Synthese (2017) 194:1303–1322
DOI 10.1007/s11229-015-0996-6

Probabilistic coherence measures: a psychological study of coherence assessment

Jakob Koscholke[1] · **Marc Jekel**[2]

Received: 1 September 2015 / Accepted: 14 December 2015 / Published online: 11 January 2016

Abstract Over the years several non-equivalent probabilistic measures of coherence have been discussed in the philosophical literature. In this paper we examine these measures with respect to their empirical adequacy. Using test cases from the coherence literature as vignettes for psychological experiments we investigate whether the measures can predict the subjective coherence assessments of the participants. It turns out that the participants' coherence assessments are best described by Roche's (Insights from philosophy, jurisprudence and artificial intelligence, 2013) coherence measure based on Douven and Meijs' (Synthese 156:405–425, 2007) average mutual support approach and the conditional probability.

Keywords Bayesian coherentism · Probabilistic coherence measures · Probabilistic support measures · Test cases · Experimental philosophy

1 Introduction

Loosely speaking, coherence is the property of propositions hanging or fitting together, dovetailing with or mutually supporting each other (cf. BonJour 1985; Olsson 2005). It is the key concept of any coherentist theory of justification or truth. Nevertheless, as many authors have pointed out, coherentists have put only little emphasis on elucidating their central concept—or in Nicholas Rescher's words "the coherence theorists

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

Marc Jekel
marc.jekel@fernuni-hagen.de

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

[2] Institut für Psychologie, FernUniversität in Hagen, 58084 Hagen, Germany

themselves have not always been too successful in explicating the nature of coherence" (Rescher 1973, p. 33). In order to overcome this supposed shortcoming, various philosophers have attempted to provide a mathematically precise explication of the concept coherence using probability theory. The results are so-called *probabilistic coherence measures* (cf. Douven and Meijs 2007; Fitelson 2003, 2004; Glass 2002; Meijs 2006; Olsson 2002; Roche 2013; Schippers 2014; Schupbach 2011; Shogenji 1999). Of course, these measures have to be examined with respect to their claim of measuring coherence adequately. So far, there have been two common ways to do that: (i) formulating adequacy constraints and proving whether they are satisfied by a measure or not, (ii) developing paradigmatic test cases providing an intuitive normative coherence assessment and testing whether a measure is in line with this assessment or not.

This paper concentrates on the second approach, although in a slightly different way. Rather than using test cases and the provided normative coherence assessment as a benchmark for probabilistic coherence measures, the test cases are used as vignettes for psychological experiments, in which participants are asked for subjective coherence assessments of specified sets of propositions. Accordingly, the results of the experiments can be used to (i) evaluate the normative coherence assessments provided by the test cases and to (ii) evaluate the suitability of the tested measures as predictors of the participants' coherence assessments. The paper is structured as follows. In Sect. 2 a collection of probabilistic coherence measures that have been proposed in the literature is introduced. In Sects. 3 and 4 the psychological study including methods and results is described. Finally, Sect. 5 discusses which conclusions can be drawn from the results.

2 Probabilistic measures of coherence

The notion of a probabilistic coherence measure can be introduced formally in a straightforward manner. Let L be a classical propositional language consisting of atomic formulas closed under some functional complete selection of classical logical connectives such as e.g. $\{\neg, \wedge\}$ and let $P : L \rightarrow [0, 1]$ be a probability function over L with conditional probability defined by $P(x_1|x_2) = P(x_1 \wedge x_2)/P(x_2)$ for any $x_2 \in L$ with $P(x_2) \neq 0$. Furthermore, let $2^L_{\geq 2}$ denote the set of all non-empty, non-singleton subsets of L and $\mathbf{P}$ the set of all probability functions over L. A probabilistic coherence measure can then be defined as a partial function $C : 2^L_{\geq 2} \times \mathbf{P} \rightarrow \mathbb{R}$ assigning real numbers to sets of propositions under some joint probability distribution. By contrast, a probabilistic measure of support, on which a coherence measure can be based, is a partial function $S : L \times L \times \mathbf{P} \rightarrow \mathbb{R}$ assigning real numbers to pairs of propositions under some probability distribution where the first argument is commonly interpreted as a hypothesis and the second as a piece of evidence. Notice that we will omit reference to a particular probability function as a separate function argument of coherence or support measures.

Still, these are only very general requirements a probabilistic coherence measure should meet. The question which probabilistic information should be taken into account by a probabilistic coherence measure in order to adequately quantify the

degree of coherence has been answered in different ways leading to different kinds of measures. They can be categorized into three groups: (i) measures that quantify coherence in terms of deviation from probabilistic independence (see Sect. 2.1), (ii) in terms of relative set-theoretic overlap (see Sect. 2.2) and (iii) in terms of average mutual support (see Sect. 2.3). In the following we briefly introduce the approaches and the resulting measures.

2.1 Deviation from independence measures

According to standard textbooks on probability theory (cf. e.g. Kolmogorov 1956), a set X of propositions $x_1, \ldots, x_n$ is said to be n-wise negatively dependent iff $P(x_1 \wedge \ldots \wedge x_n) < P(x_1) \times \ldots \times P(x_n)$, independent iff $P(x_1 \wedge \ldots \wedge x_n) = P(x_1) \times \ldots \times P(x_n)$ and positively dependent iff $P(x_1 \wedge \ldots \wedge x_n) > P(x_1) \times \ldots \times P(x_n)$. This definition can be rearranged by dividing the term on the left hand side by the term on the right hand side. Positive dependence is then defined as a value in the open interval $(1, \infty)$, independence as a value of 1 and negative dependence as a value in the half-open interval $[0, 1)$. This can be considered the basic idea underlying Shogenji's (1999) coherence measure. According to Shogenji, the degree of coherence of a finite set X of propositions can be computed by dividing the joint probability of X's propositions by the product over their marginal probabilities. This quantifies the propositions' ratio-wise deviation from their independence threshold value $\theta = 1$. This value is interpreted as neutrality such that values below θ indicate degrees of incoherence and values above θ indicate degrees of coherence:

$$C_{sho}(X) = \frac{P\left(\bigwedge_{x_i \in X} x_i\right)}{\prod_{x_i \in X} P(x_i)}$$

