Erkenn (2016) 81:155–181
DOI 10.1007/s10670-015-9734-1

ORIGINAL ARTICLE

Evaluating Test Cases for Probabilistic Measures of Coherence

Jakob Koscholke[1]

Received: 26 September 2014 / Accepted: 8 April 2015 / Published online: 17 April 2015

Abstract How can we determine the adequacy of a probabilistic coherence measure? A widely accepted approach to this question besides formulating adequacy constraints is to employ paradigmatic test cases consisting of a scenario providing a joint probability distribution over some specified set of propositions coupled with a normative coherence assessment for this set. However, despite the popularity of the test case approach, a systematic evaluation of the proposed test cases is still missing. This paper's aim is to change this. Using a custom written computer program for the necessary probabilistic calculations a large number of coherence measures in an extensive collection of test cases is examined. The result is a detailed overview of the test case performance of any probabilistic coherence measures proposed so far. It turns out that none of the popular coherence measures such as Shogenji's, Glass' and Olsson's, Fitelson's or Douven and Meijs' but two rather unnoticed measures perform best. This, however, does not mean that the other measures can be rejected straightforwardly. Instead, the results presented here are to be understood as a contribution among others to the project of finding adequate probabilistic coherence measures.

1 Introduction

Probabilistic coherence measures are functions assigning real numbers to sets of propositions under some joint probability distribution. At best, the assigned numbers represent how good the respective propositions fit or hang together, agree with or mutually support each other—briefly, how *coherent* the propositions are. At worst, they do not. Hence, the development of probabilistic measures of coherence as pursued by formal philosophers such as Douven and Meijs (2007), Fitelson (2003, (2004), Glass (2002), Meijs (2005), Olsson (2002), Roche (2013), Schippers (2014),

✉ Jakob Koscholke
jakob.koscholke@uni-oldenburg.de

[1] Philosophy Department, University of Oldenburg, 26111 Oldenburg, Germany

Schupbach (2011) or Shogenji (1999) can be understood as a search for a quantitative explication—in the sense of Carnap (1950)—of the concept of coherence. Of course, in order to be an explication at all, the *explicandum*—here, the concept of coherence—and the *explicatum*—here, a probabilistic coherence measure—have to be similar (besides the expclicatum having to be exact, fruitful and simple). To ensure this kind of similarity several authors (e.g. Bovens and Hartmann 2005; Fitelson 2004; Moretti and Akiba 2007; Siebel and Wolff 2008) have formulated adequacy constraints for probabilistic coherence measures (for an overview cf. Schippers 2014). These constraints are based on considerations regarding the relationship between the characteristics of some set of propositions such as being equivalent, inconsistent or being tied together by explanatory, probabilistic relevance, deductive entailment or other inferential relations (cf. BonJour 1985) and its degree of coherence. Adequacy constraints can therefore be considered as serving as reference points for the evaluation of a coherence measure's adequacy.

Another common practice, however, besides formulating general desiderata for probabilistic coherence measures in order to evaluate their adequacy is to employ test cases (cf. e.g. Bovens and Hartmann 2003; Meijs 2005; Siebel 2004). Test cases for probabilistic coherence measures are paradigmatic situations providing information about a specified set of propositions such that the values of a probabilistic coherence measure for this set can be computed. Most important, test cases come with a normative coherence assessment for the respective set based on considerations regarding the situation in which the set is located. The evaluation is then quite simple. If some measure's values in a certain test case are in accordance with the normative coherence assessment provided by this case and the assessment has strong intuitive support, then the measure remains a candidate for an adequate measure of coherence. If a measure's value is not in accordance with the assessment, its credibility as an adequate coherence measure decreases. Though this method is very appealing, it has merely been used for a limited number of measures in single test cases. The main reasons for this shortcoming are the time consuming computational effort and the possibility of miscalculation. Thus, for the purpose of this investigation a custom written computer program in GNU Octave (the source code is free via the author) has been used to circumvent these two problems and to be able to test any extant probabilistic coherence measure in any test case proposed so far. Additionally, the program allows for an easy implementation of future coherence measures as well as future test cases. Hence, besides evaluating probabilistic coherence measures with respect to a collection of test cases it can also be considered a subordinate aim of this paper to demonstrate the capabilities of the program.

The structure of this paper is rather straightforward. In Sect. 2 the notion of a probabilistic coherence measure is introduced formally. Then, all probabilistic coherence measures that have been proposed in the literature are presented. In addition, this collection is complemented with measures that have not been suggested as coherence measure but might be promising candidates. In Sect. 3 the notion of a test case is introduced. After that each test case is presented followed by every measure's performance in the respective case. Finally, in Sect. 4 the results are summarized and critically discussed with respect to the issue of determining the adequacy of probabilistic coherence measures.

2 Probabilistic Measures of Coherence

Before introducing the notion of a probabilistic coherence measure the necessary formal framework needs to be established first. Let L be a classical propositional language consisting of atomic formulae closed under some functional complete selection of classical logical connectives such as $\{\neg, \wedge\}$ where connectives like $\vee$ or $\rightarrow$ can be defined in terms of the selection. Let then 2^L denote the powerset of L, i.e. the set of all subsets of L, let furthermore $P : L \rightarrow [0, 1]$ be a probability function with conditional probability defined as $P(x_1|x_2) = P(x_1 \wedge x_2)/P(x_2)$ for $x_2 \in L$ with $P(x_2) \neq 0$ and let $\mathbf{P}$ denote the set of all probability functions over L. In order to define the domain of a probabilistic coherence measure a further restriction is needed, sometimes referred to as "Rescher's principle" (Olsson 2005, 17). According to this principle, "[c]oherence is [...] a feature that propositions cannot have in isolation but only in groups, containing several—i.e. at least two—propositions" (Rescher 1973, 32). Let therefore be $2^L_{\geq 2} = \{X \subseteq 2^L : |X| \geq 2\}$. A probabilistic coherence measure can then be defined as a function—more specifically, a *partial* function due to some undefined function values—$C : 2^L_{\geq 2} \times \mathbf{P} \rightarrow \mathbb{R}$ mapping pairs (X_i, P_i) onto real numbers where X_i is a set of propositions under some joint probability distribution P_i.

Now suppose we would like to assess the degree of coherence of some finite, non-empty, non-singleton set $X = \{x_1, \ldots, x_n\}$. According to Shogenji (1999), this can be done the following way: take the joint probability of X's members and divide it by the product over all marginal probabilities of the respective propositions. This quantifies the propositions' deviation from their joint probabilistic independence:

$$C_{sho}(X) = \frac{P\left(\bigwedge_{i=1}^{n} x_i\right)}{\prod_{i=1}^{n} P(x_i)}$$

In order to overcome difficulties of Shogenji's measure associated with its insensitivity for subsets of propositions as pointed out by Fitelson (2003), Schupbach (2011) has suggested the following generalization of Shogenji's measure: to assess the degree of coherence of X, apply a log-normalized version of Shogenji's coherence measure to each set X'_{ij} which is the i-th subset of X and contains $j \geq 2$ proposition. For each of them divide its coherence value by the number of sets with j members, sum up the resulting values and divide this sum by X's cardinality minus one ignoring singleton sets:

$$C_{sch}(X) = \frac{\sum_{j=2}^{n} \frac{\sum_{i=1}^{\binom{n}{j}} \log\left(C_{sho}(X'_{ij})\right)}{\binom{n}{j}}}{n-1}$$

Glass (2002) and Olsson (2002) have proposed a different account. In this case, in order to compute the degree coherence of X simply divide the probability of the

conjunction by the probability of the disjunction over X's members. Set-theoretically speaking, this can be understood as quantifying the propositions' relative overlap:

$$C_{go}(X) = \frac{P\left(\bigwedge_{i=1}^{n} x_i\right)}{P\left(\bigvee_{i=1}^{n} x_i\right)}$$

