Bayesian confirmation and commonsense notions of evidential strength

Derek Powell (dmpowell@asu.edu)
School of Social and Behavioral Sciences
4701 W. Thunderbird Rd. Glendale, AZ 85306 USA

Shyam Nair (gsnair@asu.edu)
School of Historical, Philosophical, and Religious Studies
975 S. Myrtle Ave
P.O. Box 874302
Tempe, AZ 85287 USA

Abstract

How can we quantify the degree to which a piece of evidence affects a person's belief? Philosophers investigating theories of *Bayesian Confirmation* have identified a plurality of potential measures, each with their own virtues and shortcomings. Psychologists meanwhile have largely neglected this question, which has limited their ability to understand *differential belief updating*, cases where certain individuals or groups respond to the same evidence in different ways. In this study, we examine how competing Bayesian confirmation measures track commonsense notions of evidential strength. We demonstrate how these measures can be computed from participants' belief reports, and identify cases where the measures come apart in their characterization of participants' belief updating. In so doing, this project seeks to build connections between investigations of psychological belief updating processes and formal epistemic theories of confirmation.

Keywords: Belief updating; Bayesian confirmation

Introduction

Suppose across some shared moment, a child pushes a cup from a table and watches it fall to the floor, a physics student watches a counterintuitive demonstration, a scientist compares the fit of two statistical models, and a philosopher weighs the force of an argument. When each observes their evidence, how much do those observations change their beliefs? How can this change be quantified, in principle and in practice?

On the principled front, epistemologists investigated methods for quantifying how evidence impacts beliefs by considering different theories of *Bayesian Confirmation*. These theorists have sought to define measures of the degree to which some evidence E confirms a hypothesis H relative to background knowledge K. Within this literature there is no strong consensus as to a particular "best" measure, as several competing measures are recognized as offering different virtues.

On the practical front, psychologists are often interested in testing how beliefs change given different forms of evidence. Most investigations focus on testing for the existence of belief effects: does a piece of evidence produce a change in peoples' beliefs? But other investigations seek to uncover more nuanced patterns of belief updating. In particular, many insights stand to be gained in examinations of *differential belief updating*, cases where certain individuals or groups respond to the same evidence in different ways. For example: Do Democrats and Republicans respond to scientific arguments in the same way? Are emotional appeals persuasive, and for whom? Do people more readily update their beliefs for good news than for bad?

Examinations of differential responding will generally need to quantify degrees of belief change in a way that commits to a particular measure that quantifies that degree of confirmation.[1] As mentioned, the theoretical literature has proposed many such measures, each with their own virtues. Though rightly constrained by epistemic principles, a psychological theory of confirmation has its own priorities and desiderata that might help to narrow the field.

In this paper, we consider some of the most important confirmation measures in the theoretical literature and which of them might best fulfill the desiderata for a psychological theory of evidential confirmation. We then present a study examining how competing Bayesian confirmation measures track commonsense notions of evidential strength. We demonstrate how these measures can be computed from participants' belief reports, and identify cases where the measures come apart in their characterization of participants' belief updating.

Confirmation measures

Theorists have proposed dozens of different incremental confirmation measures. Some popular measures are the following:

- Difference Measure (Carnap, 1962; Earman, 1992):

$$d(H,E) = P(H \mid E) - P(H)$$

- Log Ratio Measure (Milne, 1996):

$$r(H,E) = \log\left(\frac{P(H \mid E)}{P(H)}\right)$$

- Log likelihood Measure (Fitelson, 1999; Good, 1983):

$$l(H,E) = \log\left(\frac{P(E \mid H)}{P(E \mid \neg H)}\right)$$

- Z-measure (Crupi, Tentori, & Gonzalez, 2007):

$$z(H,E) = \begin{cases} \frac{P(H|E)-P(H)}{1-P(H)} & P(H \mid E) \geq P(H) \\ \frac{P(H|E)-P(H)}{P(H)} & P(H \mid E) < P(H) \end{cases}$$

[1] One exception could be extreme cases like "backfire," where evidence has the opposite of its intended effect for some subset of people. Such cases are likely quite rare (Wood & Porter, 2019), though they are at least theoretically possible (Jern et al., 2014).

- Normalized Difference Measure (Christensen, 1999; Joyce, 1999):

$$s(H,E) = P(H \mid E) - P(H \mid \neg E)$$

These measures all have attractive features and agree in many cases. Still, there are important (ordinal) differences between them, and each has been subject to serious criticism. Here, we highlight three main areas of theoretical concern.

