Available online at www.sciencedirect.com
ScienceDirect
Cognition 103 (2007) 107–119
COGNITION
www.elsevier.com/locate/COGNIT

# Comparison of confirmation measures ☆,☆☆

Katya Tentori [a,*], Vincenzo Crupi [a,c], Nicolao Bonini [a], Daniel Osherson [b]

[a] *CRD, DiSCoF, University of Trento, via Matteo del Ben 5, 38068 Rovereto (TN), Italy*
[b] *Department of Psychology, Princeton University, Green Hall, Washington Street, Princeton (NJ) 08540, United States*
[c] *Laboratory of Cognitive Psychology, CNRS & University of Aix-Marseille I, 3 Place Victor Hugo F-13331, Marseille, France*

Received 21 March 2005; revised 2 September 2005; accepted 2 September 2005

**Abstract**

Alternative measures of *confirmation* or *evidential support* have been proposed to express the impact of ascertaining one event on the credibility of another. We report an experiment that compares the adequacy of several such measures as descriptions of confirmation judgment in a probabilistic context.


*Keywords:* Bayesian confirmation measures; Evidential support; Inductive reasoning; Belief updating

☆ This manuscript was accepted under the editorship of Jacques Mehler.
☆☆ We thank Branden Fitelson, Douglas Medin, Lance Rips, and an anonymous referee for helpful comments on an earlier draft. This work was supported in part by a FIRB-MIUR grant awarded to the third author.
* Corresponding author.
*E-mail addresses:* katya.tentori@unitn.it (K. Tentori), vincenzo.crupi@unitn.it (V. Crupi), nicolao.bonini@unitn.it (N. Bonini), osherson@princeton.edu (D. Osherson).

0010-0277/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.cognition.2005.09.006

## 1. Evidential impact and measures of confirmation

Perhaps the reader will concur that owning a laptop is better evidence for a physics diploma than for a taxi license. Judgments like these seem central to inductive logic, so epistemologists have attempted to quantify them via scales that relate evidence $e$ (owning a laptop) to hypotheses $H$ (having a physics diploma). The measured quantity is known as the *degree of support* or *confirmation* that $e$ brings to $H$. Most confirmation measures are embedded in a Bayesian framework, and thus involve prior and posterior probabilities, and likelihoods. Among the candidates are the following.[1]

Rival Confirmation measures:

$$d(e,H) = Pr(H|e) - Pr(H) \text{ (Eells, 1982; Jeffrey, 1992)},$$

$$r(e,H) = \log\left[\frac{Pr(H|e)}{Pr(H)}\right] \text{ (Keynes, 1921; Horwich, 1982)},$$

$$n(e,H) = Pr(e|H) - Pr(e|\neg H) \text{ (Nozick, 1981)},$$

$$\ell(e,H) = \log\left[\frac{Pr(e|H)}{Pr(e|\neg H)}\right] \text{ (Good, 1984)},$$

$$c(e,H) = Pr(H \wedge e) - (Pr(e) \times Pr(H)) \text{ (Carnap, 1962)},$$

$$k(e,H) = \frac{Pr(e|H) - Pr(e|\neg H)}{Pr(e|H) + Pr(e|\neg H)} \text{ (Kemeny \& Oppenheim, 1952)}.$$

Each measure maps given evidence $e$ and hypothesis $H$ into a number meant to indicate the impact of $e$ on the credibility of $H$. Values can be either positive or negative, corresponding to confirmation and disconfirmation, respectively (zero indicates neutrality). Note that evidential impact is not the same as the ultimate credibility of $H$, usually taken to be its posterior probability $Pr(H|e)$. To illustrate, select a man from New York City at random and suppose he owns a laptop. This is better evidence for him possessing a physics diploma than a taxi license even though the probability of the latter might be greater (because there are so many taxi-drivers in New York).

In a pioneering study, Briggs and Krantz (1992) investigated the combined effect of distinct items of evidence on overall belief, using scenarios drawn from common experience (e.g., predicting compatibility of roommates). Their participants were able to judge the influence of each piece of evidence independently of the others. In contrast, the present study focusses on individual evidence in a standard chance set-up. We attempt to identify the confirmation measure that best predicts the psychological impact of samples drawn from an urn on beliefs about the urn's composition. Success in this enterprise might ultimately allow the appropriate rule of

[1] We rely on Eells and Fitelson (2002) for some literature citations.

combination to predict the overall confidence in an hypothesis supported by a multiplicity of evidence (as in Shafer, 1976).

