Psychonomic Bulletin & Review
2007, 14 (5), 877-883

Determinants of confirmation

Katya Tentori
Università degli Studi di Trento, Rovereto, Italy

Vincenzo Crupi
IUAV, Venice, Italy
Università degli Studi di Trento, Rovereto, Italy
and Université de Provence, Marseille, France

and

Daniel Osherson
Princeton University, Princeton, New Jersey

Epistemologists often suppose that the extent to which evidence *e* confirms hypothesis *H* depends on probabilities involving *e* and *H*, and nothing more. We show experimentally that human reasoners sometimes violate this assumption.

If verifying the truth of one statement increases the initial presumption for another then epistemologists often say that the first *confirms* the second.[1] This relation is distinct from the posterior probability of the second statement in light of the first. Consider, for example:

(1) (A) Bill owns a mountain bike.
(B) Bill works in an office.
(C) Steve talks into his sleeve.
(D) Steve works for the Secret Service.

The reader will perhaps concur that the probability of (1B) given (1A) exceeds that of (1D) given (1C), because there are so few Secret Service agents as compared with office workers. Yet the reader may have the additional intuition that (1C) provides more evidence for (1D) than (1A) provides for (1B)—in other words, that (1C) confirms (1D) more than (1A) confirms (1B). The present inquiry evaluates a hypothesis about the variables that determine judgments of the latter kind. To formulate the hypothesis, let us first note a striking consistency among proposals for measuring confirmation.

A variety of such measures have been advanced, including those shown in Table 1. Each proposal assumes that statements confirm each other through their probabilities. To be more precise, fix an agent A, and assume her to be *probabilistically coherent*; that is, assume that A uses numbers to represent chances in a manner consistent with the standard axioms of probability.[2] The probability function that A relies on will be denoted Pr. For statements e, H, let CONF(e, H) be the confirmation that A perceives e to offer H.

(2) **Definition**: A is *formal* if and only if CONF(e, H) depends just on $Pr(e\wedge H)$, $Pr(e\wedge\neg H)$, $Pr(\neg e\wedge H)$, and $Pr(\neg e\wedge\neg H)$.

Formality is more inclusive than the definition might suggest. Since Pr is coherent, it is equivalent to call A formal in just the case that CONF(e, H) is a function of (only) quantities that can be defined from $Pr(e\wedge H)$, $Pr(e\wedge\neg H)$, $Pr(\neg e\wedge H)$, and $Pr(\neg e\wedge\neg H)$. Such quantities include any absolute or conditional probability in which only e, H, and their negations appear—for example, $Pr(H)$, $Pr(H \mid e)$, $Pr(e \mid H)$, and all the other terms seen in Table 1.

The hypothesis to be investigated is simply

(3) **Hypothesis**: Human reasoning agents are formal.

To appreciate the content of this hypothesis, consider what would make it false. If A's perception of confirmation is affected by statement content beyond the latter's contribution to the probabilities evoked in Definition 2, then A is not formal. For example, throwing snake eyes in successive rolls (e) might confirm

H_1: the dice are loaded

H_2: you will be in foul temper at the end of the evening

to different extents, even if it turns out that

$$Pr(e\wedge H_1) = Pr(e\wedge H_2),$$
$$Pr(e\wedge\neg H_1) = Pr(e\wedge\neg H_2),$$
$$Pr(\neg e\wedge H_1) = Pr(\neg e\wedge H_2),$$

K. Tentori, katya.tentori@unitn.it

Table 1
Alternative Measures of Confirmation

$d(e, H) = Pr(H \mid e) - Pr(H)$		(Eells, 1982; Jeffrey, 1992)
$r(e, H) = \log\left[\frac{Pr(H \mid e)}{Pr(H)}\right]$		(Keynes, 1921; Horwich, 1982)
$n(e, H) = Pr(e \mid H) - Pr(e \mid \neg H)$		(Nozick, 1981)
$l(e, H) = \log\left[\frac{Pr(e \mid H)}{Pr(e \mid \neg H)}\right]$		(Good, 1984)
$c(e, H) = Pr(H \wedge e) - [Pr(e) \times Pr(H)]$		(Carnap, 1962)
$k(e, H) = \frac{Pr(e \mid H) - Pr(e \mid \neg H)}{Pr(e \mid H) + Pr(e \mid \neg H)}$		(Kemeny & Oppenheim, 1952)
$s(e, H) = Pr(H \mid e) - Pr(H \mid \neg e)$		(Christensen, 1999)
$z(e, H) = \begin{cases} \frac{Pr(H \mid e) - Pr(H)}{Pr(\neg H)} \\ \frac{Pr(H \mid e) - Pr(H)}{Pr(H)} \end{cases}$	if $Pr(H \mid e) \geq Pr(H)$ otherwise	(Crupi et al., in press)

