The Plurality of Bayesian Measures of Confirmation

and the Problem of Measure Sensitivity

Branden Fitelson[†‡]

University of Wisconsin–Madison

May 14, 1999

Contemporary Bayesian confirmation theorists measure degree of (incremental) confirmation using a variety of non-equivalent relevance measures. As a result, a great many of the arguments surrounding quantitative Bayesian confirmation theory are implicitly *sensitive to choice of measure of confirmation*. Such arguments are *enthymematic*, since they tacitly presuppose that certain relevance measures should be used (for various purposes) rather than other relevance measures that have been proposed and defended in the philosophical literature. I present a survey of this pervasive class of Bayesian confirmation-theoretic enthymemes, and a brief analysis of some recent attempts to resolve the problem of measure sensitivity.

1 Preliminaries.

1.1 Terminology, Notation, and Basic Assumptions

The present paper is concerned with the degree of incremental confirmation provided by evidential propositions E for hypotheses under test H, given background knowledge K, according to relevance measures of degree of confirmation $\mathfrak{c}$. We say that $\mathfrak{c}$ is a *relevance measure* of degree of confirmation if and only if $\mathfrak{c}$ satisfies the following constraints, in cases where E confirms, disconfirms, or is confirmationally irrelevant to H, given background knowledge K.[1]

$$(\mathcal{R}) \qquad \mathfrak{c}(H,E\mid K)\begin{cases} >0 & \text{if } \Pr(H\mid E\,\&\,K) > \Pr(H\mid K),\\ <0 & \text{if } \Pr(H\mid E\,\&\,K) < \Pr(H\mid K),\\ =0 & \text{if } \Pr(H\mid E\,\&\,K) = \Pr(H\mid K).\end{cases}$$

I will restrict my attention to the following four relevance measures of degree of confirmation: the *difference* measure d, the log-*ratio* measure r, the log-*likelihood ratio* measure l, and Carnap's (1962, §67) relevance measure $\mathfrak{r}$. The three measures d, r, and l are representative of the varieties of quantitative Bayesian confirmation theory that are currently being defended in the philosophical literature.[2] Carnap's measure $\mathfrak{r}$ (which is a very close relative of the difference measure d) is included here to illustrate that even relevance measures which are very closely related to each other can diverge in important and subtle ways. The measures d, r, l, and $\mathfrak{r}$ are defined as follows.[3]

$$d(H,E\mid K) =_{df} \Pr(H\mid E\,\&\,K) - \Pr(H\mid K)$$

$$r(H,E\mid K) =_{df} \log\left[\frac{\Pr(H\mid E\,\&\,K)}{\Pr(H\mid K)}\right]$$

$$l(H,E\mid K) =_{df} \log\left[\frac{\Pr(E\mid H\,\&\,K)}{\Pr(E\mid \bar{H}\,\&\,K)}\right]$$

$$\begin{aligned}\mathfrak{r}(H,E\mid K) &=_{df} \Pr(H\,\&\,E\,\&\,K)\cdot\Pr(K) - \Pr(H\,\&\,K)\cdot\Pr(E\,\&\,K)\\ &= \Pr(K)\cdot\Pr(E\,\&\,K)\cdot d(H,E\mid K)^{4}\end{aligned}$$

1.2 A General Overview of the Problem

Many arguments surrounding quantitative Bayesian confirmation theory presuppose that the degree to which E incrementally confirms H, given K should be measured using some relevance measure (or, class of relevance measures) $\mathfrak{c}$,

[†]Philosophy Department, University of Wisconsin, 600 N. Park Street, Madison, WI 53706.

[‡]This paper has appeared in *Philosophy of Science* 66: S362–S378. I would like to thank Marty Barrett, Ellery Eells, Malcolm Forster, Ken Harris, Mike Kruse, Elliott Sober, and, especially, Patrick Maher for useful conversations on relevant issues.

[1]I will not defend the *qualitative* Bayesian relevance notion of confirmation here (I will just *assume* it, as an underpinning for the *quantitative* issues I discuss below). Nor will I argue for the existence of a 'rational' probability function Pr of the kind required to give Bayesian confirmation theory its (objective) normative teeth. For a nice recent discussion of many of the controversies surrounding qualitative Bayesian confirmation theory, see Maher 1996.

[2]*Many* relevance measures have been proposed over the years. For a nice survey, see Kyburg 1983. The three relevance measures d, r, and l have had the most loyal following in recent years. Advocates of d include Earman (1992), Eells (1982), Gillies (1986), Jeffrey (1992), and Rosenkrantz (1994). Advocates of r include Horwich (1982), Keynes (1921), Mackie (1969), Milne (1996), and Schlesinger (1995). Advocates of l include Fitelson (1998b), Good (1984), Heckerman (1988), Horvitz and Heckerman (1986), Pearl (1988), and Schum (1994).

[3]Overbars are used to express negations of propositions (i.e. '$\bar{X}$' stands for '*not-X*'). Logarithms (of any arbitrary base greater than 1) of the ratios $\Pr(H\mid E\,\&\,K)/\Pr(H\mid K)$ and $\Pr(E\mid H\,\&\,K)/\Pr(E\mid \bar{H}\,\&\,K)$ are taken to insure that (*i*) r and l satisfy $\mathcal{R}$, and (*ii*) r and l are *additive* in various ways. Not all advocates of r or l adopt this convention (e.g. Horwich (1982)). But, because logarithms are monotonic increasing on $(0,+\infty)$, defining r and l in this way will not result in any loss of (or gain in) generality in my argumentation.