Based on the same idea, but less complicated, Meijs (2006) has suggested the following generalization of the coherence measure by Glass and Olsson: in order to assess the coherence of X, take the straight average over all values of the Glass–Olsson measure applied to every subset X_i' of X with $|X_i'| \geq 2$:

$$C_{mei}(X) = \frac{\sum_{i=1}^{(2^n-n)-1} C_{go}(X_i')}{(2^n - n) - 1}$$

A whole family of coherence measures can be obtained using an approach systematically developed by Douven and Meijs (2007). According to their approach, coherence is to be understood as average mutual support. Since there is a variety of probabilistic measures of support (for an overview cf. Crupi et al. 2007) one can easily obtain a huge collection of candidates for coherence measures based on them. The basic idea runs as follows: to assess the coherence of X, consider all pairs $(X', X'')_i$ where X' and X'' are non-empty, disjoint subsets of X. For each pair, take the conjunctions over the propositions contained in the respective set and calculate the average degree of support according to some chosen probabilistic support measure S, i.e. a two-place function that is supposed to quantify the degree to which its first argument, some proposition x_1 is supported by its second argument, another proposition x_2:

$$C_S(X) = \frac{\sum_{i=1}^{(3^n-2^{n+1})-1} S\left(\left(\bigwedge_{x_j \in X'} x_j, \bigwedge_{x_k \in X''} x_k\right)_i\right)}{(3^n - 2^{n+1}) - 1}$$

In the literature several measures of support have been suggested as a foundation for coherence measures. For his coherence measure $C_{S_{fit}}$ Fitelson (2004) uses a case-sensitive variation of Kemeny and Oppenheim's (1952) measures of factual support:

$$S_{fit}(x_1, x_2) = \begin{cases} \dfrac{P(x_2|x_1) - P(x_2|\neg x_1)}{P(x_2|x_1) + P(x_2|\neg x_1)} & \text{if } \mathrm{x_2} \nvdash \mathrm{x_1} \text{ and } \mathrm{x_2} \nvdash \neg \mathrm{x_1} \\ 1 & \text{if } \mathrm{x_2} \vdash \mathrm{x_1} \text{ and } \mathrm{x_2} \nvdash \bot \\ -1 & \text{if } \mathrm{x_2} \vdash \neg \mathrm{x_1} \end{cases}$$

Douven and Meijs (2007), by contrast, prefer Carnap's (1950) difference measure of support for their favourite coherence measure here denoted $C_{S_{car}}$:

$$S_{car}(x_1, x_2) = P(x_1|x_2) - P(x_1)$$

Notice that due to the symmetry of the average mutual support approach Douven and Meijs could also have used the counterpart of Carnap's difference measure by Mortimer (1988) where only the two arguments are interchanged. The same holds for Levi's (1962) corroboration measure which can easily be shown to be identical to Carnaps's difference measure. As we will see, this also holds for other coherence measures based on measures of evidential support. Besides their favourite coherence measure Douven and Meijs also investigated other measures of support as foundations for coherence measures without advocating them. One of them is Keynes' (1921) relevance quotient:

$$S_{key}(x_1, x_2) = \frac{P(x_1|x_2)}{P(x_1)}$$

Here again, instead of using Keynes' measure one could obtain the same coherence measure using Kuipers' (2000) symmetrically identical ratio measure or Finch's (1960) ratio measures of evidential support which would yield identical function values shifted by -1. It is also worth noticing that in the case of sets of two propositions the coherence measures resulting from using these confirmation measures are identical to Shogenji's coherence measure or shifted by -1. Another measure taken into consideration by Douven and Meijs is the well-known likelihood ratio measure by Good (1984) for which instead Joyce's (2008) odds-ratio measure could also have been used:

$$S_{goo}(x_1, x_2) = \frac{P(x_2|x_1)}{P(x_2|\neg x_1)}$$

The aforementioned support measure are based on an incremental—as opposed to an absolute—understanding of evidential support. More recently, Roche (2013) proposed his favourite candidate for a coherence measure $C_{S_{roc}}$ based on Douven and Meijs' approach employing a case-sensitive notion of absolute support, namely the conditional probability:

$$S_{roc}(x_1, x_2) = \begin{cases} P(x_1|x_2) & \text{if } \mathrm{x}_2 \nvdash \mathrm{x}_1 \text{ and } \mathrm{x}_2 \nvdash \neg \mathrm{x}_1 \\ 1 & \text{if } \mathrm{x}_2 \vdash \mathrm{x}_1 \text{ and } \mathrm{x}_2 \nvdash \bot \\ 0 & \text{if } \mathrm{x}_2 \vdash \neg \mathrm{x}_1 \end{cases}$$

Another more recent coherence measure $C_{S_{sch}}$ has been developed by Schippers (2014) based on his own measure of support:

$$S_{sch}(x_1, x_2) = \begin{cases} \dfrac{P(x_1|x_2) - P(x_1|\neg x_2)}{1 - P(x_1|\neg x_2)} & \text{if } \mathrm{P}(\mathrm{x}_1|\mathrm{x}_2) \geq \mathrm{P}(\mathrm{x}_1) \\ \dfrac{P(x_1|x_2) - P(x_1|\neg x_2)}{P(x_1|\neg x_2)} & \text{if } \mathrm{P}(\mathrm{x}_1|\mathrm{x}_2) < \mathrm{P}(\mathrm{x}_1) \end{cases}$$

The same coherence measure could have been obtained using Cheng's (1997) causal Power-PC measure. Notice that just like Cheng's measure can be understood as a

normalization of Nozick's (1981) measure, Schipper's support measure can be understood as a normalization of Christensen's (1999) measure. This already extensive collection has been expanded by Siebel and Wolff (2008). They considered further alleged coherence measures based on Douven and Meijs' approach. For instance, they used Carnap's (1950) relevance measure:

$$S_{car'}(x_1, x_2) = P(x_1 \wedge x_2) - P(x_1) \cdot P(x_2)$$

Siebel and Wolff also investigated Nozick's (1981) counterfactual likelihood difference measure for which the resulting coherence measure is identical to the measure obtained when using Christensen's (1999) counterfactual difference measure:

$$S_{noz}(x_1, x_2) = P(x_2|x_1) - P(x_2|\neg x_1)$$

Furthermore, Siebel and Wolff took into account Popper's (1954) corroboration measure as a foundation for a coherence measure:

$$S_{pop}(x_1, x_2) = \frac{P(x_2|x_1) - P(x_2)}{P(x_2|x_1) + P(x_2)} \cdot (1 + P(x_1) \cdot P(x_1|x_2))$$

And ultimately, they also included Rescher's (1958) measure of evidential support in their examination:

$$S_{res}(x_1, x_2) = \frac{P(x_1|x_2) - P(x_1)}{1 - P(x_1)} \cdot P(x_2)$$

To make the collection complete, we will also take into account coherence measures based on further measures of positive relevance. These include Crupi et al.'s (2007) so-called z-measure of evidential support which has received some attention lately:

$$S_{cru}(x_1, x_2) = \begin{cases} \dfrac{P(x_1|x_2) - P(x_1)}{1 - P(x_1)} & \text{if } \mathrm{P}(\mathrm{x}_1|\mathrm{x}_2) \geq \mathrm{P}(\mathrm{x}_1) \\ \dfrac{P(x_1|x_2) - P(x_1)}{P(x_1)} & \text{if } \mathrm{P}(\mathrm{x}_1|\mathrm{x}_2) < \mathrm{P}(\mathrm{x}_1) \end{cases}$$

Moreover, we will include Gaifman's (1979) measure as an ingredient for a coherence measure:

$$S_{gai}(x_1, x_2) = \frac{P(\neg x_1)}{P(\neg x_1|x_2)}$$

Rips' (2001) measure is also included:

$$S_{rip}(x_1, x_2) = 1 - \frac{P(\neg x_2|x_1)}{P(\neg x_2)}$$

Finally, we also take into account Shogenji's (2012) measure of justification which according to him is also a measure of evidential support:

$$S_{sho}(x_1, x_2) = \frac{\log_2 P(x_1|x_2) - \log_2 P(x_1)}{-\log_2 P(x_1)}$$

In the following it will be helpful to know some of the general properties of the introduced measures, such as their threshold value t indicating neutral coherence and their range r (Table 1).

Two things should be mentioned here. First, any measure with a half-open or open interval as its range cannot assign minimal, maximal or both degrees of coherence. Second, it should also be mentioned that for the measures put forward by Glass and Olsson, Meijs and Roche, the neutral value is to be understood in a different way compared to the other measures. While for all measures except these three the neutral value indicates the joint independence of the propositions in some set, the values of the aforementioned measures indicate an equal overlap of the propositions in question and their complement or in other words the propositions are as coherent as their negations.