Note—Each maps an evidence statement e and hypothesis H into a real number intended to measure the confirmation that e provides for H. We rely on Eells and Fitelson (2002) for some literature citations. Measure l is the log of the "Bayes factor" (Jeffrey, 2004), which may have been originally introduced by Alan Turing (according to Good, 1984).

and

$$Pr(\neg e \wedge \neg H_1) = Pr(\neg e \wedge \neg H_2).$$

Such judgments violate formality inasmuch as they imply that some variable beyond those encompassed by Definition 2 influences *A*'s judgment of confirmation. The additional variable may or may not have a probabilistic character (see the Discussion section), but it must involve more than the kind of point estimates seen in Table 1.

It is worth emphasizing that Hypothesis 3 is not connected to issues of human rationality. At least, there seems to be no *evident* defect in estimates of confirmation that depend on more than the probabilities listed in Definition 2. In contrast, normative concern is justified if people sometimes confuse confirmation with conditional probability—as suggested, for example, in Sides et al. (2002). Such confusion can lead to probabilistic incoherence—specifically, to conjunction fallacies.[3] But none of this is relevant to the present context, where coherence is assumed (and enforced by our experimental procedure).

The assumption that epistemic agents are formal has produced many philosophical insights.[4] Here we investigate the empirical question of whether human reasoners satisfy the same assumption—that is, we test Hypothesis 3. For this purpose, we constructed pairs e, H of statements with distinct content that engender the same estimates of probability. Formality requires that judgments of confirmation coincide across content.

EXPERIMENT 1

Thirty-two students (17 female) from the University of Trento participated in exchange for course credit (mean age 23). In what follows, we use *A* to denote a given participant in the study. *A* was asked to issue judgments of confirmation and probability about two scenarios, here called *rich* and *lean*, respectively.

Rich Scenario

The rich scenario involved the extraction of individuals from a random sample consisting of 100 Italian women and 100 Italian men. Each drawn individual *X* was qualified by exactly one of the predicates e_r shown in Table 2 (the "r" denotes *rich*). *A* was asked how much the information that *X* satisfies e_r influenced her opinion that *X* is male. Each such question was based on a single proposition—namely, that *X* satisfies e_r; there was no accumulation of evidence across multiple propositions. Twelve independent extractions were imagined, one for each of the 12 predicates, presented in individually randomized order. The predicates were constructed on the basis of a pilot study to be roughly balanced between weakening and strengthening the hypothesis that *X* is male.

Here is more detail about how the judgments were elicited. For each draw *X*, *A* first concurred that 1/2 was the prior probability of the hypothesis that *X* is male (by "prior" is meant prior to presenting the predicate, which served as evidence). After the evidence was given, *A* chose one of the following three descriptions of its impact.

Table 2
Predicates in the Rich Scenario

Confirming for *is a male*	Disconfirming for *is a male*
. . . likes cigars	. . . likes aerobic dance
. . . does not like shopping	. . . does not like soccer
. . . does not like ballet	. . . does not like sports cars
. . . likes small scale models	. . . likes décor magazines
. . . likes bricolage	. . . likes herb teas
. . . does not like skating	. . . does not like beer

Note—A pilot study suggested that predicates in the left column would tend to confirm the hypothesis that the drawn individual is male, whereas predicates in the right column would be disconfirmatory to a similar extent. The pilot study also suggested that predicates higher in the left-hand list would be more confirmatory of the hypothesis, and that predicates higher in the right-hand list would be more disconfirmatory. All predicates are translated from Italian.

(4) (A) *weakens* (*at least a little) my belief that X is male*

(B) *has no influence (not even a little) on my belief that X is male*

(C) *strengthens* (*at least a little) my belief that X is male*

If *A* chose Answer 4A or 4C, she was asked to quantify the judgment by indicating a position on a scale marked from −10 (*completely weakens my belief*) to +10 (*completely strengthens my belief*), passing through 0 (*has no influence at all on my belief*). The indicated number served as *A*'s estimate of confirmation for the trial. If (4B) was chosen, the estimate was taken to be 0. The whole scale was visible in each trial, but only the relevant half was made available after choice of either (4A) or (4C). Twelve judgments of confirmation were thus collected for *A*, one per predicate.