[4]It is perhaps easiest to think of Carnap's $\mathfrak{r}$ as a kind of *covariance* measure. Indeed, when K is *tautologous*, we have: $\mathfrak{r}(H,E\mid K) = \Pr(H\,\&\,E\,\&\,K)\cdot\Pr(K) - \Pr(H\,\&\,K)\cdot\Pr(E\,\&\,K) = \Pr(H\,\&\,E) - \Pr(H)\cdot\Pr(E) = \mathrm{Cov}(H,E)$. In general, $\mathfrak{r}(H,E\mid K) = \Pr(K)^2\cdot\mathrm{Cov}(H,E\mid K)$.

where $\mathfrak{c}$ is taken to have certain quantitative properties. We say that an argument $\mathcal{A}$ of this kind is *sensitive to choice of measure* if $\mathcal{A}$'s validity varies, depending on which of the four relevance measures d, r, l, or $\mathfrak{r}$ is used in $\mathcal{A}$. If $\mathcal{A}$ is valid *regardless* of which of the four relevance measures d, r, l, or $\mathfrak{r}$ is used in $\mathcal{A}$, then $\mathcal{A}$ is said to be *in*sensitive to choice of measure (or, simply, *robust*).[5]

Below, I will show that seven well-known arguments surrounding contemporary Bayesian confirmation theory are sensitive to choice of measure. I will argue that this exposes a weakness in the theoretical foundation of Bayesian confirmation theory which must be shored-up. I call this problem *the problem of measure sensitivity*. After presenting a survey of measure sensitive arguments, I will examine some recent attempts to resolve the measure sensitivity problem. I will argue that, while some progress has been made toward this end, we still do not have an adequate or a complete resolution of the measure sensitivity problem. Specifically, I will show that the many defenders of the difference measure have failed to provide compelling reasons to prefer d over the two alternative measures l and $\mathfrak{r}$. Thus, a pervasive problem of measure sensitivity still remains for many modern advocates and practitioners of Bayesian confirmation theory.

2 Contemporary Examples of the Problem

It is known (although, apparently, not that widely) that *no pair of the four measures d, r, l, and $\mathfrak{r}$ is ordinally equivalent*. That is, each of these four measures can impose *distinct orderings* over sets of hypotheses and collections of evidence.[6] I have not seen many discussions concerning the measure sensitivity of concrete arguments surrounding Bayesian confirmation theory.[7] In this section, I will show that a wide variety of well-known arguments surrounding Bayesian confirmation theory are sensitive to choice of measure.

2.1 Gillies's Rendition of the Popper-Miller Argument

Gillies (1986) reconstructs the infamous argument of Popper and Miller (1983) for the "impossibility of inductive probability" in such a way that it trades essentially on the following *additivity* property of the difference measure d:

(1) $$d(H, E \mid K) = d(H \vee E, E \mid K) + d(H \vee \bar{E}, E \mid K).$$

[5]One can invent more or less stringent varieties of measure sensitivity. For instance, one could call an argument "measure sensitive" (in a *very strict* sense) if $\mathcal{A}$ is valid with respect to *some conceivable* relevance measure $\mathfrak{c}_1$, but invalid with respect to some other *conceivable* relevance measure $\mathfrak{c}_2$. Of course, such a strict notion of sensitivity would probably not be very interesting, since highly gerrymandered relevance measures can undoubtedly be concocted to suit arbitrary purposes. I am employing a much more restrictive notion of measure sensitivity which works only with measures that have actually been used and defended in the literature.

[6]Rosenkrantz (1981, Exercise 3.6) discusses the ordinal non-equivalence of d, r, and l; and Carnap (1962, §67) talks about some important ordinal differences between $\mathfrak{r}$, d, and r (Carnap does not compare $\mathfrak{r}$ with l). See Krantz, Luce, Suppes, and Tversky 1971, Ch. 1 for a theoretical treatment of ordinal equivalence between abstract quantitative measures.

[7]Two notable exceptions are Redhead 1985 and Mortimer 1988, §11.1.

The details of Gillies's Popper-Miller argument are not important here. All that matters for my present purposes is that the additivity property depicted in (1) is required for Gillies's rendition of the Popper-Miller argument against Bayesianism to go through.

Redhead (1985) points out that *not* all Bayesian relevance measures have this requisite additivity property. Specifically, Redhead (1985) notes that the log-ratio measure r does *not* satisfy (1). It follows that the Popper-Miller argument is *sensitive to choice of measure*. Gillies (1986) responds to Redhead's point by showing that the log-ratio measure r is *not* an adequate Bayesian relevance measure of confirmation. Gillies argues that the ratio measure r is inferior to the difference measure d because r fails to cope properly with cases of *deductive evidence* (see § 3.1 for more on this telling argument against r). Unfortunately, however, Gillies fails to recognize that Redhead's criticism of the Popper-Miller argument can be significantly strengthened *via* the following theorem (see the Appendix for proofs of all Theorems):

Theorem 1. *l does not have the additivity property expressed in* (1).[8]

Moreover, as we will see below in § 3.1, the log-likelihood ratio measure l is immune to Gillies's criticism of r. So, pending some good reason to prefer d over l, Gillies's reconstruction of the Popper-Miller argument does not seem to pose a serious threat to Bayesian confirmation theory (charitably reconstructed).

2.2 Rosenkrantz and Earman on "Irrelevant Conjunction"

Rosenkrantz (1994) offers a Bayesian resolution of "the problem of irrelevant conjunction" (a.k.a. "the tacking problem") which trades on the following property of the difference measure d:

(2) If $H \models E$, then $d(H \,\&\, X, E \mid K) = \Pr(X \mid H \,\&\, K) \cdot d(H, E \mid K)$.

I won't bother to get into the details of Rosenkrantz's argument. It suffices, for my present purposes, to note that it depends sensitively on property (2). As a result, Rosenkrantz's argument does *not* go through if one uses r or l, instead of d, to measure degree of confirmation. The proof of the following theorem demonstrates this strong measure sensitivity of Rosenkrantz's approach:

Theorem 2. *Neither r nor l has the property expressed in* (2).[9]

Consequently, Rosenkrantz's account of "irrelevant conjunction" is adequate only if the difference measure d is to be preferred over the two alternative relevance measures r and l. Like Gillies, Rosenkrantz (1981, Exercise 3.6) does provide good reason to prefer d over r (see § 3.1). However, he explicitly admits that he knows of "no compelling considerations that adjudicate between" the difference measure d and the log-likelihood ratio measure l. This leaves Rosenkrantz in a rather uncomfortable position. As I will discuss below, Rosenkrantz is not

[8]Interestingly, Carnap's relevance measure $\mathfrak{r}$ *does* satisfy (1). This follows straightaway from (1), and the fact that $\mathfrak{r}(H, E \mid K) = \Pr(K) \cdot \Pr(E \,\&\, K) \cdot d(H, E \mid K)$.

alone in this respect. I know of no arguments (much less, compelling ones) that have been proposed to demonstrate that d should be preferred over l.