Besides these rather general properties the philosophical motivations underlying the presented measures shall not be discussed here. For the philosophical backgrounds and further properties of the presented measures the reader is referred to the original papers in which the respective measures have been proposed. It is, however, worth noticing that the variety of motivations underlying the different proposals shows that the proponents aimed at explicating different aspects of the concept of coherence. This indicates that there might be more more than a single probabilistic coherence measure. Recent results by Schippers (2014) also point in this direction suggesting that we should be pluralists with respect to the concept of coherence and probabilistic measures of coherence. Nevertheless, it is important to notice that the following investigation does not rely on this assumption. Rather, the

Table 1 Neutral point t and range r

	t	r
C_{sho}	1	$[0, \infty)$
C_{sch}	0	$(-\infty, \infty)$
C_{go}	.5	$[0, 1]$
C_{mei}	.5	$[0, 1]$
$C_{S_{fit}}$	0	$[-1, 1]$
$C_{S_{car}}$	0	$[-1, 1)$
$C_{S_{key}}$	1	$[0, \infty)$
$C_{S_{goo}}$	1	$[0, \infty)$
$C_{S_{roc}}$	.5	$[0, 1]$
$C_{S_{sch}}$	0	$[-1, 1]$
$C_{S_{car'}}$	0	$(-1, 1)$
$C_{S_{noz}}$	0	$[-1, 1]$
$C_{S_{pop}}$	0	$[-1, 1]$
$C_{S_{res}}$	0	$[-1, 1]$
$C_{S_{cru}}$	0	$[-1, 1]$
$C_{S_{gai}}$	1	$[0, \infty)$
$C_{S_{rip}}$	0	$(-\infty, 1]$
$C_{S_{sho}}$	0	$[-\infty, 1]$

results of this investigation can contribute to the question of pluralism with respect to the concept of coherence, e.g. by examining whether there are classes of test cases with the same underlying coherence intuition in which certain measures fail while succeeding in others with a different intuition. For critical discussions of the presented coherence measures see e.g. Akiba (2000), Fitelson (2003), Moretti and Akiba (2007), Schippers (2014), Siebel (2004, 2005), Siebel and Wolff (2008). For discussions of support measures see e.g. Crupi et al. (2007), Eells and Fitelson (2002), Tentori et al. (2007). Having presented the test candidates we may now turn to the test cases.

3 Test Cases and Results

The initially established formal framework enables us to introduce the notion of a test case more precisely. Let again a pair (X_i, P_i) denote some set of propositions X_i under a joint probability distribution P_i. Such pairs might be thought of as situations in which X_i's propositions have probabilities according to P_i. Moreover, let A denote an assessment of the degree of coherence of some set of propositions in a certain situation, e.g. $C(X_i) \lessgtr \theta_C$ where θ_C is a threshold value of a specific measure C, $C(X_i) \lessgtr C(X_j)$ where two sets under different probability distributions are compared or $C(X_i') \lessgtr C(X_i'')$ with $X_i', X_i'' \subset X_i$ and $X_i' \cap X_i'' \neq \emptyset$ where subsets of a set under some distribution is examined. Then a test case is a pair $T = (\{(X_1, P_1), \ldots, (X_n, P_n)\}, \{A_1, \ldots, A_m\})$ of a set of situations and a set of coherence assessments.

In the following subsections the collection of 18 probabilistic coherence measure introduced in Sect. 2 is submitted to 11 test cases from the literature. For each test case the course will be as follows. First, the scenario together with the set of propositions and the expected coherence assessment is described. After that the calculated coherence values for every probabilistic coherence measure are presented. Notice that since some function values are very small but nevertheless relevant these values will be represented in scientific notation where e.g. $-4.26\text{e}{-05}$ stands for -0.0000426. Finally, these results are evaluated with respect to the coherence assessment provided for the respective case. In the last subsection of this section the results are summarized and discussed.

Throughout this investigation positive test case results, i.e. results that agree with the provided coherence assessment are rewarded with a score of 1 while negative results receive a score of 0. Notice that, following Siebel and Wolff (2008) undefined function values such as $-\infty$ or ∞ are treated as if the respective measure remained silent regarding a coherence assessment and are marked "NaN" standing for "not a number". In such cases the score will be 0. Also notice that the plausibility of coherence assessments provided for each test case are not going to be subject of discussion here. This task is left for future research. The aim of this section is to evaluate the performance of each measure in each test case under the assumption that all normative coherence assessments are equally rational. In Sect. 3.12, however, a tentative solution for this shortcoming is offered.

3.1 Akiba's Die Case

Akiba (2000) has developed a test case that is supposed to show that Shogenji's coherence measure fails to handle certain sets of propositions adequately. The problematic cases Akiba points out are sets that do not only contain a finite number of propositions but also deductive consequences of these propositions. His intuition here is that if two sets with the same cardinality consist of some proposition and furthermore each set contains some proposition that is logically entailed by the proposition they both have in common, then the degree of coherence of both sets should be the same. Akiba's test case runs as follows:

> **Situation:** Imagine tossing a fair die and consider the following three predictions about the outcome:
>
> x_1: The die will come up 2.
> x_2: The die will come up 2 or 4.
> x_3: The die will come up 2 or 4 or 6.

According to Akiba, the sets $X_1 = \{x_1, x_2\}$ and $X_2 = \{x_1, x_3\}$ should be equal with respect to their degrees of coherence since both x_2 and x_3 are deductive consequences of x_1. Let us take a look at the values (Table 2).

Table 2 Results for Akiba's die case

	X_1	X_2	Score
C_{sho}	3	2	0
C_{sch}	0.477	0.301	0
C_{go}	0.5	0.333	0
C_{mei}	0.5	0.333	0
$C_{S_{fit}}$	0.833	0.714	0
$C_{S_{car}}$	0.5	0.333	0
$C_{S_{key}}$	3	2	0
$C_{S_{goo}}$	NaN	NaN	0
$C_{S_{roc}}$	0.75	0.667	0
$C_{S_{sch}}$	0.75	0.667	0
$C_{S_{car'}}$	0.111	0.0833	0
$C_{S_{noz}}$	0.65	0.467	0
$C_{S_{pop}}$	0.604	0.426	0
$C_{S_{res}}$	0.15	0.133	0
$C_{S_{cru}}$	0.7	0.6	0
$C_{S_{gai}}$	NaN	NaN	0
$C_{S_{rip}}$	0.7	0.6	0
$C_{S_{sho}}$	0.807	0.693	0

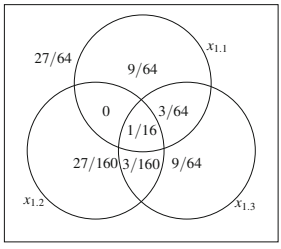
These result are truly devastating. Akiba's test case is not only a problem for Shogenji's measure of coherence but for all considered probabilistic coherence measures. Not a single measure satisfies Akiba's normative coherence assessment. Most measures assign X_1 a higher degree of coherence than X_2. The average mutual support measures based on Good's likelihood-ratio measures and Gaifman's support measure fail because they have non-defined function values for X_1 and X_2. It is worth noticing that a coherence measure based on the joint probability, i.e. $C(X) = P\left(\bigwedge_{x_i \in X} x_i\right)$ would master Akiba's test case. But as Olsson (2013) has shown this would be an implausible proposal for a probabilistic coherence measure. It is also worth noticing that these results does not necessarily have to be interpreted as a failure of all coherence measures. Instead, it might indicate that Akiba's coherence assessment could be incorrect. However, as indicated before this question will not be discussed here.

3.2 BonJour's Raven Case

Laurence BonJour's contribution to the systematic development of coherentism cannot be underestimated. His theory of empirical knowledge can be referred to as the cornerstone of modern theories of coherentist justification. In his seminal *The Structure of Empirical Knowledge* BonJour (1985) confronts us with an example, that has often been used to demonstrate a set of coherent versus a set of incoherent propositions. Bovens and Hartmann (2003) developed a probability distribution for this example to the effect that it can serve as a test case for probabilistic coherence measures. Consider the following two sets of propositions under the respective joint probability distributions shown in the diagrams:

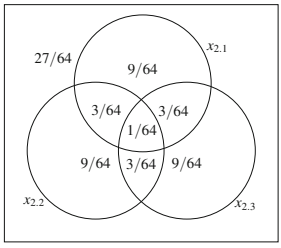
Situation 1:

$x_{1.1}$: All ravens are black.
$x_{1.2}$: This bird is a raven.
$x_{1.3}$: This bird is black.