As Fitelson (2003) and Schupbach (2011) have pointed out, Shogenji's measure suffers from a lack of subset sensitivity when applied to a set of more than two propositions. This is due to the fact that for any set X consisting of n propositions there are probability distributions such that there are subsets of X which are i-wise negatively dependent, independent or dependent but not j-wise negatively dependent, independent or dependent for $i \neq j$ where $i, j \leq n$ (cf. Pfeiffer 1990). Therefore, Schupbach (2011) has suggested the following alternative generalization: to assess the degree of coherence of X, apply a log-normalized version of C_{sho} to each set X'_{ij} which is the i-th subset of X and contains $j \geq 2$ propositions. 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:

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

Although the measure is more fine-grained it is still based on the idea of measuring coherence in terms of their deviation from independence. However, due to the log-

normalization the threshold value of Schupbach's measure for neutrality is $\theta = 0$, such that values in $(-\infty, 0)$ indicate degrees of incoherence and values in $(0, \infty)$ indicate degrees of coherence.

2.2 Relative overlap measures

Glass (2002) and Olsson (2002) have proposed a different measure of coherence that is based on a set-theoretically inspired understanding of coherence. Here, the joint probability over all propositions $x_1 \ldots, x_n$ in some set X is interpreted as their overlapping set-theoretic surface. Likewise, the probability that any of these propositions is true is interpreted as their total set-theoretic surface. In order to compute the degree of coherence of X Glass and Olsson suggest to 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:

$$C_{go}(X) = \frac{P\left(\bigwedge_{x_i \in X} x_i\right)}{P\left(\bigvee_{x_i \in X} x_i\right)}$$

It is easy to see that the measure has the codomain [0, 1] where 0 means no overlap at all and 1 means identity of overlap and total surface of the propositions. But unlike the two measures mentioned before the threshold θ cannot be based on probabilistic independence. One can, however, argue that the threshold is .5 in the case of two propositions x_1, x_2, since values below this threshold would indicate that x_1 coheres better with $\neg x_2$ than with x_2. In any case, Bovens and Hartmann (2003) have shown that this measure has similar problems with respect to subset-sensitivity like Shogenji's measure. In order to overcome these difficulties Meijs (2006) has suggested the following alternative: in order to assess the coherence of X, take the straight average over all C_{go} values applied to every subset X_i' of X with $|X_i'| \geq 2$:

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

This measure is obviously more fine-grained but it is easy to see that the codomain [0, 1] and the threshold θ remain the same.

2.3 Average mutual support measures

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. And since there is a variety of probabilistic measures of support (for overviews cf. Crupi et al. 2007; Festa 2012) 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:

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

For his coherence measure Fitelson (2004) has chosen a case-sensitive variation of Kemeny and Oppenheim's (1952) measure of factual support. The values of the resulting coherence measure are in $[-1, 1]$ with $\theta = 0$:

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

Douven and Meijs (2007) have investigated three further support measures as foundations for probabilistic coherence measures, namely Carnap's (1950) difference measure with codomain $[-1, 1)$ and $\theta = 0$, Keynes' (1921) relevance quotient and Good's (1984) likelihood ratio measure both with codomain $[0, \infty)$ and $\theta = 1$. Notice that due to commutativity or ordinal equivalence (up to identity) one would obtain identical coherence measures if one used Levi's (1962) corroboration measure or Mortimer's (1988) confirmation measure instead of Carnap's difference measure, Kuipers' (2000) confirmation measure or Finch's (1960) confirmation measure +1 instead of Keynes' and finally Joyce's (2008) odds-ratio measure instead of Good's likelihood-ratio measure. Douven and Meijs' favourite is the coherence measure based on Carnap's difference measure:

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

Siebel and Wolff (2008) have extended the collection of candidate measures by taking into account Carnap's (1950) relevance measure with codomain $(-1, 1)$ and $\theta = 0$, Nozick's (1981) counterfactual likelihood difference measure, Popper's (1954) corroboration measure and Rescher's (1958) measure of evidential support, all three with values in $[-1, 1]$ with $\theta = 0$:

$$S_{car'}(x_1, x_2) = P(x_1 \wedge x_2) - P(x_1) \cdot P(x_2)$$
$$S_{noz}(x_1, x_2) = P(x_2|x_1) - P(x_2|\neg x_1)$$
$$S_{pop}(x_1, x_2) = \frac{P(x_2|x_1) - P(x_2)}{P(x_2|x_1) + P(x_2)} \cdot (1 + P(x_1) \cdot P(x_1|x_2))$$
$$S_{res}(x_1, x_2) = \frac{P(x_1|x_2) - P(x_1)}{1 - P(x_1)} \cdot P(x_2)$$

A more recent proposal is due to Roche (2013). His favourite coherence measure is based on Douven and Meijs' approach and a case-sensitive version of absolute—as opposed to incremental—support, namely the conditional probability. The codomain of the resulting measure obviously is [0, 1] but just like in the case of the Glass-Olsson measure $\theta = 0.5$ although the interpretation of this value differs from the interpretation of the θ values of the other measures. Here, values above θ mean that some proposition is supported to a stronger degree than its negation, while values below indicate the opposite.

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

Another recent coherence measure has been developed by Schippers (2014) and is based on his own measure of support. The values of this measure are in $[-1, 1]$ with $\theta = 0$. Notice that one can obtain the very same coherence measure by using the so-called *power PC* measure by Cheng (1997).