Information about psychologically correct confirmation measures might be expected from the extensive literature on *causal induction* (see Perales & Shanks, 2003). The iterated trials that arise in the study of causal induction, however, seem not to be adapted to the single-case perspective of evidential confirmation. Thus, the "probabilistic contrast model" of causal induction (Cheng & Novick, 1990) offers $n$ (defined above) to measure the strength of $H$ as a cause of evidence $e$. (In the causal induction literature, the notation $\Delta P$ is used in place of $n$.) But suppose an urn contains 100 balls labeled $1 \cdots 100$, and consider the events:

$H =$ the label belongs to $\{1 \cdots 20\}$,

$e_1 =$ the label belongs to $\{1 \cdots 10\}$,

$e_2 =$ the label belongs to $\{1 \cdots 11\} \cup \{100\}$.

Then $n(e_1,H) = .5$ and $n(e_2,H) = .54$ which seems to reverse the intuition that $e_1$ confirms $H$ more than $e_2$ does (since $e_1$ implies $H$ whereas $e_2$ does not). The same problematic predictions arise if $n$ is normalized through division by $1 - Pr(e|\neg H)$, as suggested by the *power PC* model of Cheng (1997).[2]

POWER PC MODEL:

$$p(e,H) = \begin{cases} \frac{n(e,H)}{1-Pr(e|\neg H)} & \text{if } n(e,H) \geqslant 0, \\ \frac{n(e,H)}{Pr(e|\neg H)} & \text{otherwise.} \end{cases}$$

Similar difficulties afflict other measures in our list. For example, $r$ and $c$ are symmetric in their arguments $e$, $H$ whereas simple examples suggest that $e$ may confirm $H$ to a different extent than $H$ confirms $e$ (see below). We are nonetheless inclined to leave in play all the measures exhibited above since ordinary judgment sometimes contradicts the intuitions of experts.[3]

The measure $c$ may be stated in the equivalent (but less intuitive) form:

$$c(e,H) = \frac{Pr(e|H)Pr(H)}{Pr(H|e)} \times [Pr(H|e) - Pr(H)].$$

With this adjustment, all the measures are functions of $Pr(H)$, $Pr(H|e)$, $Pr(e|H)$, and $Pr(e|\neg H)$.

[2] So far as we know, researchers in the area of causal induction have advanced no claims about confirmation. It is worth noting that easy examples show "X confirms $Y$" to imply neither that X causes $Y$ nor that $Y$ causes X. (Thus, a rash on the left arm and the right arm may be evidence for each other without either causing the other.) It is also easy to see that "X causes $Y$" does not imply that $Y$ confirms X more than slightly. (For example, a meteor strike might cause the shattering of a tea cup whereas the shattered tea cup is not much evidence for a meteor strike.) On the other hand, "X causes $Y$" does seem to imply that X confirms $Y$.

[3] An illustration is the "conjunction fallacy". See Tentori, Bonini, and Osherson (2004).

Table 1
Correlations between confirmation measures in a Monte Carlo test

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| *d* | 1 | .822 | .930 | .936 | .939 | .970 | .901 | .736 |
| *r* | .822 | 1 | .781 | .883 | .770 | .857 | .805 | .686 |
| *n* | .930 | .781 | 1 | .884 | .987 | .941 | .954 | .694 |
| $\ell$ | .936 | .883 | .884 | 1 | .871 | .970 | .847 | .777 |
| *c* | .939 | .770 | .987 | .871 | 1 | .927 | .940 | .700 |
| *k* | .970 | .857 | .941 | .970 | .927 | 1 | .916 | .791 |
| *p* | .901 | .805 | .954 | .847 | .940 | .916 | 1 | .671 |
| *pst* | .736 | .686 | .694 | .777 | .700 | .791 | .671 | 1 |
| | *d* | *r* | *n* | $\ell$ | *c* | *k* | *p* | *pst* |

*Note.* The correlations represent 10,000 selections of a joint distribution over the variables $e$, $H$ (chosen uniformly randomly with no zero probabilities). *pst* is the posterior probability $Pr(H|e)$.