Earman (1992) offers a similar approach to "irrelevant conjunction" which is less sensitive to choice of measure. Earman's approach relies only on the following logically weaker fact about d:

(2′) If $H \models E$, then $d(H \,\&\, X, E \mid K) < d(H, E \mid K)$.

Both the log-likelihood ratio measure l and Carnap's relevance measure $\mathfrak{r}$ satisfy (2′) (proofs omitted); but, the log-ratio measure r does *not* satisfy (2′) (see § 3.1). So, while still sensitive to choice of measure, Earman's "irrelevant conjunction" argument is *less* sensitive to choice of measure than Rosenkrantz's.

2.3 Eells on the Grue Paradox

Eells (1982) offers a resolution of the Grue Paradox which trades on the following property of the difference measure d (where $\beta =_{df} \Pr(H_1 \,\&\, E \mid K) - \Pr(H_2 \,\&\, E \mid K)$, and $\delta =_{df} \Pr(H_1 \,\&\, \bar{E} \mid K) - \Pr(H_2 \,\&\, \bar{E} \mid K)$).

(3) If $\beta > \delta$ and $\Pr(E \mid K) < \frac{1}{2}$, then $d(H_1, E \mid K) > d(H_2, E \mid K)$.

As usual, I will skip over the details of Eells's proposed resolution of Goodman's "new riddle of induction." What is important for our purposes is that (3) is *not* a property of either the log-likelihood ratio measure l or the log-ratio measure r, as is illustrated by the proof of the following theorem:

Theorem 3. *Neither r nor l has the property expressed in* (3).[9]

As a result, Eells's resolution of the Grue Paradox (which is endorsed by Sober (1994)) only works if one assumes that the difference measure d is to be preferred over the log-likelihood ratio measure l and the log-ratio measure r. Eells (pers. comm.) has described a possible reason to prefer d over r (this argument against r is discussed in § 3.2). As far as I know, Eells has offered no argument aimed at showing that d is to be preferred over l.

2.4 Horwich et al. on Ravens and the Variety of Evidence

A great many contemporary Bayesian confirmation theorists (including Horwich (1982)) have offered quantitative resolutions of the Ravens paradox *and/or* the problem of varied (or diverse) evidence which trade on the following relationship between conditional probabilities and relevance measures of confirmation.[10]

(4) If $\Pr(H \mid E_1 \,\&\, K) > \Pr(H \mid E_2 \,\&\, K)$, then $\mathfrak{c}(H, E_1 \mid K) > \mathfrak{c}(H, E_2 \mid K)$.

[9] It is easy to show that both (2) and (3) *do* hold for Carnap's $\mathfrak{r}$ (proofs omitted).

[10] An early quantitative resolution of the Ravens Paradox was given by Hosiasson-Lindenbaum (1940). Hosiasson-Lindenbaum was *not* working within a relevance framework. So, for her, it *was* sufficient to establish that $\Pr(H \mid E_1 \,\&\, K) > \Pr(H \mid E_2 \,\&\, K)$, where E_1 is a black-raven, E_2 is a non-black non-raven, H is the hypothesis that all ravens are black, and K is our background knowledge. Contemporary Bayesian relevance theorists have presupposed that this inequality is sufficient to establish that a black raven *incrementally* confirms that all ravens are black more strongly than a non-black non-raven does. As Theorem 4 shows, this is true for only *some* relevance measures. This same presupposition is also made by Bayesians who argue that (ceteris paribus) more varied sets of evidence (E_1) confirm hypotheses (H) more strongly than less varied sets of evidence (E_2) do. See Earman 1992, 69–79 for a survey of recent Bayesian resolutions of the Ravens Paradox, and Wayne 1995 for a survey of recent Bayesian resolutions of the problem of evidential diversity. As far as I know, *all* of these popular contemporary approaches are measure sensitive in the sense described here.

As it turns out (fortuitously), all three of the most popular contemporary relevance measures d, r, and l share property (4) (proofs omitted). But, Carnap's relevance measure $\mathfrak{r}$ does *not* satisfy (4), as the proof of Theorem 4 shows.

Theorem 4. $\mathfrak{r}$ *does not have the property expressed in* (4).[11]

Until we are given some compelling reason to prefer d, r, and l to Carnap's $\mathfrak{r}$ (and, to any other relevance measures which violate (4)—see fn. 11 and Appendix §D for further discussion), we should be wary about accepting the popular quantitative resolutions of the Ravens Paradox, or the recent Bayesian accounts of the confirmational significance of evidential diversity.[12]

2.5 An Important Theme in Our Examples

As our examples illustrate, several recent Bayesian confirmation theorists have presupposed the superiority of the difference measure d over one or more of the three alternative relevance measures r, l, and $\mathfrak{r}$. Moreover, we have seen that many well-known arguments in Bayesian confirmation theory depend sensitively on this assumption of d's superiority. To be sure, there are other arguments that fit this mold.[13] While there are some arguments in favor of d *as opposed to* r, there seem to be no arguments in the literature which favor d over the alternatives l and $\mathfrak{r}$. Moreover, as I will show in the next section, only one of the two popular arguments in favor of d *as opposed to* r is compelling. In contrast, several *general* arguments in favor of r, l, and $\mathfrak{r}$ *have* appeared in the literature.[14] It is precisely this kind of *general* argument that is needed to undergird the use of one particular relevance measure rather than any other.

In the next section, I will examine two recent arguments in favor of the difference measure d *as opposed to* the log-ratio measure r. While one of these arguments seems to definitively adjudicate between d and r (in favor of d), I will

[11] There are other relevance measures which violate (4). Mortimer (1988, §11.1) shows that the measure $\Pr(E \mid H \,\&\, K) - \Pr(E \mid K)$ violates (4). It also turns out that Nozick's (1981, 252) measure $\Pr(E \mid H \,\&\, K) - \Pr(E \mid \bar{H} \,\&\, K)$ violates (4). See Appendix §D for proofs.