Subsequently, for each predicate e_r, *A* was asked to estimate

- the number of men in the sample of 100 to which e_r applies,[5]

and

- the number of women in the sample of 100 to which e_r applies.

For ease of notation, we abbreviate the statement that *X* satisfies predicate e_r to just e_r. Letting H_r be the (rich scenario) hypothesis that *X* is male, the estimates elicited in the rich scenario determine

(5) $Pr(e_r \wedge H_r)$, $Pr(e_r \wedge \neg H_r)$, $Pr(\neg e_r \wedge H_r)$, $Pr(\neg e_r \wedge \neg H_r)$

for each predicate e_r. $Pr(e_r \wedge H_r)$ represents the proportion of men satisfying predicate e_r out of the total sample of 200 individuals (100 women and 100 men); $Pr(e_r \wedge \neg H_r)$ represents the proportion of women satisfying predicate e_r out of that same total sample; and so forth. It is therefore clear that $Pr(e_r \wedge H_r) + Pr(\neg e_r \wedge H_r) = Pr(e_r \wedge \neg H_r) + Pr(\neg e_r \wedge \neg H_r) = 1/2$. This is because $Pr(e_r \wedge H_r) + Pr(\neg e_r \wedge H_r) = Pr(H_r)$, which is the prior probability of drawing a man; similarly, $Pr(e_r \wedge \neg H_r) + Pr(\neg e_r \wedge \neg H_r) = Pr(\neg H_r)$ is the prior probability of drawing a women. If *A* is formal in the sense of Definition 2, then the numbers in (5) suffice to predict $\mathrm{CONF}(e_r, H_r)$.

Lean Scenario

The lean scenario consisted of urn problems whose parameters were based on the probabilities (5) recorded in the rich scenario. For each predicate e_r of the rich scenario, *A* was asked to consider an urn with 200 balls composed as follows.

- $100 \times Pr(e_r \wedge H_r)$ red striped balls
- $100 \times Pr(\neg e_r \wedge H_r)$ red spotted balls
- $100 \times Pr(e_r \wedge \neg H_r)$ blue striped balls
- $100 \times Pr(\neg e_r \wedge \neg H_r)$ blue spotted balls

Because $Pr(e_r \wedge H_r) + Pr(\neg e_r \wedge H_r) = Pr(e_r \wedge \neg H_r) + Pr(\neg e_r \wedge \neg H_r) = 1/2$, the urn contains 100 red balls and 100 blue balls, corresponding to the men and women in the rich scenario. The proportion of red balls that are striped corresponds to *A*'s estimate of the number of men satisfying the predicate e_r, and likewise the proportion of striped blue balls corresponds to *A*'s estimate of the number of women satisfying e_r. Note that these proportions were tailored to the individual participant *A*, relying on just *A*'s responses in the rich scenario (there was no averaging).

These numbers were communicated to *A* via a pie chart, with four regions labeled by the appropriate kind of ball and sized to reflect their respective fraction of the total. It was explicitly stated that the urn contained 200 balls evenly divided between red and blue. The subdivision of red and blue into striped and spotted was reflected solely by relative size of their pie slices. (Dots were regularly spaced along the circumference of the chart but no numbers were displayed.)

Relative to this urn, *A* was presented with the hypothesis H_l that a drawn ball is red and concurred that the prior probability of H_l is 1/2. Then, *A* was asked to estimate the impact on H_l of learning that the drawn ball is striped. The latter fact serves as evidence e_l in the lean scenario. The same options (Answers 4A–4C) were employed to elicit judgments, followed by the same scale as before. The pie chart remained on the screen until the estimate of impact was collected. As a manipulation check, participants were subsequently asked to estimate the proportions of each type of ball in the urns of the lean scenario (with pie charts present). For each urn, these estimates determine $Pr(e_l \wedge H_l)$, $Pr(e_l \wedge \neg H_l)$, $Pr(\neg e_l \wedge H_l)$, and $Pr(\neg e_l \wedge \neg H_l)$.