Situation 2:

$x_{2.1}$: This chair is brown.
$x_{2.2}$: Electrons are negatively charged.
$x_{2.3}$: Today is Thursday.



Since the set $X_2 = \{x_{2.1}, x_{2.2}, x_{2.3}\}$ consists of propositions that have nothing to do with each other, whereas the propositions in $X_1 = \{x_{1.1}, x_{1.2}, x_{1.3}\}$ are tied

together by relevance or entailment relations, BonJour in his original example as well as Bovens and Hartmann in their probabilistic version argue that X_2 should be less coherent than X_1 (Table 3).

As the table indicates, only two measures fail in this test case, namely the average mutual support measures based on Good's likelihood-ratio measure and on Gaifman's evidential support measure. Both measures have non-defined function values for X_1. All the other measures are doing a good job. Moreover notice that as a further plus all measures assign values indicating incoherence or neutral coherence to X_2.

3.3 Bovens and Hartmann's Tweety Case

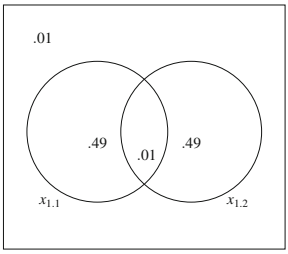
In their *Bayesian Epistemology* Bovens and Hartmann (2003) have presented a variety of test cases for probabilistic coherence measures. One of them is a variation of the classic Tweety example which is often discussed in the context of logics of non-monotonic reasoning (cf. Brewka 1991). In Bovens and Hartmann's version this case is an example of how adding a piece of information to an existing set of information can increase the coherence of all the information taken together. Moreover, this test case aims at pointing out a general problem of the Glass–Olsson coherence measure which will become obvious. The test case runs as follows: imagine a pet named "Tweety" and consider the following two situations in which you receive information about Tweety with probabilities according to the diagrams:

Table 3 Results for BonJour's raven case

	X_1	X_2	Score
C_{sho}	3.72	1	1
C_{sch}	0.334	0	1
C_{go}	0.108	0.027	1
C_{mei}	0.176	0.114	1
$C_{S_{fit}}$	0.402	0	1
$C_{S_{car}}$	0.207	0	1
$C_{S_{key}}$	2.15	1	1
$C_{S_{goo}}$	NaN	1	0
$C_{S_{roc}}$	0.42	0.203	1
$C_{S_{sch}}$	0.292	0	1
$C_{S_{car'}}$	0.0299	0	1
$C_{S_{noz}}$	0.232	0	1
$C_{S_{pop}}$	0.336	0	1
$C_{S_{res}}$	0.0376	0	1
$C_{S_{cru}}$	0.269	0	1
$C_{S_{gai}}$	NaN	1.17	0
$C_{S_{rip}}$	0.315	0.1	1
$C_{S_{sho}}$	0.392	0	1

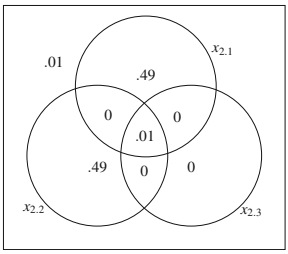
Situation 1:

$x_{1.1}$: Tweety is a bird.
$x_{1.2}$: Tweety is a ground dweller.



Situation 2:

$x_{2.1}$: Tweety is a bird.
$x_{2.2}$: Tweety is a ground dweller.
$x_{2.3}$: Tweety is a penguin.



According to Bovens and Hartmann, the set $X_2 = \{x_{2.1}, x_{2.2}, x_{2.3}\}$ should be judged more coherent than the set $X_1 = \{x_{1.1}, x_{1.2}\}$ since given our background knowledge about penguins the information that Tweety is a penguin entails that Tweety is a bird and that Tweety is a ground dweller. The values for all measures are as follows (Table 4).

Quite obvious, this test case is a problem for Glass' and Olsson's measure since the it treats both sets of propositions as equally coherent. It is therefore reasonable to prefer Meijs' generalized version of the Glass–Olsson measure as the proper generalization since it masters the test case. As before, the two measures based on Good's and Gaifman's support measures fail in this test case because they have non-defined function values for X_2.

3.4 Bovens and Hartmann's Tokyo Murder Case

Another test case from Bovens and Hartmann (2003) is more extensive. In contrast to the preceeding cases this one provides five different situations and three different coherence assessments. Imagine the following scenario: a murder has occured in Tokyo but the corpse has not been found yet. Draw a grid over the map of the city consisting of 100 numbered squares with each square having the same probability of being the location the corpse is to be found. Now consider the following situations s_i where $i \in \{1, \ldots, 5\}$ in which two independent and equally reliable witnesses make reports $x_{i.1}$ and $x_{i.2}$ about the location of the corpse. The suspected location is a closed interval of the respective square numbers as given in the table below (Table 5).

Bovens and Hartmann give the following intuitive coherence assessments: $X_1 = \{x_{1.1}, x_{1.2}\}$ should be more coherent than $X_2 = \{x_{2.1}, x_{2.2}\}$ or $X_3 = \{x_{3.1}, x_{3.2}\}$. The

Table 4 Results for Bovens and Hartmann's Tweety case

	X_1	X_2	Score
C_{sho}	0.04	4	1
C_{sch}	−1.4	0.168	1
C_{go}	0.0101	0.0101	0
C_{mei}	0.0101	0.0151	1
$C_{S_{fit}}$	−0.96	0.398	1
$C_{S_{car}}$	−0.48	0.255	1
$C_{S_{key}}$	0.04	18	1
$C_{S_{goo}}$	0.0204	NaN	0
$C_{S_{roc}}$	0.02	0.51	1
$C_{S_{sch}}$	−0.98	0.343	1
$C_{S_{car'}}$	−0.24	−0.035	1
$C_{S_{noz}}$	−0.96	0.101	1
$C_{S_{pop}}$	−0.932	0.287	1
$C_{S_{res}}$	−0.48	−0.0733	1
$C_{S_{cru}}$	−0.96	0.343	1
$C_{S_{gai}}$	0.51	NaN	0
$C_{S_{rip}}$	−0.96	0.343	1
$C_{S_{sho}}$	−4.64	−0.224	1

Table 5 Situations for Bovens and Hartmann's Tokyo Murder case

	s_1	s_2	s_3	s_4	s_5
$x_{i.1}$	50–60	22–55	20–61	41–60	39–61
$x_{i.2}$	51–61	55–90	50–91	51–70	50–72

sets $X_4 = \{x_{4.1}, x_{4.2}\}$ and $X_5 = \{x_{5.1}, x_{5.2}\}$ should have similar degrees of coherence. Let us take a look at the results (Table 6).

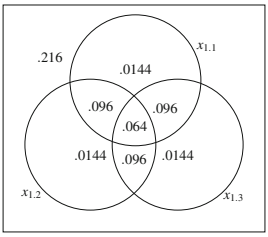
Apparently, every coherence measure masters this test case. We can therefore turn to the next case. Nevertheless, notice that it would have been possible and desirable to have more coherence assessments than the ones provided by Bovens and Hartmann. For instance, they could have indicated, in which of the situations the reports are the most and the least coherent.

3.5 Bovens and Hartmann's Culprit Case

A third test case by Bovens and Hartmann (2003) runs as follows: imagine that we would like to identify a culprit in a murder case. Now consider the following three situations in which we are confronted with reports from independent and equally reliable witnesses:

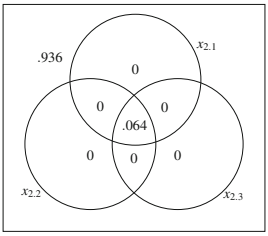
Situation1:

$x_{1.1}$: The culprit was a woman.
$x_{1.2}$: The culprit had a Danish accent.
$x_{1.3}$: The culprit drove a Ford.



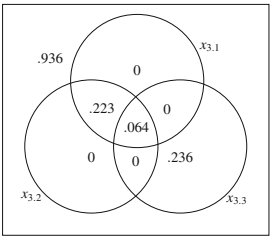
Situation 2:

$x_{2.1}$: The culprit wore Coco Chanel shoes.
$x_{2.2}$: The culprit had a French accent.
$x_{2.3}$: The culprit drove a Renault.