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

Finally, Koscholke (2015) has added four further cadidate measures to the investigation, namely Crupi's (2007) z-measure with values in $[-1, 1]$ and $\theta = 0$, Gaifman's (1979) measure in $[0, \infty)$ with $\theta = 1$, Rips' (2001) measure and Shogenji's (2012) justification measure which according to him is also a measure of evidential support both with values in in $(-\infty, 1]$ and $\theta = 0$:

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

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

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

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

Notice that the codomains and the θ values of the confirmation measures carry over to the coherence measures that are based on the respective measures. It is also worth noticing that although many of the presented confirmation measures have ordinal equivalent versions, the resulting coherence measures are not necessarily ordinally equivalent. Having introduced the candidate measures we may now turn to the experiments.

3 Methods

For the experiments a collection of test cases from Koscholke (2015) has been employed as vignettes. These test cases include Akiba's dice case (cf. Akiba 2000), Bovens and Hartmann's Tweety and their Tokyo murder cases (cf. Bovens and Hartmann 2003), Glass' dedecahedron case (cf. Glass 2005), Meijs' samurai and his rabbit case (cf. Meijs 2005), Meijs and Douven's plane lottery case (cf. Meijs and Douven 2007), Schupbach's robber case (cf. Schupbach 2011), Siebel's pickpocketing case (cf. Siebel 2004) and Siebel and Schippers' inconsistent testimony case (cf. Schippers and Siebel 2014). An overview of the employed test cases is given in Appendix 1, the test case results for each measure in Appendix 2.

Notice that Harris and Hahn (2009) have provided a very similar study to the one presented here. However, they only investigated the empirical adequacy of Bovens and Hartmann's (2003) coherence quasi-ordering and only for a modified version of their Tokyo murder case. The present study can therefore be understood as an extension of Harris and Hahn's project with respect to coherence measures and with respect to test cases.

3.1 Participants

57 participants (36 female, mean age = 25.8) were recruited from the Decision Lab Subject Pool of the University of Göttingen using the online recruiting tool ORSEE (cf. Greiner 2004). Participants received a show-up fee of 7 Euros (approx. USD 9.50) or course-credit.

3.2 Procedure and materials

The participants answered three questionnaires online no later than twelve hours before they arrived for the main study in the lab. The questionnaire included a translation of the brief form of the preference for consistency scale (cf. Cialdini et al. 1995) consisting of nine items, the numeracy scale (cf. Weller et al. 2013) consisting of fourteen items, and the cognitive reflection test (cf. Frederick 2005) consisting of three items. In the lab participants were presented the ten test cases in random order. Except for Bovens and Hartmann's (2003) Tokyo murder case and Siebel's (2004) pickpocketing case, each test case consists of two sets of propositions. Participants were first asked to indicate in which of the two sets the propositions fit together better or if they fit together equally well. Then participants were asked to use a continuous slider ranging from −100 to 100 to indicate the degree to which the propositions for each set fit together. In Bovens and Hartmann's Tokyo murder case participants were asked to rank order the five sets of propositions according to how well the propositions fit together. Here, they also had to rate the degree of coherence of each set of propositions using the slider. For Siebel's (2004) pickpocketing case the participants were asked if the propositions fit together or not. Then again the participants had to use the slider to evaluate how well the propositions fit together. Finally, participants were asked to provide demographic data and received a written debriefing.

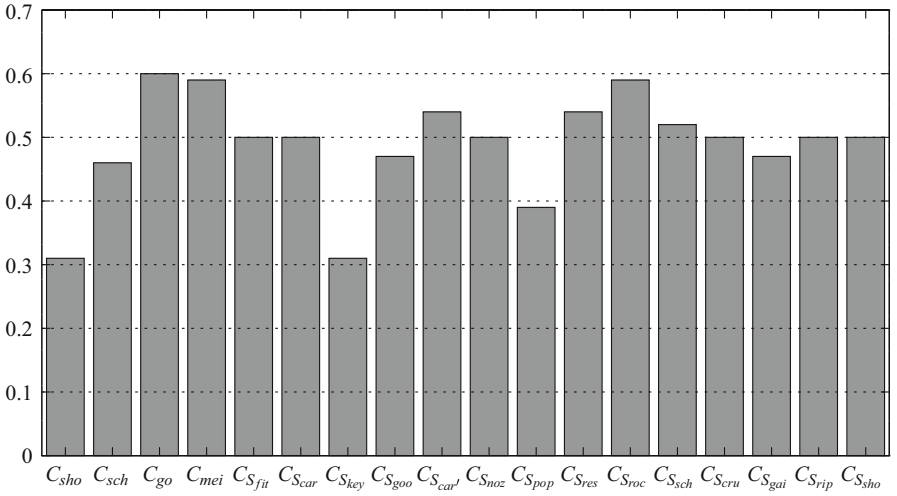

Fig. 1 Correctly predicted choices

3.3 Assessment of predictive accuracy

We assessed three variables to evaluate how well the coherence measures predict participants' coherence assessments. We recorded if participants chose the first or second set of propositions as more coherent or if participants chose that the sets were equally coherent. The first variable *choices* (see Sect. 4.1) is the agreement between participants' choices and the coherence assessments of each measure. For Bovens and Hartmann's Tokyo murder case we recorded the coherence ranking participants gave to the five sets of propositions. The second variable *ranking* (see Sect. 4.2) is the percentage of participants who ranked propositions according to the rankings given by the measures. We also recorded the continuous coherence judgments participants gave for each set of propositions in each test case. The third variable *judgments* (see Sect. 4.3) is the fit between the observed judgments and coherence predictions as assessed in a mixed-linear-regression model for each measure as explained in more detail below.