To begin, Eells and Fitelson (2002) use symmetries related to confirmation to evaluate measures. Taking **c** to be an arbitrary confirmation measure, they consider the following symmetry conditions:

- Evidence Symmetry (ES): $\mathbf{c}(H,E) = -\mathbf{c}(H,\neg E)$
- Communitivity Symmetry (CS): $\mathbf{c}(H,E) = \mathbf{c}(E,H)$
- Hypothesis Symmetry (HS): $\mathbf{c}(H,E) = -\mathbf{c}(\neg H,E)$

Eells and Fitelson argue ES and CS are not generally valid or desirable while HS is. To see this, they consider an example in which a card is drawn from a fair deck. Let H be the claim that the card is black and E be the claim that the card is a seven of spades. That the card is a seven of spades (E) is decisive evidence that the card is black (H). But that the card is not a seven of spades ($\neg E$) is not decisive evidence against the claim that the card is black (H). This falsifies ES, which claims that the degree to which E confirms H is equal to the degree to which $\neg E$ disconfirms.

Additionally, that the card is black (H) is relatively weak evidence that it is a seven of spades (E). This falsifies CS which claim the degree to which E confirms H is the degree to which H confirms E.

On the other hand, the card being a seven of spades (E) maximally disconfirms that claim that the card is not black ($\neg H$). So this suggests HS is true. Indeed, other authors (Tentori, Crupi, Bonini, & Osherson, 2007) concur with Eells and Fitelson that HS is not subject to counterexample.

Based on this, Eells and Fitelson conclude the correct confirmation measure will validate HS, but not ES or CS. They show that d and l do this. But r fails to validate HS and incorrectly validates CS. And s incorrectly validates ES and CS and fails to validate HS. More recently, (Crupi et al., 2007) have shown z correctly validates HS, but not ES or CS.

Crupi, Tentori, and Gonzales (2007) have argued that there is a more general family of desirable symmetry conditions. This larger family includes the condition HS as one of its members, but it also includes some more subtle conditions such as:

$$\text{if } P(H \mid E) < P(H),\, \mathbf{c}(H,E) = \mathbf{c}(E,H)$$

The argument in favor of this more general family of symmetries is sufficiently complex so as to be beyond the scope of this paper. Nevertheless, what is important is that the authors show that only z satisfies these more general symmetries.

Next, Tentori, Crupi, Bonini, and Oshershon (2007) present problems for d and l related to whether they are natural generalizations of logical entailment. If E entails H, then plausibly confirmation for H by E is maximal. But this does not hold for d. To see this, let H_1 be the claim that the card is a seven, H_2 be the claim that the card is a spade, and E be the claim that the card is a seven of spades. Here E provides maximal support for both H_1 and H_2, but $d(H_1,E) > d(H_2,E)$ due to the different prior expectations $P(H_1)$ and $P(H_2)$.

l also faces trouble with generalizing logical entailment. This is because it is undefined in cases where E entails H or entails $\neg H$ (because it involves division by 0 in the first case and $log(0)$ in the second case).

Finally, Fitelson (2001) observes that it is desirable for confirmation measures to have an additive feature in cases where pieces of evidence are independent. To state Fitelson's claim, it helps to introduce some terminology. Given a probability distribution P, let P_E be defined as $P_E(\cdot) = P(\cdot \mid E)$. And given a confirmation measure **c** defined by P, let $\mathbf{c}_E$ be exactly like **c** except defined by P_E. Finally let us say:

> E_1 and E_2 are *confirmationally independent regarding H according to* **c** exactly if $\mathbf{c}(H,E_1) = \mathbf{c}_{E_2}(H,E_1)$ and $\mathbf{c}(H,E_2) = \mathbf{c}_{E_1}(H,E_2)$

When E_1 and E_2 are confirmationally independent regarding H, the confirmation measure says that updating on one piece of evidence makes no difference to the degree of confirmation provided to H by the other piece of evidence.