Situation 3:

$x_{3.1}$: The culprit wore Coco Chanel shoes.
$x_{3.2}$: The culprit had a French accent.
$x_{3.3}$: The culprit drove a Ford.



According to Bovens and Hartmann, the set $X_1 = \{x_{1.1}, x_{1.2}, x_{1.3}\}$ is less coherent than the set $X_2 = \{x_{2.1}, x_{2.2}, x_{2.3}\}$ since in the second situation the reports fit together better than in the first. The more interesting set however is $X_3 = \{x_{3.1}, x_{3.2}, x_{3.3}\}$ because it seems unclear whether it is more or less coherent compared to the other sets. In one respect, $x_{3.1}$ and $x_{3.2}$ fit together very well, but in the other $x_{3.3}$ does not really fit together with $x_{3.1}$ or $x_{3.2}$. However, since Bovens and Hartmann suspend judgement with respect to X_3 we follow them and only present the values for this set but do not take them into consideration when evaluating the measures (Table 7).

Again, the two average mutual support measures based on Good's and Gaifman's probabilistic measures of support fail in this test case due to non-defined function values for X_2 and X_3. Every other coherence measure masters the test case. We can therefore turn to the next one.

3.6 Glass' Dodecahedron Case

Glass (2005) has offered a variation of the aforementioned die case by Akiba (2000). His intention here is to point out the difficulty that most probabilistic measures of coherence heavily depend on unconditional probabilities. Glass argues that when assessing the coherence of some set of propositions the relations holding between the propositions that are given by their conditional probabilities are more

Table 6 Results for Bovens and Hartmann's Tokyo murder case

	X_1	X_2	X_3	X_4	X_5	Score
C_{sho}	8.26	0.0817	0.68	2.5	2.27	1
C_{sch}	0.917	−1.09	−0.167	0.398	0.356	1
C_{go}	0.833	0.0145	0.167	0.333	0.353	1
C_{mei}	0.833	0.0145	0.167	0.333	0.353	1
$C_{S_{fit}}$	0.976	−0.896	−0.288	0.6	0.57	1
$C_{S_{car}}$	0.799	−0.321	−0.134	0.3	0.292	1
$C_{S_{key}}$	8.26	0.0817	0.68	2.5	2.27	1
$C_{S_{goo}}$	80.9	0.0547	0.552	4	3.65	1
$C_{S_{roc}}$	0.909	0.0286	0.286	0.5	0.522	1
$C_{S_{sch}}$	0.908	−0.945	−0.448	0.429	0.442	1
$C_{S_{car'}}$	0.0879	−0.112	−0.0564	0.06	0.0671	1
$C_{S_{noz}}$	0.898	−0.494	−0.232	0.375	0.379	1
$C_{S_{pop}}$	0.863	−0.857	−0.213	0.471	0.435	1
$C_{S_{res}}$	0.0988	−0.173	−0.0972	0.075	0.0871	1
$C_{S_{cru}}$	0.898	−0.918	−0.32	0.375	0.379	1
$C_{S_{gai}}$	9.79	0.669	0.812	1.6	1.61	1
$C_{S_{rip}}$	0.898	−0.495	−0.232	0.375	0.379	1
$C_{S_{sho}}$	0.957	−2.39	−0.444	0.569	0.557	1

Table 7 Results for Bovens and Hartmann's culprit case

	X_1	X_2	X_3	Score
C_{sho}	1	244	1.78	1
C_{sch}	−9.64e−17	1.79	0.121	1
C_{go}	0.0816	1	0.101	1
C_{mei}	0.208	1	0.325	1
$C_{S_{fit}}$	−1.73e−17	1	0.298	1
$C_{S_{car}}$	−2.78e−17	0.936	0.177	1
$C_{S_{key}}$	1	15.6	1.93	1
$C_{S_{goo}}$	1	NaN	NaN	0
$C_{S_{roc}}$	0.34	1	0.462	1
$C_{S_{sch}}$	−4.63e−17	1	0.0817	1
$C_{S_{car'}}$	−1.39e−17	0.0599	0.0219	1
$C_{S_{noz}}$	−1.39e−17	1	0.244	1
$C_{S_{pop}}$	−4.02e−17	0.936	0.162	1
$C_{S_{res}}$	−1.85e−17	0.064	0.0253	1
$C_{S_{cru}}$	−6.94e−17	1	0.127	1
$C_{S_{gai}}$	1.33	NaN	NaN	0
$C_{S_{rip}}$	0.143	1	0.372	1
$C_{S_{sho}}$	−8.4e−17	1	0.104	1

important than their unconditional probabilities. His test case runs as follows. Imagine two situations:

Situation 1: A fair die is rolled. Consider the following predictions:

$x_{1.1}$: The die will come up 2.
$x_{1.2}$: The die will come up 2 or 4.

Situation 2: A fair dodecahedron is rolled. Consider the following predictions:

$x_{2.1}$: The dodecahedron will come up 2.
$x_{2.2}$: The dodecahedron will come up 2 or 4.

The main difference between both situations is that the unconditional probabilities of the predictions have changed. However, Glass' intuition is that the coherence of the sets $X_1 = \{x_{1.1}, x_{1.2}\}$ and $X_2 = \{x_{2.1}, x_{2.2}\}$ should be equal (Table 8).

This test case seems is a problem for all coherence measures except the Glass–Olsson measure, its generalized version suggested by Meijs and the average mutual support measures proposed by Roche and Schippers. All other measures do not satisfy Glass' coherence assessment. Notice that again the two measures based on Good's and Gaifman's evidential support measures fail due to non-defined function values for X_1 and X_2.

3.7 Meijs' Samurai Sword Case

Meijs (2005) has provided a test case in which a set of propositions has to be evaluated in two different situations. Meijs' is not particularly precise about the intuition behind this test case. However, it seems to based on the assumption that the coherence of a set of proposition should be influenced by the propositions relative overlap. The test case is based on the following scenario: imagine that a murder occurred in a big city and we are interested in finding the murderer. The two situations are as follows:

Situation 1: There are ten million independent and equally likely suspects. 1059 suspects are Japanese, 1059 suspects own a Samurai sword, nine suspects are Japanese and own a Samurai sword. Now consider the following two propositions:

$x_{1.1}$: The murderer is Japanese.
$x_{1.2}$: The murderer owns a Samurai sword.

Situation 2: There are 100 independent and equally likely suspects. Ten suspects are Japanese, ten suspects own a Samurai sword, nine suspects are Japanese and own a Samurai sword. Again, consider the two propositions:

$x_{2.1}$: The murderer is Japanese.
$x_{2.2}$: The murderer owns a Samurai sword.

Table 8 Results for Glass' dodecahedron case

	X_1	X_2	Score
C_{sho}	3	6	0
C_{sch}	0.477	0.778	0
C_{go}	0.5	0.5	1
C_{mei}	0.5	0.5	1
$C_{S_{fit}}$	0.833	0.917	0
$C_{S_{car}}$	0.5	0.625	0
$C_{S_{key}}$	3	6	0
$C_{S_{goo}}$	NaN	NaN	0
$C_{S_{roc}}$	0.5	0.5	1
$C_{S_{sch}}$	0.75	0.75	1
$C_{S_{car'}}$	0.111	0.0694	0
$C_{S_{noz}}$	0.65	0.705	0
$C_{S_{pop}}$	0.604	0.789	0
$C_{S_{res}}$	0.15	0.0795	0
$C_{S_{cru}}$	0.7	0.727	0
$C_{S_{gai}}$	NaN	NaN	0
$C_{S_{rip}}$	0.7	0.727	0
$C_{S_{sho}}$	0.807	0.861	0

According to Meijs' relative overlap intuition the set $X_1 = \{x_{1.1}, x_{1.2}\}$ is less coherent than the set $X_2 = \{x_{2.1}, x_{2.2}\}$. As we can see in the function values, only few coherence measures are not in accordance with this coherence assessment (Table 9).