4 Results

4.1 Choices

Most measures can predict participants' choices better than chance, i.e. 33 % for three choice-options.[1] Correctly predicted choices range from 31 to 60 % between the measures. The three best measures—C_{go}, C_{mei} and $C_{S_{roc}}$—perform equally well around 59% to 60% of correctly predicted choices (see Fig. 1).

[1] We excluded the Siebel and Schippers' inconsistent testimony case from all analyses because most measures have undefined function values in this test case (see Appendix 2).

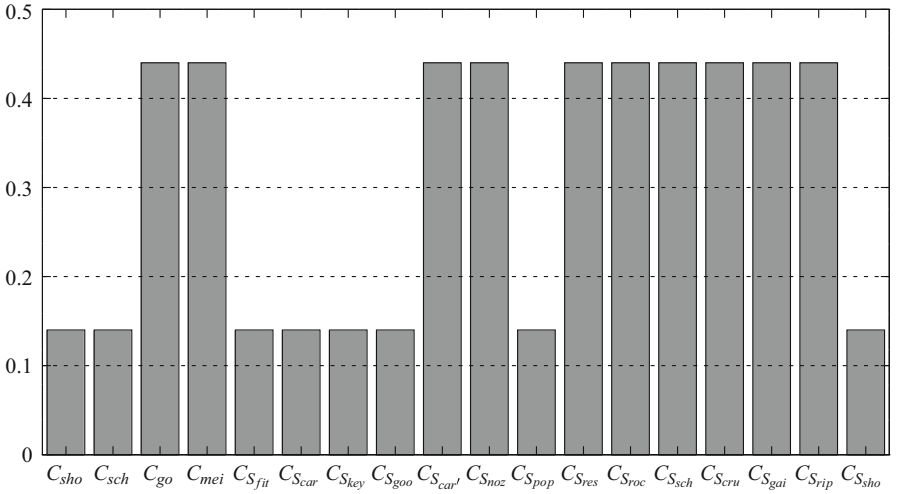

Fig. 2 Correctly predicted rankings

4.2 Rankings

Only six particpants (i.e. 11 %) rank-ordered the five pairs of propositions in Bovens and Hartmann's Tokyo murder test case in the way predicted by 44 % of all measures, i.e. rank-order: 1, 5, 4, 3, 2. A majority of 19 participants (33 %) used a similar ranking differing only in the ranking of the final two pairs of propositions, i.e. rank-order 1, 5, 4, 3, 2 versus 1, 5, 4, 2, 3. Thus, if we allow for one error in the ranking of the final two propositions, 44 % of the measures predict 44 % of participants correctly. Furthermore, allowing a switch in the second and third ranking (i.e. rank-order 1, 4, 5, 2, 3 or 1, 4, 5, 3, 2), the remaining 56 % of measures predict another 14 % of all participants' rankings. Overall, 68 % of participants behave (although not perfectly) in line with at least one of the measures. This also means that a considerable percentage of participants (i.e. 32 %) do not rank-order the pairs of propositions in accordance with any measure. The three best measures for predicting choices—C_{go}, C_{mei} and $C_{S_{roc}}$—also predict the ranking that a majority of participants gave quite well (see Fig. 2).

4.3 Judgments

The investigated coherence measures differ in the assessment of coherence in various test cases, which results in a unique profile for each measure (corresponding to the rows of Table 2 in Appendix 2). We used these profiles as predictors for participants' continuous coherence judgments to test how well the measures can account for the judgments. Conceptually, for each measure j we fitted the profile to the coherence judgments of all participants. To account for the different scaling between predictions and judgments ranging from -100 to 100 used in the study, the regression weight of a factor b_{ij} was estimated from the data to expand the profile. To account for differences in the extent of scaling for each participant i and to also account statistically

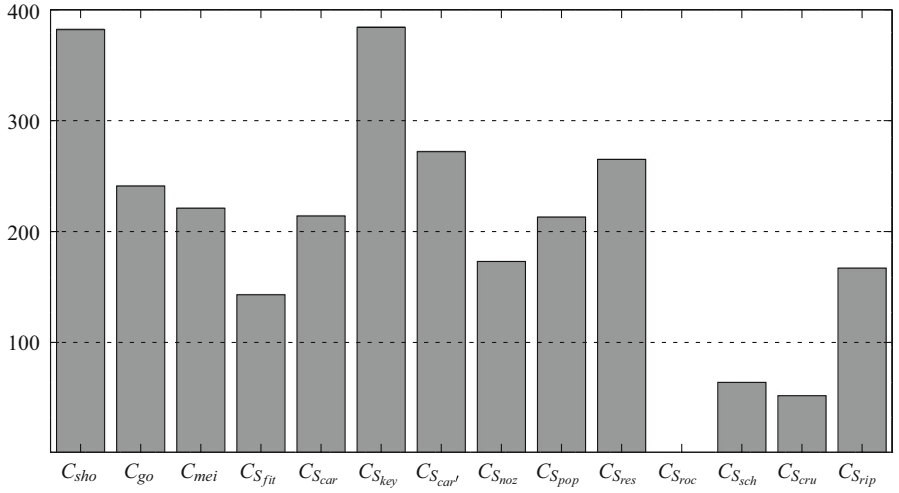

Fig. 3 Bayesian information criterion $\Delta\text{BIC} = \text{BIC}_j - \text{BIC}_{C_{S_{roc}}}$ (i.e. difference of BIC for each measure and best fitting measure by Roche 2013). Notice that C_{sch}, $C_{S_{goo}}$, $C_{S_{gai}}$ and $C_{S_{sho}}$ do not provide a BIC score since they are undefined for some test cases

for repeated ratings from the same participants, b_{ij} consists of the sum of a value shared by all participants and an individual value estimated from the judgments of all participants and participant i. We further accounted for the direction of the predictions by subtracting the neutrality value from all values for each measure to restrict predictions above the neutrality value to judgments above 0, below the neutrality value to judgments below 0, and predictions identical to the neutrality value to zero-judgments. Technically, this can be achieved by including the prediction profile of a measure after subtracting the neutrality value in a mixed-linear-regression without an intercept as a fixed effect and a random effect for each participant.