Participants in the experiment were run individually. The design required that lean scenario probabilities match the corresponding rich scenario probabilities. Participants thus confronted the 12 rich scenarios prior to the lean. Within each scenario, judgments about evidential impact were always elicited before probability estimates. In both scenarios (rich and lean), confirmation judgments were preceded by practice problems. All questions were posed through a computer interface that composed the urns and pie charts of the lean scenario on the basis of answers recorded in the rich scenario. There was ample opportunity to revisit each answer before proceeding.

Overall, the 32 participants estimated the following quantities for each of 12 corresponding evidence pairs e_r, e_l.

(6) (A) In the rich scenario, for the hypothesis H_r = "the selected person is male": $\mathrm{CONF}(e_r, H_r)$, $Pr(e_r \wedge H_r)$, $Pr(e_r \wedge \neg H_r)$

(B) In the lean scenario, for the hypothesis H_l = "the selected ball is red": $\mathrm{CONF}(e_l, H_l)$, $Pr(e_l \wedge H_l)$, $Pr(e_l \wedge \neg H_l)$

In all cases, participants acknowledged that $Pr(H_r) = Pr(H_l) = 1/2$, so $Pr(\neg e_r \wedge H_r)$, $Pr(\neg e_r \wedge \neg H_r)$, $Pr(\neg e_l \wedge H_l)$, and $Pr(\neg e_l \wedge \neg H_l)$ are determined as well.

Results

Let us call each triple of judgments in (6A) or (6B) a *data set*. There were $32 \times 12 = 384$ data sets per scenario. Each implies a value for $Pr(H_r \mid e_r)$ or $Pr(H_l \mid e_l)$.[6] We excluded 11 data sets in the rich scenario because either

$$\mathrm{CONF}(e_r, H_r) > 0 \text{ and } Pr(H_r \mid e_r) < Pr(H_r)\ (= 1/2)$$

or

$$\mathrm{CONF}(e_r, H_r) < 0 \text{ and } Pr(H_r \mid e_r) > Pr(H_r).$$

Similarly, we excluded 4 data sets in the lean scenario because either

$$\mathrm{CONF}(e_l, H_l) > 0 \text{ and } Pr(H_l \mid e_l) < Pr(H_l)\ (= 1/2),$$

or

$$\mathrm{CONF}(e_l, H_l) < 0 \text{ and } Pr(H_l \mid e_l) > Pr(H_l).$$

Such judgments are unintelligible since they imply that evidence e may confirm (disconfirm) hypothesis H despite decreasing (increasing) its initial credibility.[7] The 15 excluded data sets represent less than 2% of the total. There remain 369 matched pairs of data sets—that is, involving corresponding e_r, e_l from which neither (6A) nor (6B) was excluded. Subsequent analyses employ just these matched pairs.

Let us first determine whether, as intended,

$$\begin{aligned}(7)\ & Pr(e_r \wedge H_r) = Pr(e_l \wedge H_l),\\ & Pr(e_r \wedge \neg H_r) = Pr(e_l \wedge \neg H_l),\\ & Pr(\neg e_r \wedge H_r) = Pr(\neg e_l \wedge H_l),\\ & Pr(\neg e_r \wedge \neg H_r) = Pr(\neg e_l \wedge \neg H_l).\end{aligned}$$

Table 3 presents the averages of $Pr(e_r \wedge H_r)$, $Pr(e_l \wedge H_l)$, $Pr(e_r \wedge \neg H_r)$, and $Pr(e_l \wedge \neg H_l)$ across all participants. For matched e_r, e_l, we also computed the difference between a given participant's estimate of $Pr(e_r \wedge H_r)$ versus $Pr(e_l \wedge H_l)$ and between her estimate of $Pr(e_r \wedge \neg H_r)$ versus $Pr(e_l \wedge \neg H_l)$. A total of 94% of the differences were zero (as intended by the design of the experiment). The mean absolute difference for the remaining pairs was only .009, which presumably resulted from the lack of explicit numerical information in the pie charts of the lean scenario. There were 12 average differences between estimates of $Pr(e_r \wedge H_r)$ versus $Pr(e_l \wedge H_l)$, and 12 between estimates of $Pr(e_r \wedge \neg H_r)$ versus $Pr(e_l \wedge \neg H_l)$, for matched e_r, e_l. None of these were revealed to be reliable at $p < .05$ by either paired t test or Wilcoxon signed-rank test for related measures.