Quite obviously, Shogenji's measure fails in this test case. And since both X_1 and X_2 contain 2 propositions, the average mutual support measure based on Keynes' relevance quotient is identical to Shogenji's coherence measure and must therefore also fail. Moreover, since in the case of 2 propositions Schupbach's measure is ordinally equivalent to Shogenji's measure being a simple log-transformation, Schupbach's measure must fail, too. Furthermore, the average mutual support measure based on Popper's measure of evidential support does not master this test case, either. It is worth noticing that despite the relative overlap intuition the test case is driven by the test case is also mastered by many average mutual support measures.

3.8 Meijs' Albino Rabbit Case

Meijs (2006) has constructed a test case in order to show that Fitelson's measure of coherence provides counter-intuitive results for certain sets of propositions. The test case runs as follows: imagine a population of 102 rabbits living on an island and consider the following two situations:

Table 9 Results for Meijs' samurai sword case

	X_1	X_2	Score
C_{sho}	80.3	9	0
C_{sch}	1.9	0.954	0
C_{go}	0.00427	0.818	1
C_{mei}	0.00427	0.818	1
$C_{S_{fit}}$	0.9756	0.9761	1
$C_{S_{car}}$	0.00839	0.8	1
$C_{S_{key}}$	80.3	9	0
$C_{S_{goo}}$	80.9	81	1
$C_{S_{roc}}$	0.0085	0.9	1
$C_{S_{sch}}$	0.00839	0.899	1
$C_{S_{car'}}$	8.89e−07	0.08	1
$C_{S_{noz}}$	0.00839	0.889	1
$C_{S_{pop}}$	0.975	0.872	0
$C_{S_{res}}$	8.89e−07	0.0889	1
$C_{S_{cru}}$	0.00839	0.889	1
$C_{S_{gai}}$	1.01	9	1
$C_{S_{rip}}$	0.00839	0.889	1
$C_{S_{sho}}$	0.479	0.954	1

Situation 1: 101 rabbits are grey, 101 rabbits have two ears and 100 rabbits are grey and have two ears. Randomly pick one of the rabbits and consider the following two propositions:

$x_{1.1}$: The rabbit is grey.
$x_{1.2}$: The rabbit has two ears.

Situation 2: 100 rabbits are grey, 100 rabbits have two ears and 100 rabbits are grey and have two ears. Randomly pick one of the rabbits and consider the same two propositions:

$x_{2.1}$: The rabbit is grey.
$x_{2.2}$: The rabbit has two ears.

Since the set $X_2 = \{x_{2.1}, x_{2.2}\}$ consist of logically equivalent propositions, Meijs argues that it is more coherent than the set $X_1 = \{x_{1.1}, x_{1.2}\}$. Nevertheless, the set X_1 is not so different from X_2 regarding its degree of coherence since the propositions in X_1 still have a high joint probability due to a high absolute overlap of two-eared rabbits in situation 1 (Table 10).

Apparently, this case is not only a problem for Fitelson's measure. Meijs' test case is a problem for all coherence measures except the Glass–Olsson measure, its

Table 10 Results for Meijs' albino rabbit case

	X_1	X_2	Score
C_{sho}	0.999	1.02	0
C_{sch}	−4.26e−05	0.0086	0
C_{go}	0.98	1	1
C_{mei}	0.98	1	1
$C_{S_{fit}}$	−0.00498	1	0
$C_{S_{car}}$	−9.71e−05	0.0196	0
$C_{S_{key}}$	1	1.02	0
$C_{S_{goo}}$	0.99	NaN	0
$C_{S_{roc}}$	0.99	1	1
$C_{S_{sch}}$	−0.0099	1	0
$C_{S_{car'}}$	−9.61e−05	0.0192	0
$C_{S_{noz}}$	−0.0099	1	0
$C_{S_{pop}}$	−9.71e−05	0.0196	0
$C_{S_{res}}$	−0.0098	0.98	0
$C_{S_{cru}}$	−9.8e−05	1	0
$C_{S_{gai}}$	0.99	NaN	0
$C_{S_{rip}}$	−0.0099	1	0
$C_{S_{sho}}$	−0.00995	1	0

alternative generalization by Meijs' and the average mutual measure by Roche. Every other coherence measure assesses X_1 as incoherent, which is counter-intuitive. To see this more clearly, simply inspect the neutrality values t from Table 1. Notice that again the two average mutual measures based on Good's and Gaifman's measures do not master the test case due to non-defined function values for X_2.

3.9 Meijs and Douven's Plane Lottery Case

Meijs and Douven (2005) have developed a rather complicated test case in which a person named "Kate" participates in a lottery. She enters a windowless plane that either flies to the North Pole, the South Pole or New Zealand. Kate's chances are as follows: $4/100$ for flying to the North Pole, $49/100$ for flying to the South Pole and $47/100$ for flying to New Zealand. The probability of seeing a penguin given she is on the South Pole is $10/49$, given she is in New Zealand is $1/47$ and given she is on the North Pole is 0. Suppose that after the random flight Kate leaves the plane not knowing where she has landed. She faces two equally reliable people and an animal she is unable to recognize. Now consider the following two situations, in which the two people independently provide the following information:

Situation 1:

$x_{1.1}$: The animal you see is a penguin.
$x_{1.2}$: You are on the North Pole.

Situation 2:

$x_{2.1}$: The animal you see is a penguin.
$x_{2.2}$: You are on the South Pole.

According to Meijs and Douven, the set $X_2 = \{x_{2.1}, x_{2.2}\}$ is more coherent than the set $X_1 = \{x_{1.1}, x_{1.2}\}$ since there are no penguins on the Northpole (Table 11).

This test case is a problem for two measures, namely Schupbach's coherence measure and the average mutual support measure based on Shogenji's measure of epistemic justification. Both measures fail because they use logarithms to normalize their function values but it is clear that $\lim_{x \to 0}(\log(x)) = -\infty$ is not a defined function value. It is also worth noticing that Schupbach's measure was supposed to overcome certain difficulties of Shogenji's coherence measure. In this case surprisingly Shogenji's coherence measure masters the test case while Schupbach's measure does not. Hence, the log-normalization of Schupbach's measure could be dropped in favour of a different kind of normalization.

3.10 Schupbach's Robber Case

Schupbach (2011) has presented a test case inspired by Fitelson's (2003) criticism against Shogenji's measure of coherence. The test case is supposed to show that Shogenji's measure has flaws due to the way it is generalized for sets containing

Table 11 Results for Meijs and Douven's plane lottery case

	X_1	X_2	Score
C_{sho}	0	1.86	1
C_{sch}	NaN	0.268	0
C_{go}	0	0.2	1
C_{mei}	0	0.2	1
$C_{S_{fit}}$	−1	0.587	1
$C_{S_{car}}$	−0.075	0.257	1
$C_{S_{key}}$	0	1.86	1
$C_{S_{goo}}$	0	6.24	1
$C_{S_{roc}}$	0	0.557	1
$C_{S_{sch}}$	−1	0.513	1
$C_{S_{car'}}$	−0.0044	0.0461	1
$C_{S_{noz}}$	−0.0798	0.328	1
$C_{S_{pop}}$	−1	0.37	1
$C_{S_{res}}$	−0.00476	0.0711	1
$C_{S_{cru}}$	−1	0.464	1
$C_{S_{gai}}$	0.925	3.36	1
$C_{S_{rip}}$	−0.0826	0.464	1
$C_{S_{sho}}$	NaN	0.573	0

more than two propositions. In order to overcome this problem Schupbach has offered his own alternative generalization of Shogenji's measure. The test case runs as follows: imagine eight suspects, each having the same probability of having committed a robbery. Now consider the following two situations in which three independent and equally reliable witnesses make reports about the possible robber:

Situation 1:

$x_{1.1}$: The robbery was committed by suspect 1, 2 or 3.
$x_{1.2}$: The robbery was committed by suspect 1, 2 or 4.
$x_{1.3}$: The robbery was committed by suspect 1, 3 or 4.

Situation 2:

$x_{2.1}$: The robbery was committed by suspect 1, 2 or 3.
$x_{2.2}$: The robbery was committed by suspect 1, 4 or 5.
$x_{2.3}$: The robbery was committed by suspect 1, 6 or 7.

According to Schupbach, $X_1 = \{x_{1.1}, x_{1.2}, x_{1.3}\}$ is more coherent than $X_2 = \{x_{2.1}, x_{2.2}, x_{2.3}\}$ since the agreement about who is the robber is much stronger in the first situation. Let us take a look at the measures' verdicts (Table 12).