In order to compare the measures we used the Bayesian Information Criterion (BIC) (cf. Schwarz 1978) from each regression model for each measure as an indicator of how well a measure can account for the participants' continuous coherence judgments (see Fig. 3). The results from the analysis using a software package for linear and nonlinear mixed effects models (cf. Pinheiro et al. 2013) in R (cf. R Core Team 2015) show that the measure $C_{S_{roc}}$ can account for participants' judgments best. The evidence from the data for $C_{S_{roc}}$ is extreme with a Bayes-factor of 2×10^{11} between $C_{S_{roc}}$ and the next best fitting measure (cf. Jeffreys 1961; Wagenmakers 2007). Overall, predictions based on $C_{S_{roc}}$ can describe participants' judgment ratings very well (see Fig. 4).

4.4 Ability and personality as predictors of coherence-judgments

We also analyzed the relation between individual scaling factors b_i for the measure $C_{S_{roc}}$ and the participants' ability to process numbers on the one hand and their per-

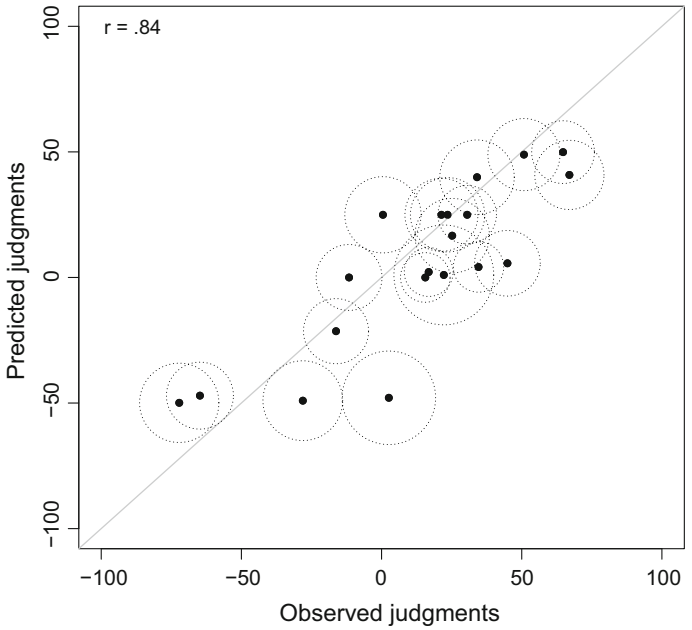

Fig. 4 Scatterplot of mean observed coherence judgments in the test cases and predicted judgments according to the best-fitting measure $C_{S_{roc}}$ (cf. Roche 2013). Note that dotted circles around means indicate 95% confidence intervals. Pearson correlation between observed and predicted means is $r = .84$ ($t(18) = 6.7$, $p < .001$)

sonality on the other hand.[2] The numeracy scale measures "the ability to understand, manipulate, and use numerical information, including probabilities" (Weller et al. 2013, p. 198) by asking participants to solve mainly statistical problems (e.g. "If Person A's chance of getting a disease is 1 in 100 in 10 years, and person B's risk is double that of A, what is B's risk?"). Since the most successful measure $C_{S_{roc}}$ in predicting participants' answers is based on conditional probabilities, we hypothesized that people who are sensitive to the measure $C_{S_{roc}}$ as reflected in a higher scaling factor b_i should also score higher on the numeracy scale.

We found a low correlation of $r = .30$ ($t(49) = 2.24$, $p < .05$) in the predicted direction (see Table 1). Closer inspection revealed that this correlation is driven by a single participant. After removing this participant from the analysis the correlation decreased to $r = .13$ ($t(48) = 0.94$, $p = .35$).

The cognitive reflection test (cf. Frederick 2005) measures if people rely on their first incorrect intuitive answer or reflect more on a task before giving an answer (e.g. "A bat and a ball cost $1.10 in total. The bat costs $1.00 more than the ball. How much does the ball cost?"). We again predicted a positive correlation between the scaling factor and the cognitive reflection test and found a low positive ($r = .11$) but insignificant correlation ($t(49) = 0.79$, $p = 0.43$). The preference for consistency

[2] Six participants did not answer the questionnaire prior to the lab study and were therefore excluded from this analysis.

Table 1 Correlations (Pearson's r) between the scaling factor b_i for $C_{S_{roc}}$, the numeracy scale, the cognitive reflection task, and the preference for consistency scale

	Ability and Personality Scales		
	Numeracy	Cognitive Reflection	Preference for Consistency
Scaling b_i for $C_{S_{roc}}$	.30*	.11	−.28*
Numeracy	(.52)	.43**	−.08
Cognitive Reflection	–	(.53)	−.19
Preference for Consistency	–	–	(.76)

Notice that values in parentheses indicate the internal consistency (Cronbach's α) of scales; * $p < .05$,** $p < .01$; N = 51 participants

scale (cf. Cialdini et al. 1995) measures individuals' preference for one's own and others' behavior being consistent and predictable (e.g. "It is important to me that those who know me can predict what I will do"). Consistent and predictable coherence judgments can be achieved by either being sensitive to coherence and thereby clearly disentangling different degrees of coherence between sets of propositions or by being insensitive to coherence and behaving similarly regarding all sets of propositions. In the analysis we found weak support for the second account: participants with a high preference for consistency show lower scaling factors ($r = -.28$; $t(49) = -2.04$, $p < .05$). Overall, the analyses show that the impact of ability and personality on the subjective coherence assessment is low.