The following Monte Carlo procedure was used to evaluate the agreement of different measures. On 10,000 trials, we uniformly randomly assigned nonzero probabilities (summing to one) to each of $e \wedge H$, $e \wedge \neg H$, $\neg e \wedge H$, and $\neg e \wedge \neg H$. These numbers imply a value of $x(e, H)$ for each of our seven confirmation measures, $x$. The Pearson correlation over the 10,000 trials was then computed for all pairs of measures; the posterior probability $Pr(H|e)$ was included for comparison. The correlations are shown in Table 1 and reveal considerable agreement among some pairs of measures. Appropriately, the least agreement is with $Pr(H|e)$. One of the high correlations seen in Table 1 has an algebraic basis. Indeed, $k$ are $\ell$ are ordinally equivalent inasmuch as $k(e, H) = \tanh(\frac{1}{2}\ell(e, H))$ which is a monotone transformation.[4]

## 2. The normative appeal of different measures

Confirmation measures are often deployed to solve philosophical puzzles, e.g., why verifying diverse consequences lends a theory more credibility than verifying similar consequences (see Earman, 1992; Horwich, 1982). Different measures appear to be needed, however, to resolve different puzzles, which undermines the hope for a unitary analysis of confirmation (Fitelson, 1999). Confirmation measures nonetheless vary in their conformity to standards advocated by epistemologists. Eells and Fitelson (2002) defend four standards that we summarize now.

Let $x$ be a given measure of confirmation. Then $x$ satisfies *evidence symmetry* (ES) if the confirmation that evidence $e$ bestows upon hypothesis $H$ is the same as the disconfirmation that the negated evidence bestows upon $H$. [If e is initially disconfirmatory then its negation should be equally confirmatory, according to (ES).] To illustrate, (ES) requires that spots confirm the diagnoses of measles to the same extent that the absence of spots disconfirms measles. The principle is abbreviated as follows

[4] We thank Branden Fitelson for this observation.

(ES) $x(e,H) = -x(\neg e,H)$.

Although the spots/measles example may accord with (ES), other examples definitely do not. Thus, drawing a Jack maximally confirms the hypothesis that a face card was drawn (indeed, implies it) but failing to draw a Jack only partially disconfirms the same hypothesis. Eells and Fitelson (2002) therefore conclude that an adequate measure of confirmation should *violate* (ES) for at least some choices of $e$ and $H$.

Similarly, elementary examples discredit both of the following principles, rendering their violation (for some choices of $e$ and $H$) a virtue for confirmation measures.

(CS) $x(e,H) = x(H,e)$,

(TS) $x(e,H) = x(\neg e,\neg H)$.

The first is *commutativity symmetry*, the second *total symmetry.* Both succumb to the Jack/face-card example used earlier. In contrast, the following principle of *hypothesis symmetry* is immune to any counterexample that Eells and Fitelson (and the present authors) have been able to devise, and plausibly relates confirmation/disconfirmation to hypothesis polarity.

(HS) $x(e,H) = -x(e,\neg H)$.

Eells and Fitelson (2002) therefore advocate conformity to (HS). They go on to classify confirmation measures in terms of agreement with their recommendations. Extending their analysis to $k$ and $p$, the following fact may be demonstrated.

> First Normative comparison: Of the seven confirmation measures described above, only $d$, $\ell$, and $k$ conform to (HS) but not to (ES), (CS), and (TS). That is, only $d$, $\ell$, and $k$ agree with the recommendations that Eells and Fitelson (2002) offer for confirmation measures.

Although $d$ and $\ell$ are endorsed by the foregoing comparison, they are questionable on other grounds. Thus, $d(e,H)$ need not assume a maximal value even if $e$ implies $H$. For example, let $e$ represent drawing the ace of spades, and let $H_1$ and $H_2$ represent drawing a black suit and a spade, respectively. Then $d(e,H_1) = .50 < .75 = d(e,H_2)$ even though $e$ entails $H_1$. This behavior is strange inasmuch as entailment seems to underwrite optimal confirmation of $H_1$ by $e$. Measure $\ell$ is likewise unsatisfactory since it is not even defined when $e$ entails $H$. In contrast, $k(e,H_1) = k(e,H_2) = 1.0$, the maximum possible value for $k$. More generally:

> Second Normative comparison: Of the seven confirmation measures described above, only $k$ reaches a maximal value when $e$ entails $H$ (the same value for all hypotheses $H$).

Combining the two comparisons suggests a normative preference for $k$. We are thus in a position to ask whether the most *normatively justified* confirmation measure is among the most *psychologically descriptive.* This will be the case if $k$ predicts the

felt impact of evidence as well as any other confirmation measure. At issue are the feelings arising in philosophically "naive" individuals; the intuitions of experts merely set the normative stage.

## 3. Experimental test of the confirmation measures

Our experiment relied on a chance set-up familiar from early experiments on probability judgment (notably, Philips & Edwards, 1966). All of the evidence presented to participants was "designated", to use the terminology of Briggs and Krantz' (1992) study of confirmation; that is, none was meant to be ignored.