Thus, almost exactly, the urns in the lean scenario satisfied (7). In conjunction with the acknowledgment by all participants that $Pr(H_r) = Pr(H_l) = 1/2$, this allows formality to be tested by comparing judgments of confirmation between the rich and lean scenarios. The median values of $\mathrm{CONF}(e_r, H_r)$ and $\mathrm{CONF}(e_l, H_l)$ are shown in Table 4. In 11 of the 12 matched pairs, the difference in confirmation was significant by Wilcoxon signed-rank test ($p < .05$). Thus, a given piece of evidence has distinct impact on the respective hypotheses of the rich and lean scenarios, despite the equality of relevant probabilities.

On the other hand, the median lean confirmation values appear to be a dilation of the rich, and the two are significantly correlated (Kendall's τ-b = .79, $N = 12$, $p < .01$). The median correlation ($N = 32$) between confirmations from the same participant in the two scenarios is .74 (again using Kendall's τ-b; some of these correlations are of length 10 or 11 because of the 15 excluded data sets).

Because probabilities could be controlled only in the lean scenario (via urns), lean estimates always came after rich estimates. Before drawing conclusions from Experiment 1, we must therefore consider the possibility that the greater dispersion seen for estimates of confirmation in the lean scenario is due to its position in the procedure. Experiment 2 controlled for this potential confound.

EXPERIMENT 2 (CONTROL)

Thirty-two students—the same number as in Experiment 1—participated in Experiment 2, who were again recruited from the University of Trento in exchange for course credit (27 female, mean age 21). None had participated in Experiment 1. The students were asked to issue judgments of confirmation and probability about urns in

Table 3
Mean Estimates of $Pr(e \wedge H)$ and $Pr(e \wedge \neg H)$ for Each Predicate and Each Scenario

Predicate	Mean $Pr(e_r \wedge H_r)$	Mean $Pr(e_l \wedge H_l)$	Mean $Pr(e_r \wedge \neg H_r)$	Mean $Pr(e_l \wedge \neg H_l)$	N
likes cigars	.22	.22	.04	.04	31
does not like shopping	.34	.34	.06	.06	29
does not like ballet	.38	.38	.15	.15	30
likes small scale models	.26	.26	.09	.09	31
likes bricolage	.26	.26	.17	.18	31
does not like skating	.30	.30	.17	.17	30
likes aerobic dance	.08	.08	.31	.31	31
does not like soccer	.10	.10	.35	.35	32
does not like sports cars	.09	.09	.25	.25	31
likes décor magazines	.12	.12	.34	.34	30
likes herb teas	.13	.13	.32	.32	31
does not like beer	.08	.08	.26	.26	32

Note—The last column shows the number of participants (out of 32) who contributed to the mean. As explained in the text, 15 data sets were excluded from the analysis, resulting in 369 matched pairs. None of the differences between corresponding probabilities in the two scenarios reach significance by either a Wilcoxon sign test or a t test for related measures.

Table 4
Median of CONF(e, H) for Each Predicate e
Across Rich and Lean Scenarios

Predicate	Median CONF(e_r, H_r)	Median CONF(e_l, H_l)	z Score for Difference	N
likes cigars	6	7	$z = 3.4^*$	31
does not like shopping	4	8	$z = 4.4^*$	29
does not like ballet	4	6	$z = 3.3^*$	30
likes small scale models	4	7	$z = 3.3^*$	31
likes bricolage	3	6	$z = 2.2^*$	31
does not like skating	1.5	4	$z = 3.6^*$	30
likes aerobic dance	−6	−7	$z = -2.3^*$	31
does not like soccer	−4	−7	$z = -3.2^*$	32
does not like sports cars	−4	−4	$z = -1.3$	31
likes décor magazines	−3	−5	$z = -2.9^*$	30
likes herb teas	−3	−5	$z = -4.1^*$	31
does not like beer	−2.5	−5	$z = -4.1^*$	32

Note—Ratings used a scale from −10 to 10 (see the text). The last column shows the number of participants (out of 32) who contributed to the median. $^*p < .05$ by Wilcoxon sign test.

the lean scenario; no material from the rich scenario was presented. Specifically, each participant *B* in Experiment 2 was paired with a unique participant *A* in Experiment 1. The 12 urns constructed for *A* (on the basis of *A*'s responses to the rich scenario) served as stimuli for *B*. The procedure was identical to the lean part of Experiment 1 (in particular, urns were presented to *B* in the same order as for *A*).