[12] See Fitelson 1996 and Wayne 1995 for independent reasons to be wary of Horwich's (1982) account of the confirmational significance of evidential diversity.

[13] Kaplan (1996) offers several criticisms of Bayesian confirmation theory which presuppose the adequacy of the difference measure d. He then suggests (76, fn. 73) that all of his criticisms will also go through for all other relevance measures that have been proposed in the literature. But, one of his criticisms (84, fn. 86) does *not* apply to measure r.

[14] Milne (1996) argues that r is "the one true measure of confirmation." Good (1984), Heckerman (1988), and Schum (1994) all give general arguments in favor of l. And, Carnap (1962, §67) gives a general argument in favor of $\mathfrak{r}$. In Fitelson (1998b), I discuss each of these arguments in some depth, and I provide my own argument for the log-likelihood ratio l.

argue that neither of them will help to adjudicate between d and l, or between d and $\mathfrak{r}$. As a result, defenders of the difference measure will need to do further logical work to complete their enthymematic confirmation-theoretic arguments.

3 Two Arguments *Against* r

3.1 The "Deductive Insensitivity" Argument Against r

Rosenkrantz (1981) and Gillies (1986) point out the following fact about r:

(5) If $H \models E$, then $r(H, E \mid K) = r(H \& X, E \mid K)$, for any X.

Informally, (5) says that, *in the case of deductive evidence, $r(H, E \mid K)$ does not depend on the logical strength of H*. Gillies (1986) uses (5) as an argument against r, and in favor of the difference measure d. Rosenkrantz (1981) uses (5) as an argument against r, but he cautiously notes that *neither d nor l* satisfies (5). It is easy to show that $\mathfrak{r}$ doesn't have property (5) either (proof omitted).

I think Gillies pinpoints what is so peculiar and undesirable about (5) quite well, when he explains that

> On the Bayesian, or, indeed, on any inductivist position, the more a hypothesis H goes beyond [deductive] evidence E, the less H is supported by E. We have seen [in (5)] that r lacks this property that is essential for a Bayesian measure of support. (1986, 112; my brackets)

I agree with Gillies and Rosenkrantz that this argument provides a rather compelling reason to abandon r in favor of *either d or l or $\mathfrak{r}$*.[15] But, it says nothing about *which* of d, l, or $\mathfrak{r}$ should be adopted. So, this argument does not suffice to shore-up all of the measure sensitive arguments we have seen. Hence, it does not constitute a complete resolution of the problem of measure sensitivity.

3.2 The "Unintuitive Confirmation" Argument Against r

Several recent authors, including Sober (1994) and Schum (1994), have criticized r on the grounds that r sanctions "unintuitive" quantitative judgments about degree of confirmation in various (hypothetical) numerical examples.[16] For instance, Sober (1994) asks us to consider a hypothetical case involving a single collection of evidence E, and two hypotheses H_1 and H_2 (where, K is taken to be *tautologous*, and is thus suppressed) such that:

$$\Pr(H_1 \mid E) = 0.9 \qquad \Pr(H_1) = 0.09$$
$$\Pr(H_2 \mid E) = 0.0009 \qquad \Pr(H_2) = 0.00009$$

In such a case, we have the following pair of probabilistic facts:

$$(\dagger) \qquad \begin{aligned} d(H_1, E) = 0.81 &\gg d(H_2, E) = 0.00081 \\ r(H_1, E) = \log(10) &= r(H_2, E) \end{aligned}$$

It is then argued, by proponents of d, that (†) exposes a highly "unintuitive" feature of r, since this is case in which—"intuitively"—E confirms H_1 to a greater degree than E confirms H_2. But, according to r, E confirms both H_1 and H_2 to exactly the same degree. Therefore, this example is purported to *rule out* r (but *not* d, since d gets the "intuitively correct" answer here).

I am not too worried about (†), for two reasons. First, (†) can be only a reason to favor the difference measure over the ratio measure (or vice versa[17]); it has little or no bearing on the relative adequacy of either l or $\mathfrak{r}$. It is clear from the definitions of the measures that Carnap's $\mathfrak{r}$ is bound to *agree* with d's "intuitive" answer in such cases. Hence, $\mathfrak{r}$ is *immune* from the "unintuitive confirmation" criticism. Moreover, the log-likelihood ratio measure l certainly *could* agree with the "intuitively" correct judgments in such cases (depending on how the details needed to fix the *likelihoods* get filled-in). Indeed, Schum (1994, Ch. 5) argues nicely that the log-likelihood ratio measure l is largely immune to the kinds of "scaling effects" exhibited by r and d in (†). Unfortunately, neither Eells nor Sober (1994) nor Schlesinger (1995) considers how the measures l and $\mathfrak{r}$ might cope with their alleged counter-examples.

Second, there seems to be little or no *independent* support offered for the crucial premise of this argument. The argument is persuasive only if it is granted that the *intuitive* degree to which E confirms H_1 *is* greater than the *intuitive* degree to which E confirms H_2. The only reason that I have seen offered in support of this claim (e.g. Sober (1994)) is that $d(H_1, E) \gg d(H_2, E)$. But, this just seems to *beg the question*; it simply presupposes that the *intuitive* amount to which E confirms H *is* accurately gauged by the *difference* measure, and *not* by the *ratio* measure (or, by some other measure altogether). What we need here are *independent reasons* for believing *precisely this*!

4 Summary of Results

We have discussed three measure sensitive arguments which are aimed at showing that certain relevance measures are *in*adequate, and we have seen four mea-

[15] There are *lots* of compelling reasons to reject r. For instance, note that, according to r, the degree to which E confirms H need not be the same as the degree to which E *dis*confirms $\bar{H}$. That is, there are cases in which $r(H, E \mid K) \neq -r(\bar{H}, E \mid K)$. This sort of asymmetry is highly unintuitive, and is not shared by any of the other relevance measures under consideration.

[16] Sober (1994) borrows this line of criticism from Ellery Eells. Eells (pers. comm.) has voiced numerical examples of various kinds to illustrate the "unintuitive" consequences of r. I have chosen the present example because I think it is representative of the kind of examples these authors seem to have in mind.

