politics of knowledge*. Berkeley, CA: University of California Press.
- Fallis, D., & Frické, M. (2002). Indicators of accuracy of consumer health information on the Internet: A study of indicators relating to information for managing fever in children in the home. *Journal of the American Medical Informatics Association, 9*(1), 73–79.
- Flanagin, A. J., & Metzger, M. J. (2007). The role of site features, user attributes, and information verification behaviors on the perceived credibility of web-based information. *New Media & Society, 9,* 319–342.
- Frické, M., Fallis, D., Jones, M., & Luszko, G. M. (2005). Consumer health information on the Internet about carpal tunnel syndrome: Indicators of accuracy. *The American Journal of Medicine, 118*(2), 168–174.
- Friedman, S. M., Dunwoody, S., & Rogers, C. L. (1999). *Communicating uncertainty: Media coverage of new and controversial science*. New York: Routledge.
- Gauchat, G., O'Brien, T., & Mirosa, O. (2017). The legitimacy of environmental scientists in the public sphere. *Climatic Change, 143,* 297–306.
- Gelfert, A. (2011). Expertise, argumentation, and the end of inquiry. *Argumentation, 25*(3), 297–312.
- Glickman, M. E., Rao, S. R., & Schultz, M. R. (2014). False discovery rate control is a recommended alternative to Bonferroni-type adjustments in health studies. *Journal of Clinical Epidemiology, 67,* 850–857.
- Goldman, A. (2001). Experts: Which ones should you trust? *Philosophy and Phenomenological Research, 63*(1), 85–111.
- Halfon, S. (2006). The disunity of consensus: International policy coordination as socio-technical practice. *Social Studies of Science, 36*(5), 783–807.
- Hardwig, J. (1985). Epistemic dependence. *The Journal of Philosophy, 82*(7), 335–349.
- Hartman, R., Dieckmann, N. F., Stantsy, B., Sprenger, A., & DeMarree, K. (2017). Modeling attitudes toward science: Development and validation of an attitudes toward science scale (ATS). *Basic and Applied Social Psychology, 39,* 358–371.
- Harvard School of Public Health. Deciphering media stories on diet. Downloaded 23 August 2016 from https://www.hsph.harvard.edu/nutritionsource/media/.
- Hayes, A. F. (2018). *Introduction to mediation, moderation, and conditional process analysis: A regression-based*

approach (2d ed.). New York: Guilford Press.

- Hayes, A. F., & Cai, L. (2007). Using heteroskedasticity-consistent standard error estimators in OLS regression: An introduction and software implementation. *Behavior Research Methods, 39,* 709–722.
- Irwin, A., & Wynne, B. (Eds.). (1996). *Misunderstanding science? The public reconstruction of science and technology*. New York, NY: Cambridge University Press.
- Jasanoff, S. (2005). *Designs on nature: Science and democracy in Europe and the United States* Princeton, NJ: Princeton University Press.
- Jasanoff, S., & Wynne, B. (1998). Science and decision-making. In S. Rayner & E. Malone (Eds.),*Human choice and climate change: Vol. 1: The societal framework* (pp. 1–88). Columbus, OH: Battelle Press.
- Johnson, B. B. (2003). Further notes on public response to uncertainty in risks and science.*Risk Analysis, 23,* 781–789.
- Johnson, B. B., & Dieckmann, N. F. (2018). Lay Americans' views of why scientists disagree with each other.*Public Understanding of Science, 27,* 824–835.
- Johnson, B. B., & Slovic, P. (1995). Explaining uncertainty in health risk assessment: Initial studies of its effects on risk perception and trust. *Risk Analysis, 15,*85–94.
- Johnson, B. B., & Slovic, P. (1998). Lay views on uncertainty in environmental health risk assessment.*Journal of Risk Research, 1,* 261–279.
- Kajanne, A., & Pirttilä-Backman, A.-M. (1999). Laypeople's viewpoints about the reasons for expert controversy regarding food additives. *Public Understanding of Science, 8,* 303–315.
- Kuklinski, J. H., Metlay, D. S., & Kay, W. D. (1982). Citizen knowledge and choices on the complex issue of nuclear energy. *American Journal of Political Science, 11,*615–642.
- Kutrovátz, G. (2010). Trust in experts: Contextual patterns of warranted epistemic dependence.*Balkan Journal of Philosophy, 2*(1), 57–68.
- Ledford, C. J. W., Willett, K. L., & Kreps, G. L. (2012). Communicating immunization science: The genesis and evolution of the national network for immunization information. *Journal of Health Communication, 17,*105–122.
- Logan, R. A. (2001). Science mass communication: Its conceptual history.*Science Communication, 23,* 135–163.
- Lysaght, T., & Kerridge, I. (2012). Rhetoric, power and legitimacy: A critical analysis of the public policy disputes surrounding stem cell research in Australia (2005–6). *Public Understanding of Science, 21,* 195–210.
- Matheson, D. (2005). Conflicting experts and dialectical performance: Adjudication heuristics for the layperson. *Argumentation, 19,* 145–158.
- Maxim, L., & Mansier, P. (2014). How is scientific credibility affected by communicating uncertainty? The case of endocrine disrupter effects on male fertility. *Human and Ecological Risk Assessment, 20,* 201–223.
- Miller, B. (2016). Scientific consensus and expert testimony in courts: Lessons from the Bendectin litigation. *Foundations of Science, 21*(1), 15–33.
- Rabinovich, A., & Morton, T. A. (2012). Unquestioned answers or unanswered questions: Beliefs about science guide responses to uncertainty in climate change in risk communication. *Risk Analysis, 32,* 992–1002.
- Rose, K. M., Howell, E. L., Su, L. Y.-F., Xenos, M. A., Brossard, D., & Scheufele, D. A. (2019). Distinguishing scientific