Fitelson claims that the following is desideratum for a confirmation measure:

> If E_1 and E_2 are confirmationally independent regarding H according to **c**, then $\mathbf{c}(H,E_1 \wedge E_2) = \mathbf{c}(H,E_1) + \mathbf{c}(H,E_2)$

It can be shown that d, r, and l satisfy this condition. But it can be shown that nothing like this holds for s.

z also does not satisfy this condition. However, it can be shown for z, that the relevant conditions holds wherever E_1 and E_2 both individually confirm H or E_1 and E_2 both individually disconfirm H. But in cases where E_1 and E_2 "point in different directions", the additivity condition does not hold (Fitelson, 2021).

Overall, criteria related to symmetries, generalizing of entailment, and additivity of independent evidence provide support for certain measures and count against others. We believe that l and z fare best based on these principled considerations, though the arguments are not entirely decisive.

Desiderata for a psychological theory of confirmation

The foregoing principled considerations define desiderata for an epistemological confirmation measure. What then for a psychological theory of confirmation? A psychological theory ought to be based upon these epistemological principles, but has its own considerations that will lead them to be weighted differently. We will argue that l provides the most satisfying criteria for a psychological theory of confirmation.

First, psychological studies using self reports and examining naturalistic sources of evidence (i.e. beyond the realm of balls and urns) can derive measures for prior and posterior beliefs for some hypothesis H, but it is typically not obvious how to identify other evidential or intermediate belief values, such as $P(H|\neg E)$. Thus, it is more or less a prerequisite that the measures be estimable only from these measures (obviating measures like s). d, r, and z are quite obvious to estimate from their definitions. l can also be directly estimated from prior and posterior belief reports, leveraging the log-odds form of Bayes' Rule:

$$\log O(H|E) = \log O(H) + \log\left(\frac{P(E \mid H)}{P(E \mid \neg H)}\right)$$

So that l can be estimated as:

$$l(H,E) = \log O(H|E) - \log O(H)$$

Second, certain principled desiderata seem less important for a psychological theory. One criticism against l (and d) is that it does not naturally generalize logical entailment. However, this concern is of substantially less importance for psychological theory, where we might more reasonably hypothesize there is a discontinuity between inductive and deductive modes of reasoning.[2]

Finally, a psychological theory of confirmation should serve a meaningful role in the larger program of Bayesian cognitive science and psychology. Bayesian cognitive scientists have argued that much of higher-order reasoning is subserved by generative probabilistic mental models (Battaglia, Hamrick, & Tenenbaum, 2013; Chater et al., 2020; Tenenbaum, Kemp, Griffiths, & Goodman, 2011). These mental models represent people's understanding of a domain, allowing them to make inferences and predictions and to reason about new evidence. For psychologists working in this paradigm, describing these mental models is a chief concern and examining differential belief updating is one potentially powerful lens through which they might be better understood. In some cases we might be interested in whether certain individuals update their beliefs more or less rationally, e.g. by examining whether motivational factors influence the degree to which people revise their beliefs (e.g. Hahn & Harris, 2014; Möbius, Niederle, Niehaus, & Rosenblat, 2022; Powell, 2022; Sharot, Korn, & Dolan, 2011). In other cases, we might use differential belief updating to test whether people have different intuitive theories or auxiliary beliefs regarding the evidence (Gershman, 2019; Jern, Chang, & Kemp, 2014). Identifying such cases could offer important clues to the larger mental models people use to reason about evidence in important domains such as vaccination decisions (e.g. Powell, Weisman, & Markman, 2023), climate change (e.g. Cook & Lewandowsky, 2016; Schotsch & Powell, 2022), and other major issues.

First, evaluations of the rationality of human belief updating must be made with respect to some normative standard. Typically, these concern simple cases where the observed evidence has a known impact, described by Bayes' Rule in terms of the likelihood of the data given the hypotheses in contention (e.g. Edwards, 1968). A question is whether people update their beliefs according to this likelihood. Here, comparisons according to l will typically offer the most direct answer, as this measure is derived entirely from the likelihoods.

Second, psychologists investigating people's mental models of a domain, or otherwise seeking to compare human behavior against a rational Bayesian standard, ought to be especially interested in examining when mental models or auxiliary beliefs specify different theories of the evidence (Gershman, 2019).

Consider Alice and Bob, a doctor and patient awaiting the results of a diagnostic test for an uncommon medical condition (H). The patient, Bob, is quite nervous, and holds a prior belief that he has the condition $P_B(H) = .50$. In contrast, Alice knows that the condition is really quite rare and so holds a much more skeptical prior, $P_A(H) = .10$. However, both doctor and patient agree perfectly in their understanding of the diagnostic test, understanding it to have specificity and sensitivity of .90, i.e. $P(E|H) = .90$ and $P(E|\neg H) = .10$. Fortunately for Bob, the test returns negative. Both Alice and Bob revise their beliefs rationally according to Bayes' Rule. Bob's posterior belief he has the condition is now .10 and Alice's is approximately .01.