### 3.1. Materials and procedure

Participants were interviewed individually. Two opaque urns were presented and their contents described. The two urns were composed as follows.

| Urn | Number of black balls | Number of white balls |
|---|---|---|
| *A* | 30 | 10 |
| *B* | 15 | 25 |

A diagrammatic reminder of this information was left in view throughout the procedure.

It was explained that a coin toss would be used to choose one of the urns, but that the outcome would remain hidden. The coin was tossed, one urn was covertly selected, the other set aside. Ten random extractions without replacement from the chosen urn were then carried out in front of the participant. All random events in the procedure were genuine results of chance; different participants thus faced different urns and different extractions. The extracted balls remained in view, lined up on the table in their order of selection.

To construct an *impact scale*, strips of paper were printed with a dotted line below the following labels, spaced evenly from left to right. (All materials are translated from Italian.)

- weakens my conviction extremely
- weakens my conviction a lot
- slightly weakens my conviction
- has no effect on my conviction
- slightly strengthens my conviction
- strengthens my conviction a lot
- strengthens my conviction extremely

The dotted line extended beyond the labels, allowing judgments of arbitrary extremity. Note that our scale represents strengthening and weakening of belief symmetrically (unlike the scale used in Briggs & Krantz, 1992, for somewhat different purposes). After each extraction, six tasks were performed.

**Task 1**. Participants marked a position on the impact scale to indicate the influence of the latest draw on their current conviction that urn $A$ had been selected. Fresh copies of the scale were used for each extraction, with successive scales left on the table through the 10 extractions. To underline that the judgments concerned changes in the participant's *current* conviction (not his conviction before the first extraction), the neutral point on the scale used in trial $n + 1$ was vertically aligned with the mark made in trial $n$. On a given scale, the signed distance from the neutral point was used to quantify impact.

The instructions accompanying use of the scale stated that it was continuous and unbounded in both directions. We also emphasized that the distances among the marks left by the 10 successive evaluations would be compared, hence were meaningful.

**Task 2**. A fresh copy of the scale was then used to indicate the evidential impact for urn $B$. (Our rival confirmation measures do not generally impose a simple relation between the answers for the two urns.)

**Tasks 3 and 4**. Participants next responded to the following questions.

> In view of the color of the extracted ball, what probability do you now assign to urn $A$ having been selected?
> In view of the color of the extracted ball, what probability do you now assign to urn $B$ having been selected?

The format of the response (decimal, fraction or percentage) was at the discretion of the participant. It was pointed out each time that the sum of their responses had to equal one.

**Tasks 5 and 6**. Finally, participants responded to these questions:

> Assuming that urn $A$ was selected, what was the probability of extracting a ball of *this* color [the color actually observed] in the current trial?
> Assuming that urn $B$ was selected, what was the probability of extracting a ball of *this* color in the current trial?

Again, the format of the response was left open. It was pointed out that there was no need for the two numbers to sum to one.

The first two tasks measure the impact of the latest extraction on the hypothesis that $A$ versus $B$ was initially selected. These numbers will be denoted *Judged* $(e,A)$ and *Judged*$(e,B)$ in what follows. Tasks 3 and 4 assess the participant's (subjective)

probability for the hypothesis $A$ versus $B$ given the latest extraction. These numbers will be denoted $Pr_{subj}(A|e)$ and $Pr_{subj}(B|e)$. Tasks 5 and 6 assess the likelihood of the latest extraction assuming that $A$ versus $B$ was selected. These numbers will be denoted $Pr_{subj}(e|A)$ and $Pr_{subj}(e|B)$. The procedure yields 10 values for each of $Judged(e,A)$, $Pr_{subj}(A|e)$, and $Pr_{subj}(e|A)$, and similarly for $B$. In each case, "$e$" refers to the evidence that issues from the latest extraction.

We use $Pr_{subj}(A)$ and $Pr_{subj}(B)$ to denote the answers to the questions in Tasks 3 and 4 of the *preceding trial.* That is, $Pr_{subj}(A)$ and $Pr_{subj}(B)$ measure confidence in $A$ and $B$ prior to the latest extraction; for the first extraction, we verified that the participant recognized the probability of both $A$ and $B$ to be one-half. Hence, for each participant, there are 10 values of $Pr_{subj}(A)$ and 10 values of $Pr_{subj}(B)$, one for each extraction.

The *objective* probabilities corresponding to $Pr_{subj}(A|e)$, $Pr_{subj}(e|A)$, $Pr_{subj}(A)$ will be denoted $Pr_{obj}(A|e)$, $Pr_{obj}(e|A)$, $Pr_{obj}(A)$, respectively, and similarly for $B$. These are the probabilities implied by the randomness of the coin flip and urns. Again, for each participant there are 10 sets of these six numbers, one for each extraction.

