Memory & Cognition
2010, 38 (7), 941-950
doi:10.3758/MC.38.7.941

Broadening the study of inductive reasoning: Confirmation judgments with uncertain evidence

Tommaso Mastropasqua
University of Trento, Trento, Italy

Vincenzo Crupi
Ludwig Maximilian University, Munich, Germany

and

Katya Tentori
University College London, London, England
and University of Trento, Trento, Italy

Although evidence in real life is often uncertain, the psychology of inductive reasoning has, so far, been confined to certain evidence. The present study extends previous research by investigating whether people properly estimate the impact of uncertain evidence on a given hypothesis. Two experiments are reported, in which the uncertainty of evidence is explicitly (by means of numerical values) versus implicitly (by means of ambiguous pictures) manipulated. The results show that people's judgments are highly correlated with those predicted by normatively sound Bayesian measures of impact. This sensitivity to the degree of evidential uncertainty supports the centrality of inductive reasoning in cognition and opens the path to the study of this issue in more naturalistic settings.

Many human activities rely on people's ability to elaborate relevant inferences from limited information available. Most of these inferences are instances of inductive reasoning, meaning that the conclusions involved do not necessarily follow from given evidence or previously available background knowledge. Such conclusions may concern the future, as in the prediction of expected clinical manifestations of an ongoing infection, or the past, as in the diagnosis, based on currently observed symptoms, of a formerly contracted infection (Sloman & Lagnado, 2005). Inductive reasoning is also crucial for almost any learning activity, such as those related to word meanings, causal relationships, and many other aspects of the world (Tenenbaum, Griffiths, & Kemp, 2006). Accordingly, inductive reasoning is seen as an essential element in human intelligence (Tomic & Kingma, 1998), as well as in one of its highest achievements, scientific knowledge (Baron, 2008; Howson & Urbach, 2006).

However, current understanding of the cognitive processes involved in inductive reasoning is still limited, and a remarkable number of issues remain open. Among these, the uncertainty of evidence available for inductive reasoning will be the issue explored in the present contribution. Indeed, despite the fact that evidence in real life is often uncertain, the study of inductive reasoning has so far been confined to *certain* evidence.

Uncertainty is widely recognized as a ubiquitous challenge for human cognition and theories thereof (see, e.g., Hastie & Dawes, 2001; Jeffrey, 1992; Oaksford & Chater, 2007). Nonetheless, major theoretical accounts of inductive reasoning typically assume some evidence to be known with certainty and to play a crucial role. Bayesianism is no exception, at least in its textbook versions (Hartmann, 2008): A Bayesian agent is supposed to evaluate hypotheses by probabilistically conditionalizing on data that are acquired as certain.

Psychological research on inductive reasoning has also largely focused on ascertained evidence. For instance, from seminal inquiries up to more recent developments, the categorical induction paradigm presents participants with the consideration of a hypothesis/conclusion (e.g., "birds have an ulnar artery") as possibly supported by an allegedly known fact given as a premise (e.g., "robins have an ulnar artery") (see, e.g., Blok, Medin, & Osherson, 2007; Blok, Osherson, & Medin, 2007; Heit, 1998; Kemp & Tenenbaum, 2009; Medin, Coley, Storms, & Hayes, 2003; Osherson, Smith, Wilkie, López, & Shafir, 1990).

As useful as it may be for epistemological analysis, the assumption that evidence is certain amounts to a rather crude simplification in psychological terms; it is rarely met in real settings. In a murder trial, for instance, the defendant's presence at the scene of the crime may

K. Tentori, katya.tentori@unitn.it

be highly relevant for the hypothesis of guilt, but it can hardly be completely ascertained in a court of law. At best, a DNA test or reliable testimony can make this element of evidence very probable. Indeed, in a variety of real-life situations, people need to assess the impact of a piece of evidence without its probability reaching extreme values.

Now that considerable amounts of data and theorizing have been accumulated on inductive reasoning from certain evidence, it seems of interest to extend empirical investigation beyond the limits of this framework, in order to address how uncertain evidence is employed in hypothesis evaluation. In what follows, we present two experiments concerning assessments of the inductive impact of uncertain evidence. More precisely, we investigate how inductive confirmation (sometimes also called *inductive strength*) is assessed when evidence is uncertain. It will be seen shortly how confirmation judgments relate to and differ from estimates of posterior probability. In order to settle the appropriate normative benchmark for the experimental tasks employed, we will need to illustrate the relevant theoretical framework that extends the basic Bayesian account to the uncertain evidence case.