Results

In the present experiment, let H_c be the hypothesis that a drawn ball is red, and let e_c be the evidence that the drawn ball is striped ("c" signifies "control"). Corresponding to each of the 12 rich predicates from Experiment 1, a given participant in Experiment 2 evaluated the following data set (which corresponds to [6B] in Experiment 1).

For the hypothesis H_c = "the selected ball is red,"

$$\text{CONF}(e_c, H_c),\ Pr(e_c \wedge H_c),\ Pr(e_c \wedge \neg H_c).$$

None of the judgments in Experiment 2 exhibited the anomalies

$$\text{CONF}(e_c, H_c) > 0 \text{ and } Pr(H_c \mid e_c) < Pr(H_c)\ (= 1/2)$$

or

$$\text{CONF}(e_c, H_c) < 0 \text{ and } Pr(H_c \mid e_c) > Pr(H_c).$$

In contrast, recall that 15 data sets were withdrawn from Experiment 1 because of such anomalies. To ensure comparability between the two experiments, the corresponding 15 data sets were withdrawn from the data of Experiment 2.

For each of the 12 rich scenarios from Experiment 1, we compared the mean estimates for $Pr(e_c \wedge H_c)$ of Experiment 2 with the mean for $Pr(e_l \wedge H_l)$ from the lean scenario of Experiment 1. Likewise, we compared $Pr(e_c \wedge \neg H_c)$ with $Pr(e_l \wedge \neg H_l)$. The two sets of numbers were nearly identical, with no reliable difference in any of the 12 comparisons (via *t* test). The same was true of the respective medians (Mann–Whitney *U* test). The congruence of estimates between the two experiments is not surprising, since they were based on pie charts for the same urns. Similarly, there were no significant differences for any predicate between $Pr(e_c \wedge H_c)$ and $Pr(e_r \wedge H_r)$ or between $Pr(e_c \wedge \neg H_c)$ and $Pr(e_r \wedge \neg H_r)$.

Table 5
Median Values of CONF(e_c, H_c) for Each Predicate e, and z Scores for Mann–Whitney *U* Tests of CONF(e_c, H_c) Versus CONF(e_l, H_l) and CONF(e_r, H_r)

Predicate From the Rich Scenario of Experiment 1	Median CONF(e_c, H_c)	z Score for CONF(e_c, H_c) Versus CONF(e_l, H_l)	z Score for CONF(e_c, H_c) Versus CONF(e_r, H_r)	N
likes cigars	7	$z = -0.5$	$z = -2.2^*$	31
does not like shopping	8	$z = -0.4$	$z = -4.6^*$	29
does not like ballet	6	$z = -0.6$	$z = -3.7^*$	30
likes small scale models	6	$z = -0.6$	$z = -2.6^*$	31
likes bricolage	4	$z = -0.4$	$z = -1.0$	31
does not like skating	4	$z = -0.2$	$z = -3.0^*$	30
likes aerobic dance	−7	$z = -0.02$	$z = -2.0^*$	31
does not like soccer	−7	$z = -0.1$	$z = -3.3^*$	32
does not like sports cars	−6	$z = -0.7$	$z = -1.6$	31
likes décor magazines	−7	$z = -1.4$	$z = -3.8^*$	30
likes herb teas	−6	$z = -0.4$	$z = -3.0^*$	31
does not like beer	−6	$z = -0.5$	$z = -3.9^*$	32

Note—Ratings used a scale from −10 to 10. The last column shows the number of participants (out of 32) who contributed to the medians (the same for all three scenarios). For comparison of CONF(e_r, H_r) versus CONF(e_l, H_l), see Table 4. $^*p < .05$ by Mann–Whitney *U* test.

More important are the confirmation estimates. The second column of Table 5 shows for each scenario the median estimate provided by participants in Experiment 2. (The same information for Experiment 1 appears in Table 4.) For a given predicate, the medians of CONF(e_c, H_c) versus CONF(e_l, H_l) and CONF(e_c, H_c) versus CONF(e_r, H_r) were compared via Mann–Whitney U test. The z scores for these comparisons are also shown in Table 5. It may be seen that across the 12 predicates, none of the comparisons between CONF(e_c, H_c) and CONF(e_l, H_l) reach significance. By contrast, 10 of the 12 comparisons between CONF(e_c, H_c) and CONF(e_r, H_r) are significant [as compared with 11 for CONF(e_l, H_l) versus CONF(e_r, H_r)]. In other words, the impact of evidence in the lean scenario was nearly identical when assessed by itself versus after the rich scenario, and different in both cases from the corresponding impact in the rich scenario.