Psychologically, we might ask whether Alice and Bob responded to the evidence in different ways? That is, we might ask for a comparison of their confirmation under some measure, confirmation simply being the degree to which a belief was updated by evidence. Within the Bayesian program however, this psychological question might also be seen as asking whether Alice and Bob differ in their mental models of the evidence ($\mathcal{M}$), i.e. whether their likelihood distribution $P_{\mathcal{M}_a}(E|H) = P_{\mathcal{M}_b}(E|H)$ (where here we use E and H to indicate R.V.). Comparisons of l will most directly answer this psychological question.

[2] It is also worth noting that the failures of l to generalize logical entailment can be quite naturally alleviated by defining l in terms of its limits of $+\infty$ or $-\infty$ wherever it would otherwise be undefined.

Branden Fitelson suggests to us a more principled way of achieving the same result that is based on a correction of an observation made by I.J. Good (1975). It is known that the following measure is ordinally equivalent to l:

$$K(H,E) = \frac{Pr(E \mid H) - Pr(E \mid \neg H)}{Pr(E \mid H) + Pr(E \mid \neg H)}$$

Further, it is known that putting the extreme cases aside:

$$l(H,E) = 2 \times \text{ArcTanh}\,(K(H,E))$$

But since according to the standard definition $\text{ArcTanh}(1) = +\infty$ and $\text{ArcTanh}(-1) = -\infty$. We may take the above to be a fully general new definition of l that is defined in the extreme cases.

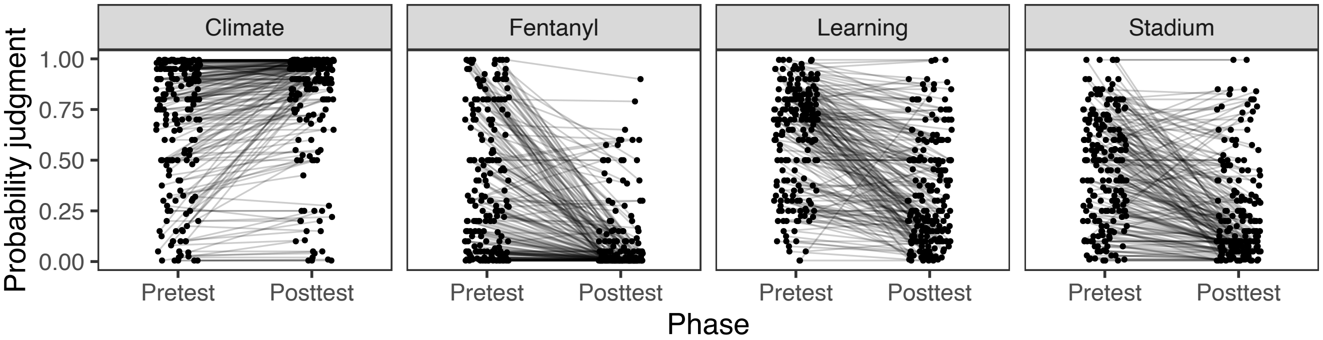

Figure 1: Beliefs before and after exposure to evidence across topics for all participants. Each participants' pretest and posttest belief reports are represented with two points connected by a line.

The present study

To demonstrate how different confirmation measures can come apart and therefore the need to commit to such a measure to make meaningful claims about differential belief updating, we examined how these measures map on to commonsense notions of evidential strength. Just as people readily deploy commonsense or folk-psychological notions of psychological concepts like "belief" or "desire", they are likewise perfectly willing to discuss the persuasiveness of different sources of evidence. By measuring how people's beliefs changed in light of several pieces of evidence, and how those individuals rated the persuasiveness of that evidence, we sought to identify the measure of confirmation to which these ordinary language ascriptions best correspond.

To our knowledge, research from Crupi, Tentori and colleagues (Crupi et al., 2007; Tentori et al., 2007) has produced the only empirical data addressing this type of question previously. Using an abstract "urns" task, they compared participant's probability judgments and their ratings of evidential impact following their observations of ball draws. Translating participants' probability judgments into various measures of confirmation. Analyses of their data found that l and z (Crupi et al., 2007) were the measures most strongly correlated with participants' impact judgments, with some evidence that z provided the strongest correlation.