As intended by Schupbach, this test case is a problem for Shogenji's coherence measure since the measure treats both X_1 and X_2 as equal regarding their coherence. Schupbach's alternative generalization of Shogenji's measure, however, does the trick just like most of the other measures. Quite surprisingly, the average mutual support measure based on Keynes' relevance quotient even treats X_2 as more coherent than the X_1. Again, as in several other test cases, the average mutual support measures based on Good's and Gaifman's support measures do not master the test case due to non-defined function values for X_1 and X_2.

3.11 Siebel's Pickpocketing Robber Case

The last test case is due to Siebel (2004). It is supposed to point out a general problem for Fitelson's average mutual support measure based on a variation of Kemeny and Oppenheim's factual support measure. Siebel's intuition here is that propositions which cannot be false together in a certain situation can nevertheless be coherent. The test case is rather simple. Imagine the following situation:

Situation: There are ten independent and equally likely suspects for a murder. Eight suspects committed a robbery, eight suspects committed a pickpocketing and six committed both. Now consider the following two propositions:

x_1: The murderer committed a robbery.
x_2: The murderer committed a pickpocketing.

Table 12 Results for Schupbach's robber case

	X_1	X_2	Score
C_{sho}	2.37	2.37	0
C_{sch}	0.312	0.162	1
C_{go}	0.25	0.143	1
C_{mei}	0.438	0.186	1
$C_{S_{fit}}$	0.582	0.255	1
$C_{S_{car}}$	0.198	0.188	1
$C_{S_{key}}$	1.56	1.78	0
$C_{S_{goo}}$	NaN	NaN	0
$C_{S_{roc}}$	0.542	0.5	1
$C_{S_{sch}}$	0.458	0.25	1
$C_{S_{car'}}$	0.0703	0.0312	1
$C_{S_{noz}}$	0.392	0.133	1
$C_{S_{pop}}$	0.256	0.242	1
$C_{S_{res}}$	0.11	0.0411	1
$C_{S_{cru}}$	0.311	0.254	1
$C_{S_{gai}}$	NaN	NaN	0
$C_{S_{rip}}$	0.533	0.276	1
$C_{S_{sho}}$	0.419	0.308	1

Since there is a big absolute overlap of pickpocketing robbers, Siebel sees no reason why the set $X = \{x_1, x_2\}$ should be judged incoherent. Albeit, apparently most measures violate this intuition (Table 13).

These results are similar to the ones for Meijs' albino rabbit case. The only measures mastering the test case are the Glass–Olsson measure, Meijs' alternative generalization of this measure and Roche's average mutual absolute support measure. All other measures fail due to the fact that they judge the set X to be incoherent. This can be seen inspecting the neutrality values t from Table 1.

3.12 Results

In the previous subsections a collection of 18 alleged probabilistic coherence measures have been investigated with respect to their performances in 11 test cases. The results for each measure in each test case T_i are presented in the following table. As before, a score of 1 indicates a positive test case result while 0 indicates a negative (Table 14).

The information provided by this table are twofold. First, the table indicates which measures are the most successful. To find these measures, simply inspect the lines for a low number of zeros or a high number of ones. The most successful measures are Meijs' (2005) generalized version of the Glass–Olsson measure (cf. Glass 2002; Olsson 2002) and Roche's (2013) average mutual support measure

Table 13 Results for Siebel's pickpocketing robber case

	X	Score
C_{sho}	0.937	0
C_{sch}	−0.028	0
C_{go}	0.6	1
C_{mei}	0.6	1
$C_{S_{fit}}$	−0.143	0
$C_{S_{car}}$	−0.05	0
$C_{S_{key}}$	0.937	0
$C_{S_{goo}}$	0.75	0
$C_{S_{roc}}$	0.75	1
$C_{S_{sch}}$	−0.25	0
$C_{S_{car'}}$	−0.04	0
$C_{S_{noz}}$	−0.25	0
$C_{S_{pop}}$	−0.0516	0
$C_{S_{res}}$	−0.2	0
$C_{S_{cru}}$	−0.0625	0
$C_{S_{gai}}$	0.8	0
$C_{S_{rip}}$	−0.25	0
$C_{S_{sho}}$	−0.289	0

based on a case-sensitive variation of the posterior probability. The weakest measures are the two average mutual support measures based on Good's (1984) likelihood-ratio measure and on Gaifman's (1979) measure of evidential support. Second, the table indicates which test cases rule out the most measures. To find these, simply inspect the columns for a low number of zero or a high number of ones. The test cases in which most measures fail are Akiba's (2000) die case, Meijs' (2005) albino rabbit case, Glass' (2005) dodecahedron case and Siebel's (2004) pickpocketing robber case.

To summarize the results of the antecedent investigation and in order to have a rough quantitative overview of the overall performance of each measure we calculated the relative score of each measure which is simply defined as the number of mastered test cases divided by the total number of test cases. This score is represented by the y-axis of the bar plot below. Calculating the relative score in this manner, however, relies on the assumption that all test cases including their corresponding coherence assessment are equally plausible since they has the same impact on the relative score. This problem has already been mentioned in Sect. 3. Here is a tentative approach to weaken this problem. Each out of n test cases T_i can be assigned a weight $w_i \in [0, 1]$ according to its plausibility such that $\sum_{i=1}^{n} w_i = 1$. The values of the weights can then be adapted according to further philosophical considerations regarding the plausibility of the provided coherence assessments. For instance, the weight for a certain test case could be chosen depending on the similarity to other test cases and the number of such cases. The values can also be adapted according to empirical findings in cognitive-psychological tasks of

Table 14 Summary of the results

	T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8	T_9	T_{10}	T_{11}
C_{sho}	0	1	1	1	1	0	0	0	1	0	0
C_{sch}	0	1	1	1	1	0	0	0	0	1	0
C_{go}	0	1	0	1	1	1	1	1	1	1	1
C_{mei}	0	1	1	1	1	1	1	1	1	1	1
$C_{S_{fit}}$	0	1	1	1	1	0	1	0	1	1	0
$C_{S_{car}}$	0	1	1	1	1	0	1	0	1	1	0
$C_{S_{key}}$	0	1	1	1	1	0	0	0	1	0	0
$C_{S_{goo}}$	0	0	0	1	0	0	1	0	1	0	0
$C_{S_{roc}}$	0	1	1	1	1	1	1	1	1	1	1
$C_{S_{sch}}$	0	1	1	1	1	1	1	0	1	1	0
$C_{S_{car'}}$	0	1	1	1	1	0	1	0	1	1	0
$C_{S_{noz}}$	0	1	1	1	1	0	1	0	1	1	0
$C_{S_{pop}}$	0	1	1	1	1	0	0	0	1	1	0
$C_{S_{res}}$	0	1	1	1	1	0	1	0	1	1	0
$C_{S_{cru}}$	0	1	1	1	1	0	1	0	1	1	0
$C_{S_{gai}}$	0	0	0	1	0	0	1	0	1	0	0
$C_{S_{rip}}$	0	1	1	1	1	0	1	0	1	1	0
$C_{S_{sho}}$	0	1	1	1	1	0	1	0	0	1	0

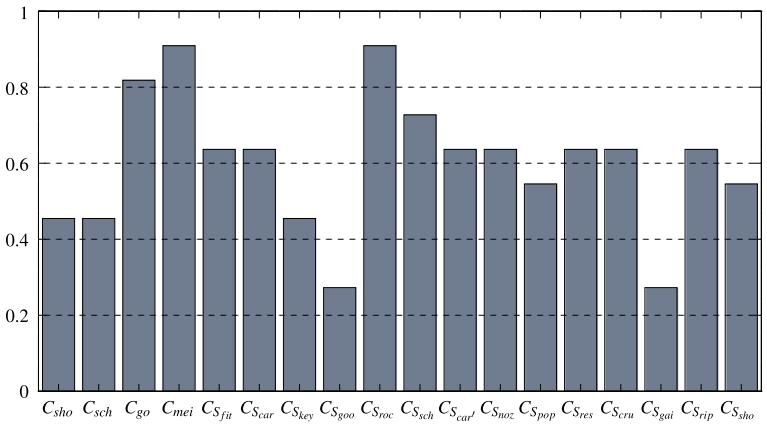


Fig. 1 Relative scores for all tested coherence measures

coherence assessments such as examined by Harris and Hahn (2009) or Jekel and Koscholke (2013) which showed that lay people have quite strong coherence intuitions when facing a test case like the ones presented above. This weighting procedure can thus be considered a promising approach to making the relative score both philosophically and empirically more accurate and to allow for incorporation

of future research. The plot below, however, shows the relative scores where each weight $w_i = w_j$ for $i,j \leq n$ and concludes this section (Fig. 1).