5 Conclusion

The evaluation of the psychological experiments clearly shows that there are probabilistic coherence measures performing better in predicting subjective coherence assessments in the employed test cases than other measures. In particular, one measure standing out from the crowd is Roche's (2013) coherence measure based on Douven and Meijs' average mutual support approach and the conditional probability. This measure shows decent results with respect to comparative coherence assessments (see Sects. 4.1, 4.2) as well as absolute, continuous coherence judgments (see Sect. 4.3).

It is, however, important to notice that this does not mean that measures showing a weak performance as predictors of subjective coherence assessments should be completely disregarded as inadequate. First, being able to predict subjective coherence assessments for a specific case does not ensure that the predicted coherence assessments themselves are correct. It might turn out that based on philosophical considerations the subjective coherence assessments for a certain scenario need to be corrected and as a consequence might be better captured by a measure that has wrongly been disregarded. Second, the empirical adequacy of a probabilistic coherence measure is only one component among others—e.g. satisfaction of certain coherence desiderata or performance in coherence-related test cases—that should be taken into account when evaluating the overall adequacy of a probabilistic coherence measure.

Interestingly enough, Roche's measure also cuts a good figure in these two respects (cf. Schippers 2014; Koscholke 2015). Therefore, this investigation can be understood as providing further, empirical support for the claim that Roche's measure is a very promising candidate for an adequate probabilistic measure of coherence.

Acknowledgements We would like to thank (in alphabetical order) Arndt Bröder, Andreas Glöckner, Björn Meder, Michael Schippers and Mark Siebel for their contributions. We would also like to thank the participants of the Operationalization Workshop 2013 in Freiburg for helpful comments. This work was supported by grant SI 1731/1-1 to Mark Siebel and grant GL 632/3-1 and BR 2130/8-1 to Andreas Glöckner and Arndt Bröder from the Deutsche Forschungsgemeinschaft (DFG) as part of the priority program "New Frameworks of Rationality" (SPP 1516).

Appendix 1: Test cases

Akiba's (2000) Die case

Imagine rolling a fair die and consider the following three statements:

S_1 : The die comes up 2.
S_2 : The die comes up 2 or 4.
S_3 : The die comes up 2 or 4 or 6.

Which pair of statements fits together better. Statement 1 and 2 or statement 1 and 3?

Bovens and Hartmann's (2003) tweety case

Situation 1: Consider a population of 100 animals. 50 out of 100 animals are birds and 50 out of 100 animals cannot fly. Among these 100 animals there is exactly one animal that is a bird and cannot fly. Randomly pick one animal and consider the following two statements:

S_1 : The picked animal is a bird.
S_2 : The picked animal cannot fly.

Situation 2: Consider a population of 100 animals. 50 out of 100 animals are birds and 50 out of 100 animals cannot fly. Among these 100 animals there is exactly one penguin and therefore a bird that cannot fly. Randomly pick one animal and consider the following three statements:

S_1 : The picked animal is a bird.
S_2 : The picked animal cannot fly.
S_3 : The picked animal is a penguin.

In which of the two situations do the respective sets of statements fit together better?

Bovens and Hartmann's (2003) Tokyo murder case

Imagine that a murder has occurred in Toyko and the corpse is still to be found. In order to search more efficiently the map of Tokyo is separated into 100 equally-sized where the probability of finding the corpse is the same for each square. Now, 5 pairs of equally reliable and independent witnesses give the following statements as witness reports:

Pair 1:
S_1 : The corpse is in squares 50 to 60.
S_2 : The corpse is in squares 51 to 61.
Pair 2:
S_1 : The corpse is in squares 22 to 55.
S_2 : The corpse is in squares 55 to 90.
Pair 3:
S_1 : The corpse is in squares 20 to 61.
S_2 : The corpse is in squares 50 to 91.
Pair 4:
S_1 : The corpse is in squares 41 to 60.
S_2 : The corpse is in squares 51 to 70.
Pair 5:
S_1 : The corpse is in squares 39 to 61.
S_2 : The corpse is in squares 50 to 72.

Which pair of statements fits together best, which worst? Can you give and ordering where the first pair is the best and the last pair the worst?

Glass' (2005) Dodecahedron case

Situation 1: You are rolling a fair die.
Situation 2: You are rolling a fair dodecahedron.

Now consider the following two statements:

S_1 : The result will be 2.
S_2 : The result will be 2 or 4.

In which of the two situation do these two statements fit together better?

Meijs' (2005) Samurai case

Situation 1: There are 10,000,000 suspects in a murder case. 1059 of them are Japanese and also 1059 own a samurai sword such that in total there are 9 suspects who are Japanese and own a samurai sword at the same time.
Situation 2: There are 100 suspects in a murder case. 10 of them are Japanese and also 10 own a samurai sword such that in total there are 9 suspects who are Japanese and own a samurai sword at the same time.

Now consider the following two statements:

S_1 : The murderer is Japanese.
S_2 : The murderer owns a samurai sword.

In which of the two situations do the two statements fit together better?

Meijs' (2006) Albino rabbit case

Situation 1: There are 102 rabbits on the first island. 101 out of these 102 rabbits are grey. Also, 101 out of 102 rabbits have two ears. In total there are 100

out of 102 rabbits which are grey and have two ears at the same time. Consequently, there is exactly one rabbit which is grey but does not have two ears and exactly one rabbits which is not grey but has two ears.

Situation 2: There are 102 rabbits on the second island, too. 100 out of these 102 rabbits are grey. Also, 100 out of 102 rabbits have two ears. In total there are 100 out of 102 rabbits which are grey and have two ears at the same time. Consequently, every grey rabbit has two ears and every rabbit that has two ears is also grey.

Now, randomly pick one rabbit and consider the following two statements:

S_1 : The rabbit is grey.
S_2 : The rabbits has two ears.