### 3.2. Participants

Twenty-six students from the University of Trento and the University of Milan-Bicocca completed the procedure. Mean age was 24 years. There were 14 men and 12 women.

## 4. Results

### 4.1. Correspondence of subjective and objective probabilities

Before considering $Judged(e,A)$ and $Judged(e,B)$, let us compare $Pr_{subj}(A|e)$ with $Pr_{obj}(A|e)$, $Pr_{subj}(e|A)$ with $Pr_{obj}(e|A)$, and $Pr_{subj}(e|B)$ with $Pr_{obj}(e|B)$. There is no need to consider $Pr_{subj}(B|e)$ since Tasks 3 and 4 constrained it to equal $1 - Pr_{subj}(A|e)$.

Concerning the first contrast, for each participant we compared the probability assigned to $A$ after the 10th extraction [denoted $Pr_{subj}(A|e_{10})$] with the corresponding objective probability [$Pr_{obj}(A|e_{10})$]. It is often reported in settings like ours that subjective posterior probabilities are *conservative* in the sense of lagging behind their objective counterparts; such estimates are smaller when objective posteriors rise and larger when they fall (Edwards, 1968; Slovic & Lichtenstein, 1971). Since all participants agreed that the prior probability (before the first extraction) was .5, we computed the following index CI, which yields a positive value for responses that are less extreme than the objective probability, suggesting a conservative bias.

$$\mathrm{CI} = \begin{cases} Pr_{obj}(A|e_{10}) - Pr_{subj}(A|e_{10}) & \text{if } Pr_{obj}(A|e_{10}) > .5 \\ Pr_{subj}(A|e_{10}) - Pr_{obj}(A|e_{10}) & \text{otherwise.} \end{cases}$$

CI was positive for 20 of the 26 participants, and its mean value of .117 ($SD = .155$) was reliably greater than 0 [$t(25) = 3.9$]. These results might reflect "anchoring" on

.5 (insufficient use of evidence), or the simple fact that $Pr_{obj}(A|e_{10})$ tended towards extreme probabilities thereby leaving little room for negative values of CI. Indeed, for 19 of the 26 participants, $Pr_{obj}(A|e_{10})$ was within .06 of either 0 or 1.

Next consider likelihoods. For each extraction $e_i$ (among the 10 extractions $e_1 \ldots e_{10}$), we computed the 26 differences $Pr_{obj}(e_i|A) - Pr_{subj}(e_i|A)$, one for each participant. All 10 average differences are positive. The grand mean is .031 ($SD = .013$), which differs reliably from zero [$t(9) = 7.4$, prob $< .01$]. Likewise, for each $e_i$ we computed $Pr_{obj}(e_i|B) - Pr_{subj}(e_i|B)$ for each participant. In this case, 9 of the 10 average differences were positive with the grand mean .023 ($SD = .018$) again reliably different from zero [$t(9) = 3.9$, prob $< .01$]. Thus, likelihoods were systematically underestimated. The mean discrepancies of .031 and .023 seem minor, however, inasmuch as objective likelihoods remained near .5 throughout the procedure. The average value of $Pr_{obj}(e|A)$ over the 260 extractions of the experiment was .520; the average for $Pr_{obj}(e|B)$ was .486.

### 4.2. Predicting subjective confirmation

We now attempt to identify which of the seven confirmation measures discussed earlier corresponds most closely to judged evidential impact. For this purpose, we assume that *Judged*(*e*,*A*) and *Judged*(*e*,*B*) lie on interval scales and consider their linear relation to rival confirmation measures.

Let $x$ stand for one of the confirmation measures, and consider a given participant. For each extraction $e_i$, we computed $x(e_i,A)$ on the basis of $Pr_{subj}(A)$, $Pr_{subj(A|e)}$, $Pr_{subj}(e|A)$,and $Pr_{subj}(e|B)$. [The latter number represents $Pr_{subj}(e|\neg A)$.] For example, we computed $d(e_i,A) = Pr_{subj}(A|e_i) - Pr_{subj}(A)$ for $i \leq 10$. The Pearson correlation between $x(e_i,A)$ and *Judged*$(e_i,A)$ was then calculated ($N = 10$). In the same way, we calculated the Pearson correlation between $x(e_i,B)$ and *Judged*$(e_i,B)$ (again, $N = 10$). Average correlations over the 26 participants are shown in Table 2.