Our study explores this question in a naturalistic rather than artificial context, exploring these relationships among beliefs and evidence related to consequential real-world domains.[3]

Methods

Participants

A total of 217 participants were recruited from Amazon Mechanical Turk (mTurk) through CloudResearch. These participants were all from the U.S. and were at least 18 years of age. Participants who failed a basic attention check question (17) were excluded from analysis. The final sample in the analyses reported below was 200 (80 female, 118 male, median age 36 years).

Materials and procedures

Four brief educational vignettes were created to correct common misconceptions about four different topics. The four topics were 1) the anthropogenic nature of climate change, 2) the dangers of skin contact with Fentanyl, 3) the effectiveness of education tailored to individual "learning styles," and 4) the economic impacts of major sports stadium construction.

Each of these vignettes were designed to provide evidence for or against a more specific belief. Participants were asked to report their beliefs in terms of probabilities. For instance, evidence about the economic impacts of sports stadiums was paired with a question asking participants to judge the probability that a new sports stadium being built in Buffalo would generate enough tax revenue to pay a return on the public investment.

Participants were asked about the topics in a randomized order. For each topic, participants made an initial probability judgment (pretest), then read the brief educational intervention, and then were asked to make a second probability judgment in light of what they had read (posttest). Then, in a second phase of the study, participants were asked to rate how persuasive they had found the evidence to be. These ratings were made on a Likert scale from "Not at all persuasive" to "Extremely persuasive".

Results

Figure 1 shows participants' probability judgments before and after reading information about each of the four topics. As intended, each piece of evidence had a substantial impact

[3] All materials and analysis code are available at `https://github.com/derekpowell/bayes-conf`.

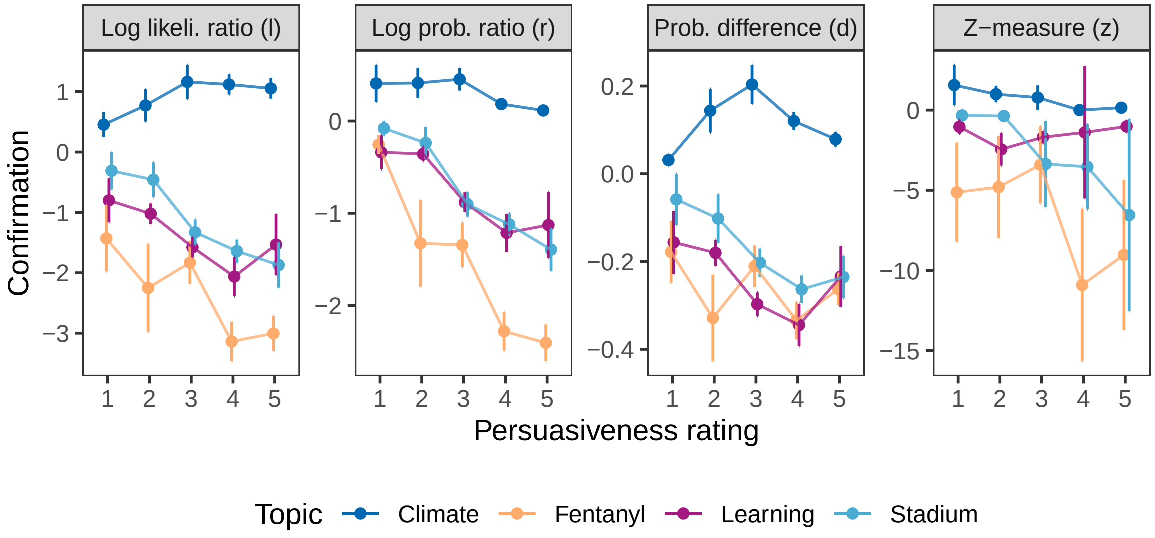

Figure 2: Average confirmation across persuasiveness ratings for each confirmation measure (facets) and topic (line colors). Error bars represent one standard error. Persuasiveness ratings concern confirmatory evidence for the climate topic and disconfirmatory evidence for all others. Thus, a well-behaved measure would be indicated by a monotonically increasing trend for climate and monotonically decreasing trends for all other topics.

on beliefs, though the magnitude of this effect appears to vary across topics.