4 Conclusion

The antecedent evaluation clearly indicates that there are two measures standing out from the crowd, namely Meijs' (2006) generalized relative overlap measure and Roche's (2013) average mutual support measure based on a case-sensitive notion of absolute support. These two measures outperform other prominent probabilistic coherence measures such as Shogenji's (1999) deviation from independence measure, Glass' (2002) and Olsson's (2002) relative overlap measure, Fitelson's (2004) coherence measure based on a variation of Kemeny and Oppenheim's (1952) measure of factual support and Douven and Meijs' (2007) favourite average mutual support measure based on Carnap's (1950) difference measure of support.

Nevertheless, we need to be be cautious with respect to the conclusions to draw from this evaluation. First, because all the results presented above rely on the assumption that each test case together with its corresponding coherence assessment is equally plausible. This assumption, as pointed out, has to be examined in future research. Still, the weighting approach suggested for the relative score of each measure enables us to incorporate future results. Second, we have to be cautious because this evaluation is not the last word on probabilistic coherence measures. Investigating the test case performance of the considered measures is only one component of evaluating their adequacy. Another component is the analysis adequacy constraints satisfied or violated by each measure (for such an overview cf. Schippers 2014). Yet another component is the investigation of their empirical adequacy, i.e. their ability to capture lay people's coherence intuitions (cf. Harris and Hahn 2009; Jekel and Koscholke 2013). Hence, this paper is not a final verdict on the adequacy of the investigated measures. It is a contribution to the enterprise of finding adequate probabilistic coherence measures.

Acknowledgments I would like to thank Michael Schippers and Mark Siebel for helpful comments or discussion. This work was supported by grant SI 1731/1-1 to Mark Siebel from the Deutsche Forschungsgemeinschaft (DFG) as part of the priority program "New Frameworks of Rationality" (SPP 1516).

References

Akiba, K. (2000). Shogenji's probabilistic measure of coherence is incoherent. *Analysis*, *60*, 356–359.

BonJour, L. (1985). *The structure of empirical knowledge*. Cambridge: Harvard University Press.

Bovens, L., & Hartmann, S. (2003). *Bayesian epistemology*. Oxford: Oxford University Press.

Bovens, L., & Hartmann, S. (2005). Why there cannot be a single probabilistic measure of coherence. *Erkenntnis*, *63*, 361–374.

Brewka, G. (1991). *Nonmonotonic reasoning: Logical foundations of commonsense*. Cambridge: Cambridge University Press.

Carnap, R. (1950). *Logical foundations of probability*. Chicago: University of Chicago Press.

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review*, *104*, 367–405.

Christensen, D. (1999). Measuring confirmation. *Journal of Philosophy*, *96*, 437–461.
Crupi, V., Tentori, K., & Gonzales, M. (2007). On Bayesian measures of evidential support: Theoretical and empirical issues. *Philosophy of Science*, *74*, 229–252.
Douven, I., & Meijs, W. (2007). Measuring coherence. *Synthese*, *156*, 405–425.
Eells, E., & Fitelson, B. (2002). Symmetries and asymmetries in evidential support. *Philosophical Studies*, *107*, 129–142.
Finch, H. A. (1960). Confirming power of observations metricized for decisions among hypotheses. *Philosophy of Science*, *27*, 293–307.
Fitelson, B. (2003). A probabilistic theory of coherence. *Analysis*, *63*, 194–199.
Fitelson, B. (2004). Two technical corrections to my coherence measure. http://fitelson.org/coherence2.
Gaifman, H. (1979). Subjective probability, natural predicates and Hempel's ravens. *Erkenntnis*, *21*, 105–147.
Glass, D. H. (2002). Coherence, explanation, and Bayesian networks. In M. O'Neill, R. F. E. Sutcliffe, C. Ryan, M. Eaton & N. J. L. Griffith (Eds.), *Artificial intelligence and cognitive science* (pp. 177–182). *13th Irish conference*, AICS 2002, Limerick, Ireland, September 2002. Berlin: Springer.
Glass, D. H. (2005). Problems with priors in probabilistic measures of coherence. *Erkenntnis*, *63*, 375–385.
Good, I. J. (1984). The best explicatum for weight of evidence. *Journal of Statistical Computation and Simulation*, *19*, 294–299.
Harris, A., & Hahn, U. (2009). Bayesian rationality in evaluating multiple testimonies: Incorporating the role of coherence. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(5), 1366–1373.
Jekel, M., & Koscholke, J. (2013). An empirical study of coherence assessment (unpublished manuscript).
Joyce, J. (2008). Bayes' theorem. http://plato.stanford.edu/archives/fall2008/entries/bayes-theorem/.
Kemeny, J., & Oppenheim, P. (1952). Degrees of factual support. *Philosophy of Science*, *1952*, 307–324.
Keynes, J. (1921). *A treatise on probability*. London: Macmillan.
Kuipers, T. A. F. (2000). *From instrumentalism to constructive realism*. Dordrecht: Reidel.
Levi, I. (1962). Corroboration and rules of acceptance. *British Journal for the Philosophy of Science*, *13*, 307–313.
Meijs, W. (2005). *Probabilistic measures of coherence*. PhD thesis, Erasmus University, Rotterdam.
Meijs, W. (2006). Coherence as generalized logical equivalence. *Erkenntnis*, *64*, 231–252.
Meijs, W., & Douven, I. (2005). Bovens and Hartmann on coherence. *Mind*, *114*, 355–363.
Moretti, L., & Akiba, K. (2007). Probabilistic measures of coherence and the problem of belief individuation. *Synthese*, *154*, 73–95.
Mortimer, H. (1988). *The logic of induction*. Paramus: Prentice Hall.
Nozick, R. (1981). *Philosophical explanations*. Oxford: Clarendon.
Olsson, E. J. (2002). What is the problem of coherence and truth? *The Journal of Philosophy*, *94*, 246–272.
Olsson, E. J. (2005). *Against coherence: Truth, probability and justification*. Oxford: Oxford University Press.
Olsson, E. J. (2013). Coherentist theories of epistemic justification. http://plato.stanford.edu/entries/justep-coherence/.
Popper, K. R. (1954). Degree of confirmation. *British Journal for the Philosophy of Science*, *5*, 143–149.
Rescher, N. (1958). Theory of evidence. *Philosophy of Science*, *25*, 83–94.
Rescher, N. (1973). *The coherence theory of truth*. Oxford: Oxford University Press.
Rips, L. J. (2001). Two kinds of reasoning. *Psychological Science*, *12*, 129–134.
Roche, W. (2013). Coherence and probability: A probabilistic account of coherence. In M. Araszkiewicz & J. Savelka (Eds.), *Coherence: Insights from philosophy, jurisprudence and artificial intelligence* (pp. 59–91). Dordrecht: Springer.
Schippers, M. (2014). Probabilistic measures of coherence: From adequacy constraints towards pluralism. *Synthese*, *191*(16), 3821–3845.
Schupbach, J. N. (2011). New hope for Shogenji's coherence measure. *British Journal for the Philosophy of Science*, *62*(1), 125–142.
Shogenji, T. (1999). Is coherence truth conducive? *Analysis*, *59*, 338–345.
Shogenji, T. (2012). The degree of epistemic justification and the conjunction fallacy. *Synthese*, *184*, 29–48.

Siebel, M. (2004). On Fitelson's measure of coherence. *Analysis*, *64*, 189–190.
Siebel, M. (2005). Against probabilistic measures of coherence. *Erkenntnis*, *63*, 335–360.
Siebel, M., & Wolff, W. (2008). Equivalent testimonies as a touchstone of coherence measures. *Synthese*, *161*, 167–182.
Tentori, K., Crupi, V., Bonini, N., & Osherson, D. (2007). Comparison of confirmation measures. *Cognition*, *103*, 107–119.