The last row of the table gives the correlation between $Pr_{subj}(A|e_i)$ and *Judged*$(e_i,A)$ as well as the correlation between $Pr_{subj}(B|e_i)$ and *Judged*$(e_i,B)$. These two

Table 2
Average correlations between judged evidential impact and confirmation measures (confirmation computed from subjective probabilities)

| Confirmation measure | Average correlation with *Judged* (*e*,*A*) | Average correlation with *Judged* (*e*,*B*) |
|---|---|---|
| $d$ | .608 | .603 |
| $r$ | .605 | .620 |
| $n$ | .692* | .712* |
| $\ell$ | .722* | .735* |
| $c$ | .581 | .581 |
| $k$ | .718* | .734* |
| $p$ | .710* | .715* |
| $Pr_{subj}(A[B]\|e)$ | .489 | .499 |

*Note.* Each number is the average of 26 correlations (one per participant); each correlation involves 10 observations. $Pr_{subj}(A[B]\|e)$ denotes $Pr_{subj}(A|e)$ or $Pr_{subj}(B|e)$ as appropriate. Starred averages are reliably greater than the average for $Pr_{subj}(A[B]\|e)$ by paired $t$-test (prob $< .01$).

Table 3
Comparison of $\ell$ and $k$ with other confirmation measures (confirmation computed from subjective probabilities)

| | $d$ | $r$ | $n$ | $c$ | $p$ |
|---|---|---|---|---|---|
| *Predicting Judged (e,A)* | | | | | |
| $\ell$ | $t = 2.2$ | $t = 2.6$ | $t = 1.8$ | $t = 2.6$ | $t = 1.9$ |
| | prob $< .05$ | prob $< .05$ | n.s. | prob $< .05$ | n.s. |
| | 17 | 16 | 20 | 18 | 19 |
| $k$ | $t = 2.1$ | $t = 2.4$ | $t = 1.8$ | $t = 2.5$ | $t = 1.6$ |
| | prob $< .05$ | prob $< .05$ | n.s. | prob $< .05$ | n.s. |
| | 17 | 16 | 18 | 18 | 17 |
| *Predicting Judged (e,B)* | | | | | |
| $\ell$ | $t = 2.2$ | $t = 1.8$ | $t = 1.3$ | $t = 2.5$ | $t = 1.3$ |
| | prob $< .05$ | n.s. | n.s. | prob $< .05$ | n.s. |
| | 18 | 17 | 15 | 18 | 18 |
| $k$ | $t = 2.1$ | $t = 1.8$ | $t = 1.4$ | $t = 2.4$ | $t = 1.5$ |
| | prob $< .05$ | n.s. | n.s. | prob $< .05$ | n.s. |
| | 18 | 17 | 17 | 18 | 17 |

*Note.* Each cell reports a paired $t$-test between the correlations obtained with the confirmation measures in the associated row and column. For each test, $N = 26$ (corresponding to the 26 participants). The correlations each involve 10 observations. The last row of each cell shows the number of participants (out of 26) for whom $\ell$ (or $k$) predicted better than the rival measure at the top of the column.

correlations would be close to unity if participants confused evidential impact with posterior probability. But they do not exceed .5. Paired $t$-tests reveal that most of the confirmation measures produced reliably greater correlations with judged evidential impact than did posterior probability (see Table 2). The results therefore suggest that participants properly distinguished posteriors from impact.

The table also shows that $\ell$ and $k$ yield the highest average correlations with *Judged* $(e,A)$ and *Judged* $(e,B)$, reaching significance at the 2% level ($t$-test for Pearson coefficients, 8 *df*). The measures $p$ and $n$ follow closely behind. By paired $t$-test, both $\ell$ and $k$ are reliably superior to $d$, $r$, and $c$ in predicting *Judged*$(e_i,A)$ (prob $< .05$), and reliably superior to $d$ and $c$ in predicting *Judged* $(e_i,B)$ (prob $< .05$). No other comparisons reach significance. See Table 3, which also shows the number of participants whose judgments were better predicted by $\ell$ and $k$.

A clearer picture emerges if the confirmation measures are computed on the basis of $Pr_{obj}(A|e)$, $Pr_{obj}(e|A)$, and $Pr_{obj}(A)$ instead of their subjective counterparts (and similarly for urn $B$). To motivate this use of objective probabilities, consider the participant's confidence in $A$ given the latest extraction $e$. The confidence is expressed only indirectly by $Pr_{subj}(A|e)$ given the distortion that might occur when underlying feelings of confidence must be quantified in an experimental setting (see Windschitl, 2002). The objective probability of $A$ given $e$ can be understood as another indirect measure of the participant's feelings inasmuch as she is attempting to perceive this very quantity and may succeed at least partially.