Several different measures can be used to quantify the degree to which participants' beliefs were affected. We focus on four indices that have been advocated for in the literature on Bayesian Confirmation measures and that can be computed from measures of belief before and after observation of evidence: Probability differences (d), log probability ratios (r), log likelihood ratios (l), and the Z-measure (z).

For each participant, each of these belief updating measures was computed for each topic. Figure 2 shows the average of these measures against participants' ratings of the persuasiveness of the evidence. By comparing different measures of participant's belief updating against their persuasiveness ratings, we can probe which confirmation measure best corresponds with common-sense notions of evidential strength.

Given that our persuasiveness scale provides a face-valid ordinal measure of persuasiveness, we should expect to observe a monotonic association between confirmation and persuasiveness ratings. Particularly damning for such a correspondence would be *population-level reversals*, cases where average confirmation values reliably reversed their ordering across levels of the ordinal scale. As shown in Figure 2, this clearly occurs for d, r, and z measures in the case of the evidence about climate change. Wilcoxon rank sum tests comparing measures for confirmation ratings at the midpoint (3) versus the high endpoint (5) reveal these differences are reliable (all $Ps < .05$)—in each case the degree of confirmation among participants rating the evidence "Extremely Persuasive" was significantly weaker than for those rating it "Moderately Persuasive". In contrast, l shows no reliable reversals for any of the topics (all $Ps > .20$). Although values of l do appear to plateau across persuasiveness levels for the climate change topic, monotonicty is not violated.

It is likely that the reversals observed for d, r, and z in the case of climate change beliefs owe to the strong correlation between persuasiveness ratings and prior beliefs (Figure 3). l is unaffected by this correlation as it is independent of prior beliefs $p(H|K)$ (though it may be somewhat affected by rounding or other measurement error, especially near the bounds of the probability scale). Since z is scaled by the prior we might have expected it to be been less affected by this correlation than d, yet it was nevertheless observed to exhibit reversals.

Discussion

Of the four Bayesian confirmation measures we examined, we found that l best tracked common sense notions of evidential strength. We assessed the viability of each of these measures by examining how they correspond with ratings of the persuasiveness of evidence. All three other measures, d, r, and z failed to consistently track persuasive ratings in a monotonic fashion. Instead, each of these measures exhibited at least one population-level reversal, where higher persuasiveness ratings were associated with reliably lower values on the confirmation measure.

Prior findings by Crupi, Tentori, and colleagues found l and z to best track human judgments of evidential impact (Crupi et al., 2007; Tentori et al., 2007) . These researchers examined the correlation between impact judgments and belief updates across many instances of evidence in a simple ball-and-urn

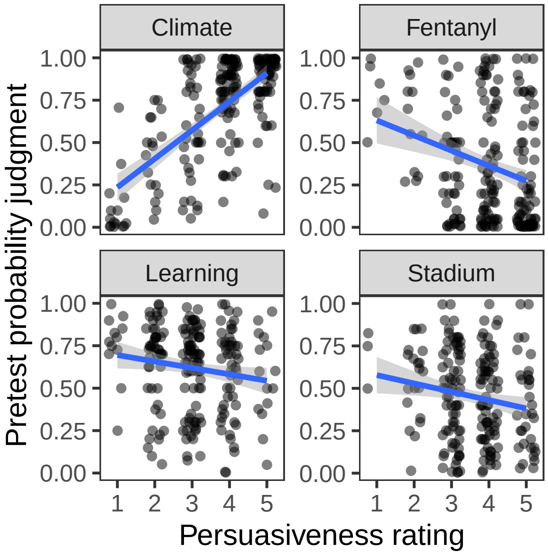

Figure 3: Scatterplots showing correlations between prior beliefs and persuasiveness ratings. To improve readability, points have been jittered along their persuasiveness ratings.

task, finding that l and z better correlated than did other measures like d and r. Our work both expands upon and refine prior findings. First, our findings extend the psychological investigation of confirmation measures to more naturalistic contexts, concerning the sort of evidence that might influence real-world decisions. Second, our findings provide empirical support for l exclusively, while more sharply demonstrating the shortcomings of d, r, and z. Our findings reveal clear failures of these measures, showing that they violate basic measurement constraints on the relationship between persuasiveness judgments and belief updating.