[17] It has been argued by Schlesinger (1995) that parallel arguments can be run "backward" *against* d and *in favor of* r. Schlesinger (1995) describes a class of examples in which the difference measure seems to give the "unintuitive" answer (and, where the key probabilistic facts are analogous to (†)). Schlesinger's examples drive home the point that the *philosophical* conclusions one draws from hypothetical, *numerical* examples of these kinds will depend crucially on what one takes the *"intuitive"* answers to be in the first place. See below.

sure sensitive arguments which presuppose the *superiority* of certain relevance measures over others. Table 1 summarizes the arguments which presuppose that certain relevance measures are superior to others, and Table 2 summarizes the arguments against various relevance measures. These tables serve as a handy reference on the measure sensitivity problem in Bayesian confirmation theory.

	Is $\mathcal{A}$ valid wrt the measure:			
Name and Section of Argument $\mathcal{A}$	d?	r?	l?	$\mathfrak{r}$?
Rosenkrantz on "Irrelevant Conjunction" (See §2.2 and Appendix §B for discussion)	Yes	No	No	Yes
Earman on "Irrelevant Conjunction" (See §2.2 for discussion)	Yes	No	Yes	Yes
Eells on the Grue Paradox (See §2.3 and Appendix §C for discussion)	Yes	No	No	Yes
Horwich et al. on Ravens & Variety (See §2.4 and Appendix §D for discussion)	Yes	Yes	Yes	No

Table 1: Four arguments which presuppose the *superiority* of certain measures.

	Is $\mathcal{A}$ valid wrt the measure:			
Name and Section of Argument $\mathcal{A}$	d?	r?	l?	$\mathfrak{r}$?
Gillies's Popper-Miller Argument (See §2.1 and Appendix §A for discussion)	Yes	No	No	Yes
"Deductive Insensitivity" Argument (See §3.1 for discussion)	No	Yes	No	No
"Unintuitive Confirmation" Argument (See §3.2 for discussion)	No	Yes[18]	No	No

Table 2: Three arguments designed to show the *inadequacy* of certain measures.

[18] As I explain in § 3.2, I do *not* think this argument is compelling, even when aimed against r. But, to be charitable, I will grant that it is, at least, *valid* when aimed against r.

5 Conclusion: Where Do We Go From Here?

In the present paper, I have shown that many well-known arguments in quantitative Bayesian confirmation theory are valid only if the difference measure d is to be preferred over other relevance measures (at least, in the confirmational contexts in question). I have also shown that there are compelling reasons to prefer d over the log-ratio measure r. Unfortunately, like Rosenkrantz (1981), I have found *no* compelling reasons offered in the literature to prefer d over the log-*likelihood* ratio measure l (or Carnap's relevance measure $\mathfrak{r}$). As a result, philosophers like Gillies, Rosenkrantz, and Eells, whose arguments presuppose that d is preferable to both l and $\mathfrak{r}$, seem compelled to produce some *justification* for using d, rather than either l or $\mathfrak{r}$, to measure degree of confirmation.[19]

In general, there seem to be two viable strategies for coping with the problem of measure sensitivity. The first strategy is to simply *avoid* the problem entirely, by making sure that one's quantitative confirmation-theoretic arguments are *robust* (i.e. *in*sensitive to choice of measure of confirmation).[20] On the other hand, if plausible robust arguments can *not* be found in some context, then one should feel compelled to give reasons why one's chosen relevance measure (or class of relevance measures) $\mathfrak{c}^*$ should be preferred over other relevance measures, the use of which would render one's argument invalid.[21]

[19] In Fitelson 1998b, I argue that this will be a difficult task, since there are some rather strong arguments *in favor of* l and *against* d.

[20] This *can* be done in some contexts (e.g., in Fitelson 1998a, I outline a new, robust Bayesian resolution of the problem of evidential diversity, and Maher (1999) gives a new, measure *in*sensitive Bayesian resolution of the Ravens Paradox, based on Carnapian inductive logic); but, I doubt that plausible, robust quantitative Bayesian accounts can *always* be found.

[21] Ideally, it would be nice to see general, *desideratum/explicatum* arguments which rule out all but a relatively small class of ordinally equivalent measures of confirmation (*i.e.*, arguments like those given by Carnap (1962, §67), Good (1984), Heckerman (1988), and Milne (1996)). Such arguments would also have the virtue of contributing in a substantive way to the *theoretical underpinning* of quantitative Bayesian confirmation theory.

Appendix

A Proof of Theorem 1

Theorem 1. *There exist probability models such that*

$$l(H, E \mid K) \neq l(H \vee E, E \mid K) + l(H \vee \bar{E}, E \mid K).$$

Proof. For simplicity, I will assume that the background knowledge K consists only of *tautologies*. Then, by the definition of l, we have the following

$$\begin{aligned} l(H \vee E, E \mid K) + l(H \vee \bar{E}, E \mid K) &= \log\left[\frac{\Pr(E \mid H \vee E)}{\Pr(E \mid \overline{H \vee E})}\right] + \log\left[\frac{\Pr(E \mid H \vee \bar{E})}{\Pr(E \mid \overline{H \vee \bar{E}})}\right] \\ &= \log\left[\frac{\Pr(E \mid H \vee E)}{\Pr(E \mid \bar{H} \,\&\, \bar{E})}\right] + \log\left[\frac{\Pr(E \mid H \vee \bar{E})}{\Pr(E \mid \bar{H} \,\&\, E)}\right] \\ &= \log\left[\frac{\Pr(E \mid H \vee E)}{0}\right] + \log\left[\frac{\Pr(E \mid H \vee \bar{E})}{1}\right] \\ &= +\infty \\ &\neq l(H, E \mid K), \text{ provided that } l(H, E \mid K) \text{ is } \textit{finite}. \end{aligned}$$

There are lots of probability models of this kind in which $l(H, E \mid K)$ is finite. Any one of these is sufficient to establish the desired result. ❑

B Proof of Theorem 2

Theorem 2. *There exist probability models in which all three of the following obtain: (i)* $H \models E$*, (ii)* $r(H \,\&\, X, E \mid K) \neq \Pr(X \mid H \,\&\, K) \cdot r(H, E \mid K)$*, and (iii)* $l(H \,\&\, X, E \mid K) \neq \Pr(X \mid H \,\&\, K) \cdot l(H, E \mid K)$.[22]