Table 4
Average correlations between judged evidential impact and confirmation measures (confirmation computed from objective probabilities)

| Confirmation measure | Average correlation with *Judged* $(e,A)$ | Average correlation with *Judged* $(e,B)$ |
|---|---|---|
| $d$ | .537 | .589 |
| r | .658* | .587 |
| $n$ | .730* | .745* |
| $\ell$ | .743* | .755* |
| c | .586 | .605 |
| $k$ | .740* | .754* |
| $p$ | .729* | .743* |
| $Pr_{obj}(A[B]\|e)$ | .488 | .508 |

*Note.* Each number is the average of 26 correlations (one per participant); each correlation involves 10 observations. $Pr_{obj}(A[B]\|e)$ denotes $Pr_{obj}(A|e)$ or $Pr_{obj}(B|e)$ as appropriate. Starred averages are reliably greater than the average for $Pr_{obj}(A[B]\|e)$ by paired $t$-test (prob $< .02$).

Table 5
Comparison of $\ell$ and $k$ with other confirmation measures (confirmation computed from objective probabilities)

| | $d$ | $r$ | $n$ | $c$ | $p$ |
|---|---|---|---|---|---|
| *Predicting Judged* $(e,A)$ | | | | | |
| $\ell$ | $t = 3.7$ | $t = 2.6$ | $t = 2.6$ | $t = 3.3$ | $t = 2.3$ |
| | prob $< .001$ | prob $< .05$ | prob $< .05$ | prob $< .005$ | prob $< .05$ |
| | 21 | 16 | 22 | 19 | 21 |
| $k$ | $t = 3.7$ | $t = 2.5$ | $t = 3.0$ | $t = 3.3$ | $t = 2.4$ |
| | prob $< .001$ | prob $< .05$ | prob $< .01$ | prob $< .005$ | prob $< .05$ |
| | 21 | 16 | 21 | 19 | 21 |
| *Predicting Judged* $(e,B)$ | | | | | |
| $\ell$ | $t = 3.2$ | $t = 3.2$ | $t = 2.2$ | $t = 2.9$ | $t = 2.0$ |
| | prob $< .005$ | prob $< .005$ | prob $< .05$ | prob $< .01$ | prob $< .06$ |
| | 20 | 16 | 19 | 20 | 19 |
| $k$ | $t = 3.2$ | $t = 3.2$ | $t = 2.8$ | $t = 2.9$ | $t = 2.3$ |
| | prob $< .005$ | prob $< .005$ | prob $< .01$ | prob $< .01$ | prob $< .05$ |
| | 20 | 16 | 20 | 20 | 19 |

*Note.* Each cell reports a paired t-test between the correlations obtained with the confirmation measures in the associated row and column. For each test, $N = 26$ (corresponding to the 26 participants). The correlations each involve 10 observations. The last row of each cell shows the number of participants (out of 26) for whom $\ell$ (or $k$) predicted better than the rival measure at the top of the column.

Hence, for each participant, each confirmation measure $x$, and each extraction $e_i$, we computed $x(e_i,A)$ on the basis of $Pr_{obj}(A|e)$, $Pr_{obj}(e|A)$, and $Pr_{obj}(A)$. For each participant the Pearson correlation was then calculated between $x(e_i,A)$ and *Judged* $(e_i,A)(N = 10)$, and similarly for urn $B$. Averages are shown in Table 4.

Once again, $\ell$ and $k$ yield the highest average correlations, and they are improved by the use of objective compared to subjective inputs. Moreover, by paired $t$-test, both $\ell$ and $k$ are reliably superior to every other measure. See Table 5. There is no reliable difference between $\ell$ and $k$. This is not surprising given that they are

monotonically related (as discussed above), and hardly discernable in the Monte Carlo procedure described earlier.

## 5. Discussion

The confirmation measure $k$ satisfies weak but apparently necessary conditions on normative adequacy, and is sole among our candidates to do so. It is also among the best predictors of evidential impact in our experiment, its only rival ($\ell$) being ordinally equivalent. This coincidence raises the possibility that judgments of confirmation have greater normative merit than direct assessments of chance, which are often observed to be deficient.[5] Questions formulated in terms of opinion-change might allow biases to cancel out. For example, the same multiplicative bias applied to the likelihood of $e$ under both $H$ and $\neg H$ would have no effect on judgment of confirmation if the latter conforms to $k$. Even in purely normative terms, the concept of confirmation can illuminate reasoning puzzles in ways supplementary to standard probabilistic analysis (see Bradley & Fitelson, 2003).