How do these empirical findings relate back to the two theoretical projects we considered at the outset, epistemological and psychological theories of confirmation? Although these theoretical concerns motivate our interest in this empirical psychological question, we see this empirical project as largely descriptive. We saw that l fares well according to a variety of principled epistemological criteria and that there are often independent reasons for psychologists to prefer l as a measure of confirmation. Our finding that l also comports with common sense notion does not address any outstanding theoretical criticisms, though we do see it as another virtue in its favor.

Finally, we consider some potential limitations and directions for future explorations. First, there are some issues related to our measure of persuasiveness that may warrant further investigation. We measured participants' assessments of evidential strength by asking them to rate how "persuasive" the evidence was. However, there could be some concerns about how participants answer this question—whether they do so as immediately (i.e., how persuasive was it to them in this instance?) or broadly (e.g. how persuasive would it be to some one else?) or counterfactually (e.g. if you were just hearing this for the first time, how persuasive would it be?).

These interpretations may raise questions related to the philosophical problem of old evidence. The problem arises in cases where one already regards $P(E)$ to be close to 1 but nonetheless sees E as important evidence for H. It may be that some subjects already know about the evidence provided by the vignette, so they can be thought of as having $P(E)$ close to 1—though E might not shift their beliefs in H in this moment, they might still regard it as highly persuasive. To be sure, cases like this are not unique to our experimental context: Glymour (1980) has argued that there are a number of prominent cases like this in the history of science. Many confirmation measures struggle to adequately deal with these kinds of examples (Christensen, 1999; Glymour, 1980), though it has been argued that l fares well with certain aspects of this problems (Eells & Fitelson, 2000).

Second, there are some potentially interesting cases our empirical study has not examined. One set of cases would be strong evidence for highly implausible or against nearly certain claims. For instance, the kind of evidence that could move someone from $P(H) = .01$ to $P(H|E) = .1$. Such a case of evidence would score a relatively large value for l (approximately equivalent to moving from $P(H) = .50$ to $P(H|E) = .90$), but would still leave a reasoner quite skeptical. It is not intuitive to imagine how people would rate such evidence, and these sorts of cases may pose problems for intuitions about l not identified here.

Finally, there are persistent biases in human probability judgments that pose a general challenge to measurement in this arena (Kahneman, 2011; e.g. Kahneman, Slovic, & Tversky, 1982). Two recent theories of probability judgments have explained a host of observed biases by proposing that people's probability judgments are shrunk toward .50 by varying degrees (Costello & Watts, 2014; Zhu, Sanborn, & Chater, 2020). This scaling could potentially induce nonmonotonicity in observed updates despite equivalent "true" updates for the measures l, r, and z, so it is possible the shortcomings of r and z may be partially explained by these biases. Nevertheless, as calculated from observed probability judgments, l provides a measure of confirmation that reliably comports with commonsense notions.

Conclusions

Rigorous investigations into *differential belief updating* demand a psychological theory of confirmation. Drawing on measures of confirmation identified in the epistemological literature, we identified l as the measure that most directly addresses the concerns of psychologists investigating Bayesian and non-Bayesian belief updating. In an experimental study, we demonstrated how this and other measures can be computed from participants' belief reports, and identify cases where the measures come apart in tracking participants' independent assessments of evidence Our findings indicate that, in addition to its theoretical virtues, l is also the measure that best characterizes commonsense notions of belief updating.

References

Battaglia, P. W., Hamrick, J. B., & Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, *110*(45), 18327–18332. http://doi.org/10.1073/pnas.1306572110

Carnap, R. (1962). *Logical foundations of probability* (2nd ed.). Chicago: The University of Chicago Press.

Chater, N., Zhu, J.-Q., Spicer, J., Sundh, J., León-Villagrá, P., & Sanborn, A. (2020). Probabilistic Biases Meet the Bayesian Brain. *Current Directions in Psychological Science*, *29*(5), 506–512. http://doi.org/10.1177/0963721420954801

Christensen, D. (1999). Measuring confirmation. *Journal of Philosophy*, *96*, 437–461.

Cook, J., & Lewandowsky, S. (2016). Rational Irrationality: Modeling Climate Change Belief Polarization Using Bayesian Networks. *Topics in Cognitive Science*, *8*(1), 160–179. http://doi.org/10.1111/tops.12186

Costello, F., & Watts, P. (2014). Surprisingly rational: Probability theory plus noise explains biases in judgment. *Psychological Review*, *121*(3), 463–480. http://doi.org/10.1037/a0037010

Crupi, V., Tentori, K., & Gonzalez, M. (2007). On bayesian measures of evidential support. *Philosophy of Science*, *74*, 229–252.