Proof. Let K include the information that we are talking about a standard deck of cards with the usual probability structure. Let E be the proposition that some card $\mathcal{C}$, drawn at random from the deck, is a black card (*i.e.*, that $\mathcal{C}$ is either a ♣ or a ♠). Let H be the hypothesis that $\mathcal{C}$ is a ♠. And, let X be the proposition that $\mathcal{C}$ is a 7. Then, we have the following salient probabilities:

$\Pr(X \mid H \,\&\, K) = \frac{1}{13}$	$\Pr(H \mid E \,\&\, K) = \frac{1}{2}$	$\Pr(H \mid K) = \frac{1}{4}$
$\Pr(E \mid H \,\&\, X \,\&\, K) = 1$	$\Pr(E \mid H \,\&\, K) = 1$	$\Pr(E \mid \bar{H} \,\&\, K) = \frac{1}{3}$
$\Pr(H \,\&\, X \mid K) = \frac{1}{52}$	$\Pr(H \,\&\, X \mid E \,\&\, K) = \frac{1}{26}$	$\Pr(E \mid \overline{H \,\&\, X} \,\&\, K) = \frac{25}{51}$

[22] Strictly speaking, this theorem is *logically stronger* than Theorem 2, which only requires that there be a probability model in which (i) and (ii) obtain, and a probability model in which (i) and (iii) obtain (but, not necessarily *the same* model). Note, also, that the X in my countermodel is, intuitively, an *irrelevant* conjunct. I think this is apropos.

Hence, this probability model is such that all three of the following obtain:

(i)
$$H \models E$$

(ii)
$$\begin{aligned} r(H \,\&\, X, E \mid K) &= \log\left[\frac{1/26}{1/52}\right] \\ &= \log(2) \\ &\neq \Pr(X \mid H \,\&\, K) \cdot r(H, E \mid K) = \frac{1}{13} \cdot \log(2) \end{aligned}$$

(iii)
$$\begin{aligned} l(H \,\&\, X, E \mid K) &= \log\left[\frac{1}{25/51}\right] \\ &= \log\left[\frac{51}{25}\right] \\ &\neq \Pr(X \mid H \,\&\, K) \cdot l(H, E \mid K) = \frac{1}{13} \cdot \log(3) \end{aligned}$$

Consequently, this probability model is sufficient to establish Theorem 2. ❑

C Proof of Theorem 3

Theorem 3. *There exist probability models in which all three of the following obtain: (i)* $\beta > \delta$ *and* $\Pr(E \mid K) < \frac{1}{2}$*, (ii)* $l(H_1, E \mid K) < l(H_2, E \mid K)$*, and (iii)* $r(H_1, E \mid K) < r(H_2, E \mid K)$.[23]

Proof. I will prove Theorem 3 by describing a class of probability spaces in which all *four* of the following obtain.[24]

$(*)$ E confirms both H_1 and H_2 (given K)

(i) $\beta > \delta$ and $\Pr(E \mid K) < \frac{1}{2}$

(ii) $l(H_1, E \mid K) < l(H_2, E \mid K)$

(iii) $r(H_1, E \mid K) < r(H_2, E \mid K)$

To this end, consider the class of probability spaces containing the three events E, H_1, and H_2 (here, we take K to be *tautologous*, for simplicity) such that the eight basic (or, atomic) events in the space have the following probabilities:

[23] Where β and δ are defined as follows: $\beta =_{df} \Pr(H_1 \,\&\, E \mid K) - \Pr(H_2 \,\&\, E \mid K)$, and $\delta =_{df} \Pr(H_1 \,\&\, \bar{E} \mid K) - \Pr(H_2 \,\&\, \bar{E} \mid K)$. And, as was the case with Theorem 2 above (see fn. 22), this theorem is, technically, *logically stronger* than Theorem 3.

[24] It crucial that our countermodel be such that $(*)$ obtains. For instance, if we were to allow E to confirm H_2 but *dis*confirm H_1, then "counterexamples" would be easy to find, but they would not be a problem for Eells's resolution of the Grue Paradox, since Eells is clearly talking about cases in which E (the observation of a large number of green emeralds, before t_0) confirms *both* H_1 (that all emeralds are green) *and* H_2 (that all emeralds are grue).

$\Pr(H_1 \,\&\, \bar{H}_2 \,\&\, \bar{E}) = \mathsf{a} = \frac{1}{16}$	$\Pr(H_1 \,\&\, H_2 \,\&\, \bar{E}) = \mathsf{b} = \frac{1}{100}$
$\Pr(\bar{H}_1 \,\&\, H_2 \,\&\, \bar{E}) = \mathsf{c} = \frac{1}{32}$	$\Pr(H_1 \,\&\, \bar{H}_2 \,\&\, E) = \mathsf{d} = \frac{21}{320}$
$\Pr(H_1 \,\&\, H_2 \,\&\, E) = \mathsf{e} = \frac{1}{8}$	$\Pr(\bar{H}_1 \,\&\, H_2 \,\&\, E) = \mathsf{f} = \frac{1}{32}$
$\Pr(\bar{H}_1 \,\&\, \bar{H}_2 \,\&\, E) = \mathsf{g} = \frac{49}{320}$	$\Pr(\bar{H}_1 \,\&\, \bar{H}_2 \,\&\, \bar{E}) = \mathsf{h} = \frac{417}{800}$

Now, we verify that the class of probability spaces described above is such that $(*)$, (i), (ii), and (iii) all obtain. To see that $(*)$ holds, note that we have both $\Pr(H_1 \mid E) > \Pr(H_1)$, and $\Pr(H_2 \mid E) > \Pr(H_2)$.

$$\Pr(H_1 \mid E) = \frac{\mathsf{d}+\mathsf{e}}{\mathsf{d}+\mathsf{e}+\mathsf{f}+\mathsf{g}} = \frac{61}{120} \approx 0.5083$$
$$\Pr(H_2 \mid E) = \frac{\mathsf{e}+\mathsf{f}}{\mathsf{d}+\mathsf{e}+\mathsf{f}+\mathsf{g}} = \frac{5}{12} \approx 0.4167$$
$$\Pr(H_1) = \mathsf{a}+\mathsf{b}+\mathsf{d}+\mathsf{e} = \frac{421}{1600} \approx 0.2631$$
$$\Pr(H_2) = \mathsf{b}+\mathsf{c}+\mathsf{e}+\mathsf{f} = \frac{79}{400} = 0.1975$$