Further research is needed to confirm the descriptive adequacy of $k$. Even in our limited urn-setting not every plausible measure of confirmation could be tested. For example, we omitted

$$s(e,H) = Pr(H|e) - Pr(H|\neg e) \text{ (Christensen, 1999)}$$

because of its reliance on $Pr(H|\neg e)$, whose assessment we did not wish to add to the experiment. It can be shown that $s$ conforms to the discredited principles (ES) and (TS) (Eells & Fitelson, 2002), and may fail to reach a maximum when $e$ entails $h$. It would therefore be interesting to compare the predictive accuracy of $s$ with that of $k$.

When people are unsure about the probabilities in play, verbally announced estimates of chance may not be directly related to feelings of confidence (Fox & Tversky, 1995; Heath & Tversky, 1991). It is therefore essential to supplement experiments like ours with more naturalistic settings that generate diverse levels of clarity about the relevant probabilities (e.g., the scenarios employed in Briggs & Krantz, 1992). Testing the generality of the present findings also requires that confirmation be measured in the presence of strong prior beliefs, perhaps emotionally charged (e.g., concerning the effectiveness of capital punishment).

It is also worth observing that the approach taken here derives confirmation from probability and likelihood. Yet it might turn out that evidential impact is psychologically prior to credence, or that each intuition typically constrains mental construction of the other. In all events, confirmation appears to be central to understanding the origin of belief in our minds, as well as its transformation through time.

[5] For example, when asked to translate evidence into probability, intuition often ignores basic principles about sample size and base rates (Griffin & Tversky, 1992). The interpretation of base-rate use remains controversial, however. See Gigerenzer and Hoffrage (1995), Koehler (1996), Stanovich (1999) for discussion.

## References


Bradley, D., & Fitelson, B. (2003). Monty hall, doomsday, and confirmation. *Analysis, 63*(1), 23–31.

Briggs, L. K., & Krantz, D. H. (1992). Judging the strength of designated evidence. *Journal of Behavioral Decision Making, 5*, 77–106.

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

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

Cheng, P. W., & Novick, L. R. (1990). A probabilistic contrast model of causal induction. *Journal of Personality and Social Psychology, 58*, 545–567.

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

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

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

Eells, E. (1982). *Rational decision and causality*. Cambridge UK: Cambridge University Press.

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

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

Fox, C., & Tversky, A. (1995). Ambiguity aversion and comparative ignorance. *Quarterly Journal of Economics, 110*, 585–603.

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: frequency formats. *Psychological Review, 102*, 684–704.

Good, I. J. (1984). The best explicatum for weight of evidence. *Journal of Statistical Computation and Simulation, 19*, 294–299.

Griffin, D., & Tversky, A. (1992). The weighing of evidence and the determinants of confidence. *Cognitive Psychology, 24*, 411–435.

Heath, C., & Tversky, A. (1991). Preference and belief: ambiguity and competence in choice under uncertainty. *Journal of Risk and Uncertainty, 4*, 5–28.

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

Jeffrey, R. (1992). *Probability and the art of judgment*. Cambridge UK: Cambridge University Press.

Kemeny, J., & Oppenheim, P. (1952). Degrees of factual support. *Philosophy of Science, 19*, 307–324.

Keynes, J. (1921). *A treatise on probability*. London: Macmillan.

Koehler, J. (1996). The base rate fallacy reconsidered: descriptive, normative and methodological challenges. *Behavioral and Brain Sciences, 19*, 1–53.

Nozick, R. (1981). *Philosophical explanations*. Oxford UK: Clarendon Press.

Perales, J., & Shanks, D. R. (2003). Normative and descriptive accounts of the influence of power and contingency on causal judgement. *The Quarterly Journal of Experimental Psychology, 56A*(6), 977–1007.

Philips, L. D., & Edwards, W. (1966). Conservatism in a simple probability inference task. *Journal of Experimental Psychology, 72*, 346–354.

Shafer, G. (1976). *A mathematical theory of evidence*. Princeton NJ: Princeton University Press.

Slovic, P., & Lichtenstein, S. (1971). Comparison of Bayesian and regression approaches to the study of information processing in judgment. *Organizational Behavior and Human Performance, 6*, 649–744.

Stanovich, K. E. (1999). *Who is rational?*. Mahwah NJ: Lawrence Erlbaum Associates.

Tentori, K., Bonini, N., & Osherson, D. (2004). The conjunction fallacy: a misunderstanding about conjunction? *Cognitive Science, 28*(3), 467–477.

Windschitl, P. (2002). Judging the accuracy of a likelihood judgment: the case of smoking risk. *Journal of Behavioral Decision Making, 15*, 19–35.