In which of the two situations do these two statements fit together better?

Meijs and Douven's (2007) plane lottery case

Imagine the following lottery. The chances are $4/100$ for flying to the North pole, $49/100$ for flying to the South pole and $47/100$ for flying to New Zealand. The probability for seeing a penguin at the North pole is 0, at the South pole it is $10/49$ and in New Zealand it is $1/47$. Now consider the following two situations in which when having landed one is confronted with two statements.

Situation 1:
S_1 : You are landing at the North pole.
S_2 : The animal you see is a penguin.
Situation 2:
S_1 : You are landing at the South pole.
S_2 : The animal you see is a penguin.

In which of the two situations do the respective statements fit together better?

Schippers and Siebel's (2015) inconsistent testimony case

Imagine there are 8 suspects for a robbery. It is certain that exactly one of them is the robber. Consider the following two situations in which two independent and equally reliable witnesses give statements about the robber:

Situation 1:
S_1 : The robbery was committed by suspect 1 or 2.
S_2 : The robbery was committed by suspect 2 or 3.
S_3 : The robbery was committed by suspect 1 or 3.
Situation 2:
S_1 : The robbery was committed by suspect 1 or 2.
S_2 : The robbery was committed by suspect 3 or 4.
S_3 : The robbery was committed by suspect 5 or 6.

In which of these two situations do the respective sets of statements fit together better?

Schupbach's (2011) Robber case

Imagine there are 10 suspects for a robbery. It is certain that exactly one of them is the robber. Consider the following two situations in which two independent and equally reliable witnesses make give statements about the robber:

Situation 1:

S_1 : The robbery was committed by suspect 1 or 2 or 3.
S_2 : The robbery was committed by suspect 1 or 2 or 4.
S_3 : The robbery was committed by suspect 1 or 3 or 4.

Situation 2:

S_1 : The robbery was committed by suspect 1 or 2 or 3.
S_2 : The robbery was committed by suspect 1 or 4 or 5.
S_3 : The robbery was committed by suspect 1 or 6 or 7.

In which of these two situations do the respective sets of statements fit together better?

Siebel's (2004) pickpocketing robber case

Imagine the following situation. There are 10 equally likely suspects for a murder. 8 out of 10 have committed a pickpocketing before, 8 out of 10 have committed a robbery and in total 6 out of 10 have committed a pickpocketing and a robbery. Now consider the following two statements:

S_1 : The murderer has committed a robbery.
S_2 : The murderer has committed a pickpocketing.

Do these two statements fit together or not?

Appendix 2: Test case results

See Table 2.

Table 2 Summary of the results

	$T_{1.1}$	$T_{1.2}$	$T_{2.1}$	$T_{2.2}$	$T_{3.1}$	$T_{3.2}$	$T_{3.3}$	$T_{3.4}$	$T_{3.5}$	$T_{4.1}$	$T_{4.2}$
C_{sho}	3	2	0.04	4	8.26	0.0817	0.68	2.5	2.27	3	6
C_{sch}	0.477	0.301	−1.4	0.168	0.917	−1.09	−0.167	0.398	0.356	0.477	0.778
C_{go}	0.5	0.333	0.0101	0.0101	0.833	0.0145	0.167	0.333	0.353	0.5	0.5
C_{mei}	0.5	0.333	0.0101	0.0151	0.833	0.0145	0.167	0.333	0.353	0.5	0.5
$C_{S_{fit}}$	0.833	0.714	−0.96	0.453	0.976	−0.896	−0.288	0.6	0.57	0.833	0.917
$C_{S_{car}}$	0.5	0.333	−0.48	0.255	0.799	−0.321	−0.134	0.3	0.292	0.5	0.625
$C_{S_{key}}$	3	2	0.04	18	8.26	0.0817	0.68	2.5	2.27	3	6
$C_{S_{goo}}$	NaN	NaN	0.0204	NaN	80.9	0.0547	0.552	4	3.65	NaN	NaN
$C_{S_{car'}}$	0.111	0.0833	−0.24	−0.035	0.0879	−0.112	−0.0564	0.06	0.0671	0.111	0.0694
$C_{S_{noz}}$	0.65	0.467	−0.96	0.182	0.898	−0.494	−0.232	0.375	0.379	0.65	0.705
$C_{S_{pop}}$	0.604	0.426	−0.932	0.287	0.863	−0.857	−0.213	0.471	0.435	0.604	0.789
$C_{S_{res}}$	0.15	0.133	−0.48	−0.0733	0.0988	−0.173	−0.0972	0.075	0.0871	0.15	0.0795
$C_{S_{roc}}$	0.75	0.667	0.02	0.51	0.909	0.0286	0.286	0.5	0.522	0.75	0.75
$C_{S_{sch}}$	0.75	0.667	−0.98	0.343	0.908	−0.945	−0.448	0.429	0.442	0.75	0.75
$C_{S_{cru}}$	0.7	0.6	−0.96	0.343	0.898	−0.918	−0.32	0.375	0.379	0.7	0.727
$C_{S_{gai}}$	NaN	NaN	0.51	NaN	9.79	0.669	0.812	1.6	1.61	NaN	NaN
$C_{S_{rip}}$	0.7	0.6	−0.96	0.343	0.898	−0.495	−0.232	0.375	0.379	0.7	0.727
$C_{S_{sho}}$	0.807	0.693	−4.64	−0.224	0.957	−2.39	−0.444	0.569	0.557	0.807	0.861