We conclude that the disparity between confirmation judgments in the rich versus lean scenarios in Experiment 1 did not derive from the lean scenario coming after the rich. Qualitatively similar estimates were seen in the lean judgments of Experiment 2, which were not accompanied by any rich estimates.

DISCUSSION

The results summarized in Table 4 suggest that human reasoners are not formal in the sense of Definition 2, for judgments of confirmation depend on more than probabilities over e and H. The missing argument may be the probability of some other event, but we suspect that it has a different character altogether.

Notice that in comparison with the sharp chances defined by urns, probabilities in the rich scenario seem more affected by personal ignorance than by objective uncertainty alone. These variables are known to influence willingness to bet (Ellsberg, 1961; Heath & Tversky, 1991). They may also lead our reasoner A to lower confidence for her distribution in the rich scenario in comparison with the lean, even though she reports the same distribution in the two cases.[8] Furthermore, suppose that A assesses confirmation consistently with her reported probabilities (via some metric in Table 1) but adjusts the outcome through multiplication with the relevant confidence level (rich vs. lean). Confirmation (either positive or negative) would then be multiplied by greater confidence in the lean scenario. It is easy to see that such a response strategy would explain the dilation phenomenon noted earlier (namely, that estimates of confirmation are more extreme in the lean scenario than in the rich; see Table 4). Variants of this hypothesis (still consistent with dilation) are easy to construct; for example, it suffices to multiply confirmation by any strictly increasing function of confidence. If human judges behave in such a manner, then some confirmation metric definable from just $Pr(e \wedge H)$, $Pr(e \wedge \neg H)$, $Pr(\neg e \wedge H)$, and $Pr(\neg e \wedge \neg H)$ may underlie estimates of evidential impact, but the metric would not be deployed mentally in the simple way suggested by Table 1.

Of course, it is also possible that confirmation judgment does not involve confidence in distributions, and that dilation must be explained as some other kind of content effect. Perhaps the ambiguous probabilities evoked by the rich scenario lead judges to hesitate about the interpretation of evidence, resulting in more conservative estimates of impact without the adjustments envisioned above. Indeed, content has been shown to intervene in many settings, including interpreting logical connectives (Newstead, Griggs, & Chrostowski, 1984; Ray, Reynolds, & Carranza, 1989), testing conditional rules (Cheng & Holyoak, 1985; Kirby, 1994), estimating probabilities (Mellers, Hertwig, & Kahneman, 2001; Sloman, Over, Slovak, & Stibel, 2003), and forming preferences among options (Goldstein & Weber, 1995).

Finally, observe that our experimental procedure was framed in terms of *strengthening or weakening belief* (see Answers 4A–4C, above). It is not guaranteed that the same pattern of results would be obtained under different wording—notably, in terms of *evidential impact*, *support*, or *change in probability*.[9] Divergent results with alternative wording would signal multiple forms of reasoning about evidence. Convergent results would reinforce the conviction that confirmation is a fundamental variable in human judgment.

AUTHOR NOTE

We thank Branden Fitelson, Douglas Medin, Jeff Rouder, and Eric-Jan Wagenmakers for very helpful discussion, and Massimo Vescovi for technical support during the data collection. The research was supported by a PRIN 2005 grant *Le dinamiche della conoscenza nella società dell'informazione* and by a grant from the SMC/Fondazione Cassa di Risparmio di Trento e Rovereto for the CIMeC (University of Trento) research project of Inductive Reasoning. Correspondence related to this article should be addressed to K. Tentori, Università degli Studi di Trento, via Matteo del Ben, 5/B, 38068 Rovereto TN, Italy (e-mail: katya.tentori@unitn.it).

REFERENCES

BONINI, N., TENTORI, K., & OSHERSON, D. (2004). A different conjunction fallacy. *Mind & Language*, **19**, 199-206.

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

CHENG, P. W., & HOLYOAK, K. J. (1985). Pragmatic reasoning schemas. *Cognitive Psychology*, **17**, 391-416.

CHRISTENSEN, D. (1999). Measuring confirmation. *Journal of Philosophy*, **96**, 437-461.

CRUPI, V., TENTORI, K., & GONZALEZ, M. (in press). On Bayesian theories of evidential support: Theoretical and empirical issues. *Philosophy of Science*.

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

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

EELLS, E., & FITELSON, B. (2002). Symmetries and asymmetries in evidential support. *Philosophical Studies*, **107**, 129-142.