knowledge: The impact of different measures of knowledge on genetically modified food attitudes. *Public Understanding of Science.* Advance online publication. doi: 10.1177/0963662518824837

- Singer, E., & Endreny, P. M. (1993). *Reporting on risk: How the mass media portray accidents, diseases, disasters, and other hazards*. New York: Russell Sage Foundation.
- Slater, M. D., & Rouner, D. (1996). How message evaluation and source attributes may influence credibility assessment and belief change. *Journalism & Mass Communication Quarterly, 73,* 974–991.
- Solomon, M. (2007). The social epistemology of NIH consensus conferences. In H. Kincaid & J. McKitrick (Eds.), *Establishing medical reality: Essays in the metaphysics and epistemology of biomedical science* (pp. 167–177). Dordrecht, The Netherlands: Springer.
- Sprecker, K. (2002). How involvement, citation style, and funding source affect the credibility of university scientists. *Science Communication, 24,* 72–97.
- Steel, B., List, P., Lach, D., & Shindler B. (2004). The role of scientists in the environmental policy process: A case study from the American West. *Environmental Science and Policy, 7,* 1–13.
- Stilgoe, J. (2007). The (co-)production of public uncertainty: UK scientific advice on mobile phone health risks. *Public Understanding of Science, 16,* 45–61.
- Sutherland, W. J., Spiegelhalter, D., & Burgman, M. A. (2013). Twenty tips for interpreting scientific claims. *Nature, 503,* 335–337.
- Svenson, O. (1981). Are we all less risky and more skillful than our fellow drivers? *Acta Psychologica, 47,* 143–148.
- Thomm, E., & Bromme, R. (2016). How source information shapes lay interpretations of science conflicts: Interplay between sourcing, conflict explanation, source evaluation, and claim evaluation. *Reading & Writing, 29,* 1629–1652.
- U.S. Census. *American Fact Finder*. Washington, D.C.: U.S. Department of Commerce. Downloaded 25 February 2016 from http://factfinder.census.gov/faces/nav/jsf/pages/index.xhtml.
- Van der Linden, S. L., Clarke, C. E., & Maibach, E. W. (2015). Highlighting consensus among medical scientists increases public support for vaccines: Evidence from a randomized experiment. *BMC Public Health, 15*(1), 1207.
- Van der Linden, S. L., Leiserowitz, A. A., Feinberg, G. D., & Maibach, E. W. (2015). The scientific consensus on climate change as a gateway belief: Experimental evidence. *PLOS ONE, 10*(2), e0118489.
- Wagenknecht, S. (2015). Facing the incompleteness of epistemic trust: Managing dependence in scientific practice. *Social Epistemology, 29*(2), 160–184.
- Wakefield, S. E. L., & Elliott, S. J. (2003). Constructing the news: The role of local newspapers in environmental risk communication. *The Professional Geographer, 55,* 216–226.
- Weinstein, N. D., & Sandman, P. M. (1993). Some criteria for evaluating risk messages. *Risk Analysis, 13,* 103–114.
- Westerman, D., Spence, P. R., & Van Der Heide, B. (2014). Social media as information source: Recency of updates and credibility of information. *Journal of Computer-Mediated Communication, 19,* 171–183.
- White, M. P., & Johnson, B. B. (2010). The Intuitive Detection Theorist (IDT) model of trust in hazard managers. *Risk Analysis, 30,* 1196–1209.
- Williams, A. E. (2012). Trust or bust: Questioning the relationship between media trust and news attention. *Journal of Broadcasting and Electronic Media, 56,* 116–131.

- Wilson, E. J., & Sherrell, D. L. (1993). Source effects in communication and persuasion research: A meta-analysis of effect size. *Journal of the Academy of Marketing Science, 21*,101–112.
- Wynne, B. (1989). Sheepfarming after Chernobyl: A case study in communicating scientific information.*Environment: Science and Policy for Sustainable Development, 31,* 10–15, 33–39.
- Yearley, S. (1994). Understanding science from the perspective of the sociology of scientific knowledge: An overview. *Public Understanding of Science, 3,* 245–258.
- Zehr, S. (2000). Public representations of scientific uncertainty about global climate change.*Public Understanding of Science, 9,* 85–103.