Table 2 continued

	$T_{5.1}$	$T_{5.2}$	$T_{6.1}$	$T_{6.2}$	$T_{7.1}$	$T_{7.2}$	$T_{8.1}$	$T_{8.2}$	$T_{9.1}$	$T_{9.2}$	T_{10}
C_{sho}	80.3	9	0.999	1.02	0	1.86	0	0	2.37	2.37	0.937
C_{sch}	1.9	0.954	−4.26e-05	0.0086	NaN	0.268	NaN	NaN	0.312	0.162	−0.028
C_{go}	0.00427	0.818	0.98	1	0	0.2	0	0	0.25	0.143	0.6
C_{mei}	0.00427	0.818	0.98	1	0	0.2	0.25	0	0.438	0.186	0.6
$C_{S_{fit}}$	0.976	0.976	−0.00498	1	−1	0.587	−0.333	−1	0.382	0.343	−0.143
$C_{S_{car}}$	0.00839	0.8	−9.71e-05	0.0196	−0.075	0.257	−0.0417	NaN	0.198	0.188	−0.05
$C_{S_{key}}$	80.3	9	1	1.02	0	1.86	0.75	NaN	1.56	1.78	0.937
$C_{S_{goo}}$	80.9	81	0.99	NaN	0	6.24	1	NaN	2.46	NaN	0.75
$C_{S_{car'}}$	8.89e-07	0.08	−9.61e-05	0.0192	−0.0044	0.0461	0	−0.0556	0.0703	0.0312	−0.04
$C_{S_{noz}}$	0.00839	0.889	−0.0099	1	−0.0798	0.328	−0.0375	NaN	0.308	0.229	−0.25
$C_{S_{pop}}$	0.975	0.872	−9.71e-05	0.0196	−1	0.37	−0.383	NaN	0.256	0.242	−0.0516
$C_{S_{res}}$	8.89e-07	0.0889	−0.0098	0.98	−0.00476	0.0711	0.00417	NaN	0.11	0.0411	−0.2
$C_{S_{roc}}$	0.0085	0.9	0.99	1	0	0.557	0.25	NaN	0.542	0.5	0.75
$C_{S_{sch}}$	0.00839	0.899	−0.0099	1	−1	0.513	−0.333	NaN	0.396	0.25	−0.25
$C_{S_{cru}}$	0.00839	0.889	−9.8e-05	1	−1	0.464	−0.375	NaN	0.311	0.254	−0.0625
$C_{S_{gai}}$	1.01	9	0.99	NaN	0.925	3.36	1.04	NaN	1.53	NaN	0.8
$C_{S_{rip}}$	0.00839	0.889	−0.0099	1	−0.0826	0.464	−0.05	NaN	0.311	0.276	−0.25
$C_{S_{sho}}$	0.479	0.954	−0.00995	1	NaN	0.573	NaN	NaN	0.419	0.308	−0.289

'NaN' indicates undefined function values which are interpreted as suspended coherence judgments (cf. Siebel and Wolff 2008). Some values are given in scientific notation where e.g. 8.89e−07 means .000000889. The indices in the heading of the table correspond to the order of the test cases given before in Appendix 1

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.

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.

Cialdini, R. B., Trost, M. R., & Newsom, J. T. (1995). Preference for consistency: The development of a valid measure and the discovery of surprising behavioral implications. *Journal of Personality and Social Psychology*, *69*, 318–328.

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.

Festa, R. (2012). For unto every one that hath shall be given. Matthew properties for incremental confirmation. *Synthese*, *184*, 89–100.

Finch, H. A. (1960). Confirming power of observations metricized for decisions among hypotheses. *Philosophy of Science*, *27*, 293–307.

Fitelson, B. (2004). Two technical corrections to my coherence measure. http://fitelson.org/coherence2.

Fitelson, B. (2003). A probabilistic theory of coherence. *Analysis*, *63*, 194–199.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, *19*, 25–42.

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 O'Neill, M., Sutcliffe, R. F. E., Ryan, C., Eaton, M., & Griffith, N. J. L. (Eds.), *Artificial intelligence and cognitive science. 13th Irish conference, AICS 2002, Limerick, Ireland, September 2002* (pp. 177–182). 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.

Greiner, B. (2004). An online recruitment system for economic experiments. In K. Kremer & V. Macho (Eds.), *Forschung und wissenschaftliches Rechnen 2003, GWDG Bericht 63* (pp. 79–93). Goettingen: Ges. fuer Wiss. Datenverarbeitung.

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.

Jeffreys, H. (1961). *Theory of probability*. Oxford: Oxford University Press.

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.

Kolmogorov, A. (1956). *Foundations of the theory of probability*. New York: AMS Chelsea Publishing.

Koscholke, J. (2015). Evaluating test cases for probabilistic measures of coherence. *Erkenntnis*. doi:10.1007/s10670-015-9734-1.

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. (2007). On the alleged impossibility of coherence. *Synthese*, *157*(3), 347–360.

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.

Pfeiffer, P. (1990). *Probability for applications*. New York: Springer.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., & R Core Team (2013). nlme: Linear and nonlinear mixed effects models. http://CRAN.R-project.org/package=nlme.

Popper, K. R. (1954). Degree of confirmation. *British Journal for the Philosophy of Science*, *5*, 143–149.
R Core Team (2015). R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing.
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.
Schippers, M., & Siebel, M. (2015). Inconsistency as a touchstone for coherence measures. *Theoria*, *30*, 11–41.
Schupbach, J. N. (2011). New hope for Shogenji's coherence measure. *British Journal for the Philosophy of Science*, *62*(1), 125–142.
Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464.
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., & Wolff, W. (2008). Equivalent testimonies as a touchstone of coherence measures. *Synthese*, *161*, 167–182.
Wagenmakers, E. J. (2007). A practical solution to the pervasive problems of *p* values. *Psychonomic Bulletin & Review*, *14*, 779–804.
Weller, J. A., Dieckmann, N. F., Tusler, M., Mertz, C. K., Burns, W. J., & Peters, E. (2013). Development and testing of an abbreviated numeracy scale: A Rasch analysis approach. *Journal of Behavioral Decision Making*, *26*, 198–212.