Earman, J. (1992). *Bayes or bust*. Cambridge: MIT Press.

Edwards, W. (1968). Conservatism in Human Information Processing. In B. Kleinmuntz (Ed.), *Formal representation of human judgment* (pp. 17–52). New York: Wiley.

Eells, E., & Fitelson, B. (2000). Measuring confirmation and evidence. *The Journal of Philosophy*, *97*, 663–672.

Eells, E., & Fitelson, B. (2002). Symmetries and asymmetries in evidential support. *Philosophical Studies*, *107*, 129–142.

Fitelson, B. (1999). The plurality of bayesian measures of confirmation and the problem of measure sensitivity. *Philosophy of Science*, *66*, S362–S378.

Fitelson, B. (2001). A bayesian account of independent evidence with applications. *Philosophy of Science*, *68*, S123–S140.

Fitelson, B. (2021). A problem for confirmation measure z. *Philosophy of Science*, *88*, 726–730.

Gershman, S. J. (2019). How to never be wrong. *Psychonomic Bulletin & Review*, *26*(1), 13–28. http://doi.org/10.3758/s13423-018-1488-8

Glymour, C. (1980). *Theory and evidence*. Princeton: Princeton University Press.

Good, I. J. (1975). Explicativity, corroboration, and the relative odds of hypotheses. *Synthese*, *30*, 39–73.

Good, I. J. (1983). *Good thinking*. Minneapolis: University of Minnesota Press.

Hahn, U., & Harris, A. J. L. (2014). What Does It Mean to be Biased. In *Psychology of Learning and Motivation* (Vol. 61, pp. 41–102). Elsevier. http://doi.org/10.1016/B978-0-12-800283-4.00002-2

Jern, A., Chang, K. K., & Kemp, C. (2014). Belief polarization is not always irrational. *Psychological Review*, *121*(2), 206–224. http://doi.org/10.1037/a0035941

Joyce, J. (1999). *The foundations of causal decision theory*. Cambridge: Cambridge University Press.

Kahneman, D. (2011). *Thinking, Fast and Slow* (1st edition). New York: Farrar, Straus; Giroux.

Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment Under Uncertainty: Heuristics and Biases*. (D. Kahneman, P. Slovic, & A. Tversky, Eds.). Cambridge University Press.

Milne, P. (1996). *Log[p(h/eb)/p(h/b)]* is the one true measure of confirmation. *Philosophy of Science*, *63*, 21–26.

Möbius, M. M., Niederle, M., Niehaus, P., & Rosenblat, T. S. (2022). Managing Self-Confidence: Theory and Experimental Evidence. *Management Science*, *68*(11), 7793–7817. http://doi.org/10.1287/mnsc.2021.4294

Powell, D. (2022). A descriptive bayesian account of optimism in belief revision. In C. Jennifer, A. Perfors, H. Rabagliati, & V. Ramenzoni (Eds.), *Proceedings of the 42nd Annual Conference of the Cognitive Science Society*.

Powell, D., Weisman, K., & Markman, E. M. (2023). Modeling and leveraging intuitive theories to improve vaccine attitudes. *Journal of Experimental Psychology: General*.

Schotsch, B., & Powell, D. (2022). Understanding intuitive theories of climate change. In J. Culbertson, A. Perfors, H. Rabagliati, & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Meeting of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Sharot, T., Korn, C. W., & Dolan, R. J. (2011). How unrealistic optimism is maintained in the face of reality. *Nature Neuroscience*, *14*(11), 1475–1479. http://doi.org/10.1038/nn.2949

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to Grow a Mind: Statistics, Structure, and Abstraction. *Science*, *331*(6022), 1279–1285. http://doi.org/10.1126/science.1192788

Tentori, K., Crupi, V., Bonini, N., & Osherson, D. (2007). Comparision of confirmation measures. *Cognition*, *103*, 107–119.

Zhu, J.-Q., Sanborn, A. N., & Chater, N. (2020). The Bayesian sampler: Generic Bayesian inference causes incoherence in human probability judgments. *Psychological Review*, *127*(5), 719–748. http://doi.org/10.1037/rev0000190