To see that (i) holds, note that $\Pr(E) < \frac{1}{2}$.

$$\Pr(E) = \mathsf{d}+\mathsf{e}+\mathsf{f}+\mathsf{g} = \frac{3}{8} = 0.375$$

And, that $\Pr(H_1 \,\&\, E) - \Pr(H_2 \,\&\, E) > \Pr(H_1 \,\&\, \bar{E}) - \Pr(H_2 \,\&\, \bar{E})$ (i.e. $\beta > \delta$).

$$\beta = \mathsf{d} - \mathsf{f} = \frac{11}{320} \approx 0.0344$$
$$\delta = \mathsf{a} - \mathsf{c} = \frac{1}{32} \approx 0.0313$$

Next, we verify that (ii) holds in our example (i.e. $l(H_1, E) < l(H_2, E)$).

$$l(H_1, E) = \log\left[\frac{(1-\mathsf{a}-\mathsf{b}-\mathsf{d}-\mathsf{e})\;(\mathsf{d}+\mathsf{e})}{(\mathsf{a}+\mathsf{b}+\mathsf{d}+\mathsf{e})\;(\mathsf{f}+\mathsf{g})}\right] = \log\left(\frac{71919}{24839}\right) \approx \log(2.895)$$
$$l(H_2, E) = \log\left[\frac{(1-\mathsf{b}-\mathsf{c}-\mathsf{e}-\mathsf{f})\;(\mathsf{e}+\mathsf{f})}{(\mathsf{b}+\mathsf{c}+\mathsf{e}+\mathsf{f})\;(\mathsf{d}+\mathsf{g})}\right] = \log\left(\frac{1605}{553}\right) \approx \log(2.902)$$

Finally, we verify that (iii) holds in our example (i.e. $r(H_1, E) < r(H_2, E)$).

$$r(H_1, E) = \log\left[\frac{\mathsf{d}+\mathsf{e}}{(\mathsf{a}+\mathsf{b}+\mathsf{d}+\mathsf{e})\;(\mathsf{d}+\mathsf{e}+\mathsf{f}+\mathsf{g})}\right] = \log\left(\frac{2440}{1263}\right) \approx \log(1.932)$$
$$r(H_2, E) = \log\left[\frac{\mathsf{e}+\mathsf{f}}{(\mathsf{b}+\mathsf{c}+\mathsf{e}+\mathsf{f})\;(\mathsf{d}+\mathsf{e}+\mathsf{f}+\mathsf{g})}\right] = \log\left(\frac{500}{237}\right) \approx \log(2.110)$$

This completes the proof of Theorem 3. ❑

D Proof of Theorem 4

Theorem 4. *There exist probability models in which both of the following obtain: (i)* $\Pr(H \mid E_1 \,\&\, K) > \Pr(H \mid E_2 \,\&\, K)$, *and (ii)* $\mathfrak{r}(H, E_1 \mid K) \leq \mathfrak{r}(H, E_2 \mid K)$.

Proof. I will prove Theorem 4 by describing a class of probability spaces in which all *three* of the following obtain.[25]

$(*)$ Each of E_1 and E_2 confirms H (given K)

(i) $\Pr(H \mid E_1 \,\&\, K) > \Pr(H \mid E_2 \,\&\, K)$

(ii) $\mathfrak{r}(H, E_1 \mid K) < \mathfrak{r}(H, E_2 \mid K)$

To this end, consider the class of probability spaces containing the three events E_1, E_2, and H (again, we take K to be *tautologous*, for simplicity) such that the eight basic (or, atomic) events in the space have the following probabilities:

$\Pr(E_1 \,\&\, \bar{E}_2 \,\&\, \bar{H}) = \mathsf{a} = \frac{1}{1000}$	$\Pr(E_1 \,\&\, E_2 \,\&\, \bar{H}) = \mathsf{b} = \frac{1}{1000}$
$\Pr(\bar{E}_1 \,\&\, E_2 \,\&\, \bar{H}) = \mathsf{c} = \frac{1}{200}$	$\Pr(E_1 \,\&\, \bar{E}_2 \,\&\, H) = \mathsf{d} = \frac{1}{100}$
$\Pr(E_1 \,\&\, E_2 \,\&\, H) = \mathsf{e} = \frac{1}{100}$	$\Pr(\bar{E}_1 \,\&\, E_2 \,\&\, H) = \mathsf{f} = \frac{1}{25}$
$\Pr(\bar{E}_1 \,\&\, \bar{E}_2 \,\&\, H) = \mathsf{g} = \frac{1}{500}$	$\Pr(\bar{E}_1 \,\&\, \bar{E}_2 \,\&\, \bar{H}) = \mathsf{h} = \frac{931}{1000}$

Now, we verify that the class of probability spaces described above is such that $(*)$, (i), and (ii) all obtain. To see that $(*)$ and (i) both hold, note that we have $\Pr(H \mid E_1) > \Pr(H)$, $\Pr(H \mid E_2) > \Pr(H)$, and $\Pr(H \mid E_1) > \Pr(H \mid E_2)$:

$$\Pr(H \mid E_1) = \frac{\mathsf{d}+\mathsf{e}}{\mathsf{a}+\mathsf{b}+\mathsf{d}+\mathsf{e}} = \frac{10}{11} \approx 0.909$$
$$\Pr(H \mid E_2) = \frac{\mathsf{e}+\mathsf{f}}{\mathsf{b}+\mathsf{c}+\mathsf{e}+\mathsf{f}} = \frac{25}{28} \approx 0.893$$
$$\Pr(H) = \mathsf{d}+\mathsf{e}+\mathsf{f}+\mathsf{g} = \frac{31}{500} = 0.062$$