ELLSBERG, D. (1961). Risk, ambiguity, and the Savage axioms. *Quarterly Journal of Economics*, **75**, 643-699.

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

GOLDSTEIN, W. M., & WEBER, E. U. (1995). Content and discontent: Indications and implications of domain specificity in preferential decision making. In J. R. Busemeyer, R. Hastie, & D. L. Medin (Eds.), *The psychology of learning and motivation* (Vol. 32, pp. 83-136). San Diego: Academic Press.

GOOD, I. J. (1984). The best explicatum for weight of evidence. *Journal of Statistical Computation & Simulation*, **19**, 294-299.

HACKING, I. (2001). *An introduction to probability and inductive logic*. Cambridge: Cambridge University Press.

HEATH, C., & TVERSKY, A. (1991). Preference and belief: Ambiguity and competence in choice under uncertainty. *Journal of Risk & Uncertainty*, **4**, 5-28.
HORWICH, P. (1982). *Probability and evidence*. Cambridge: Cambridge University Press.
JEFFREY, R. (1992). *Probability and the art of judgement*. Cambridge: Cambridge University Press.
JEFFREY, R. (2004). *Subjective probability: The real thing*. Cambridge: 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.
KIRBY, K. N. (1994). Probabilities and utilities of fictional outcomes in Wason's four-card selection task. *Cognition*, **51**, 1-28.
MELLERS, B. A., HERTWIG, R., & KAHNEMAN, D. (2001). Do frequency representations eliminate conjunction effects? An exercise in adversarial collaboration. *Psychological Science*, **12**, 269-275.
NEWSTEAD, S. E., GRIGGS, R. A., & CHROSTOWSKI, J. J. (1984). Reasoning with realistic disjunctives. *Quarterly Journal of Experimental Psychology*, **36A**, 611-627.
NOZICK, R. (1981). *Philosophical explanations*. Cambridge, MA: Harvard University Press.
RAY, J. L., REYNOLDS, R. A., & CARRANZA, E. (1989). Understanding choice utterances. *Quarterly Journal of Experimental Psychology*, **41A**, 829-848.
ROSENKRANTZ, R. (1977). *Inference, method and decision*. Dordrecht: Reidel.
SIDES, A., OSHERSON, D., BONINI, N., & VIALE, R. (2002). On the reality of the conjunction fallacy. *Memory & Cognition*, **30**, 191-198.
SKYRMS, B. (2000). *Choice and chance: An introduction to inductive logic*. Belmont, CA: Wadsworth.
SLOMAN, S. A., OVER, D., SLOVAK, L., & STIBEL, J. M. (2003). Frequency illusions and other fallacies. *Organizational Behavior & Human Decision Processes*, **91**, 296-309.
TENTORI, K., BONINI, N., & OSHERSON, D. (2004). The conjunction fallacy: A misunderstanding about conjunction? *Cognitive Science*, **28**, 467-477.

NOTES

1. The "initial presumption" terminology appears in Rosenkrantz (1977, p. 168).

2. For more on coherence, see Hacking (2001) and Skyrms (2000).

3. For recent experimental results on conjunction errors, see Bonini, Tentori, and Osherson (2004) and Tentori, Bonini, and Osherson (2004).

4. For examples, see Horwich (1982) and Earman (1992). Doubt that the same confirmation measure can be deployed to resolve different puzzles is expressed in Fitelson (1999).

5. This number was double-checked by also requesting an estimate of the number of men (or women) to which e_r *does not* apply, and verifying that the two numbers sum to 100.

6. For example, $Pr(H_r \mid e_r) = Pr(e_r \wedge H_r) / [Pr(e_r \wedge H_r) + Pr(e_r \wedge \neg H_r)]$.

7. Since $Pr(H_r) = 1/2$, it is easy to verify that $Pr(H_r \mid e_r) > Pr(H_r)$ iff $Pr(e_r \wedge H_r) > Pr(e_r \wedge \neg H_r)$. The probabilities of the latter conjunctions were elicited directly from participants. The same remarks apply to the H_l, e_l.

8. For example, confidence might be lower for the rich scenario because the reported probabilities result from a mixture of alternative distributions whose priors (functioning as weights) have greater variance than for the lean scenario.

9. We owe this point to Branden Fitelson.

(Manuscript received July 24, 2006;
revision accepted for publication December 11, 2006.)