And, to see that (ii) holds, note that $\mathfrak{r}(H, E_1) < \mathfrak{r}(H, E_2)$.[26]

$$\mathfrak{r}(H, E_1) = (\mathsf{a}+\mathsf{b}+\mathsf{d}+\mathsf{e}) \cdot \left[\frac{\mathsf{d}+\mathsf{e}}{\mathsf{a}+\mathsf{b}+\mathsf{d}+\mathsf{e}} - (\mathsf{d}+\mathsf{e}+\mathsf{f}+\mathsf{g})\right] = \frac{4659}{250000} \approx 0.0186$$
$$\mathfrak{r}(H, E_2) = (\mathsf{b}+\mathsf{c}+\mathsf{e}+\mathsf{f}) \cdot \left[\frac{\mathsf{e}+\mathsf{f}}{\mathsf{b}+\mathsf{c}+\mathsf{e}+\mathsf{f}} - (\mathsf{d}+\mathsf{e}+\mathsf{f}+\mathsf{g})\right] = \frac{727}{15625} \approx 0.0465$$

This completes the proof of Theorem 4, as well as the Appendix. ❑

[25] It is important that our countermodel satisfy $(*)$. In the ravens paradox, it should be granted that both a black raven (E_1) and a non-black non-raven (E_2) may confirm that all ravens are black (H). Similarly, it should be granted that both a "varied" (or "diverse") set of evidence (E_1) and a "narrow" set of evidence (E_2) can confirm a hypothesis under test (H). Wayne (1995) presents a "counterexample" to Horwich's (1982) Bayesian account of evidential diversity which fails to respect this constraint. See Fitelson 1996 for details.

[26] This is also a model in which both $\Pr(E_1 \mid H) - \Pr(E_1) < \Pr(E_2 \mid H) - \Pr(E_2)$, and $\Pr(E_1 \mid H) - \Pr(E_1 \mid \bar{H}) < \Pr(E_2 \mid H) - \Pr(E_2 \mid \bar{H})$ (check this!). So, the relevance measures of both Mortimer (1988, §11.1) and Nozick (1981, 252), respectively, *also* violate (4).

References

Carnap, R. (1962), *Logical Foundations of Probability*, 2nd ed. Chicago: University of Chicago Press.

Earman, J. (1992), *Bayes or Bust: A Critical Examination of Bayesian Confirmation Theory*. Cambridge, MA: MIT Press.

Eells, E. (1982), *Rational Decision and Causality*. Cambridge: Cambridge University Press.

Fitelson, B. (1996), "Wayne, Horwich, and Evidential Diversity", *Philosophy of Science* 63: 652–660.

———. (1998a), "A Bayesian Account of Independent inductive Support with an Application to Evidential Diversity", unpublished manuscript.

———. (1998b), "Independent Inductive Support and Measures of Confirmation", unpublished manuscript.

Gillies, D. (1986), "In Defense of the Popper-Miller Argument", *Philosophy of Science* 53: 110–113.

Good, I. (1984), "The Best Explicatum for Weight of Evidence", *Journal of Statistical Computation and Simulation* 19: 294–299.

Heckerman, D. (1988), "An Axiomatic Framework for Belief Updates", in L. Kanal and J. Lemmer, (eds.), *Uncertainty in Artificial Intelligence 2*. New York: Elsevier Science Publishers, 11–22.

Horvitz, E. and D. Heckerman (1986), "The Inconsistent Use of Certainty Measures in Artificial Intelligence Research", in L. Kanal and J. Lemmer, (eds.), *Uncertainty in Artificial Intelligence 1*. New York: Elsevier Science Publishers, 137–151.

Horwich, P. (1982), *Probability and Evidence*. Cambridge: Cambridge University Press.

Hosiasson-Lindenbaum, J. (1940), "On Confirmation", *Journal of Symbolic Logic* 5: 133–148.

Jeffrey, R. (1992), *Probability and the Art of Judgment*. Cambridge: Cambridge University Press.

Kaplan, M. (1996), *Decision Theory as Philosophy*. Cambridge: Cambridge University Press.

Keynes, J. (1921), *A Treatise on Probability*. London: Macmillan.

Krantz, D., R. Luce, P. Suppes, and A. Tversky (1971), *Foundations of Measurement*, vol. 1. New York: Academic Press.

Kyburg, H. (1983), "Recent Work in Inductive Logic", in T. Machan and K. Lucey, (eds.), *Recent Work in Philosophy*. Lanham: Rowman & Allanheld, 87–150.

Mackie, J. (1969), "The Relevance Criterion of Confirmation", *The British Journal for the Philosophy of Science* 20: 27–40.

Maher, P. (1996), "Subjective and Objective Confirmation", *Philosophy of Science* 63: 149–174.

———. (1999), "Inductive Logic and the Ravens Paradox", *Philosophy of Science* 66: 50–70.

Milne, P. (1996), "$log[p(h/eb)/p(h/b)]$ is the One True Measure of Confirmation", *Philosophy of Science* 63: 21–26.

Mortimer, H. (1988), *The Logic of Induction*. Paramus: Prentice Hall.

Nozick, R. (1981), *Philosophical Explanations*. Cambridge, MA: Harvard University Press.

Pearl, J. (1988), *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Francisco: Morgan Kauffman.

Popper, K. and D. Miller (1983), "The Impossibility of Inductive Probability", *Nature* 302: 687–688.

Redhead, M. (1985), 'On the Impossibility of Inductive Probability', *The British Journal for the Philosophy of Science* 36: 185–191.

Rosenkrantz, R. (1981), *Foundations and Applications of Inductive Probability*. Atascadero, CA: Ridgeview.

———. (1994), "Bayesian Confirmation: Paradise Regained", *The British Journal for the Philosophy of Science* 45: 467–476.

Schlesinger, G. (1995), "Measuring Degrees of Confirmation", *Analysis* 55: 208–212.

Schum, D. (1994), *The Evidential Foundations of Probabilistic Reasoning*. New York: John Wiley & Sons.

Sober, E. (1994), "No model, No Inference: A Bayesian Primer on the Grue Problem", in D. Stalker, (ed.), *Grue! The New Riddle of Induction*. Chicago: Open Court, 225–240.

Wayne, A. (1995), "Bayesianism and Diverse Evidence", *Philosophy of Science* 62: 111–121.