Jeffrey's Rule of Conditionalization

Consider a pair of complementary hypotheses of interest, h and not-h (extending the following treatment to any richer partition is straightforward). In the Bayesian framework, it is assumed that, at a given time t, the belief state of an agent is represented by a probability function P_t defined over h and not-h. It may occur that, from time t to $t+1$, the agent experiences a change in opinion concerning a further statement e, provided that $P_t(e)$ is not extreme to begin with—that is, $0 < P_t(e) < 1$. One important question is, then, how should the agent's beliefs in h and not-h change as a consequence?

Up to the 1960s, Bayesians had a ready answer only for the special case in which, at time $t+1$, the agent has come to believe that e is certainly true, so that $P_{t+1}(e) = 1$, and, correspondingly, $P_{t+1}(\text{not-}e) = 0$. "Classical" or "strict" Bayesian conditionalization postulates that

$$\text{if } P_{t+1}(e) = 1, \text{ then } P_{t+1}(h) = P_t(h \mid e). \tag{1}$$

However, it may surely also occur that the agent's degree of belief in e changes from time t to $t+1$ without reaching certainty. What would the value of $P_{t+1}(h)$ be then? Jeffrey (1965, chap. 11) suggested a natural way to generalize classical Bayesian conditionalization (see also Jeffrey, 2004, pp. 53–55). In Jeffrey conditionalization, it is assumed that

$$P_{t+1}(h) = P_t(h \mid e)P_{t+1}(e) + P_t(h \mid \text{not-}e)P_{t+1}(\text{not-}e). \tag{2}$$

Thus, $P_{t+1}(h)$ is computed as an average of the "old" conditional probabilities of h on e versus not-e, weighted by the current probabilities of e and not-e, respectively.

It is easy to see that Jeffrey conditionalization is a proper generalization of the classical Bayesian account. In fact, when e does become certainly true, so that $P_{t+1}(e) = 1$, Equation 2 immediately reduces to Equation 1. In all other cases, a change in belief about e prompts updating of the prior probability $P_t(h)$ to a new value $P_{t+1}(h)$, which is not identical to the conditional $P_t(h \mid e)$ but, rather, lies between the latter and $P_t(h \mid \text{not-}e)$.

From a theoretical standpoint, Jeffrey's generalized rule of conditionalization is both elegant and plausible. Indeed, in virtue of the theorem of total probabilities, this updating rule turns out to be mandatory through mere probabilistic coherence once it is assumed that $P_t(h \mid e) = P_{t+1}(h \mid e)$ and $P_t(h \mid \text{not-}e) = P_{t+1}(h \mid \text{not-}e)$—a condition called *rigidity* (Jeffrey, 1965, chap. 11) or *invariance* (Jeffrey, 2004, p. 52; for a discussion of the rigidity condition in psychology, see Oaksford & Chater, 2007, pp. 113ff).

Concerns have been recurrently raised that the Jeffrey conditionalization lacks a form of commutativity, suggesting dependence on the mere order of occurrence of allegedly identical episodes of uncertain learning from experience (for a discussion, see Lange, 2000). This worry has been more recently dispelled, however, by Wagner's (2002) proof that, once "identical learning" is appropriately formalized, Jeffrey's rule does commute across order.

Along with Jeffrey's, another influential treatment of probability updating on uncertain evidence was devised by Pearl (1988). Labeled the *method of virtual evidence*, it exploits the powerful formalism of Bayesian networks. It is worth noting, thus, that Chan and Darwiche (2005) provided mathematical results to the effect that one can neatly translate any of Jeffrey's and Pearl's machinery into the other.

From Conditionalization to Inductive Confirmation

As explained in the previous section, Jeffrey's rule of conditionalization provides an answer to the question, How can the probability of a hypothesis h be updated in light of uncertain evidence e? That is, How can $P_{t+1}(h)$ be computed from P_t when $0 < P_{t+1}(e) < 1$? Posterior probability is, of course, a crucial notion in the study of inductive reasoning, but it does not exhaust the topic. In particular, within a probabilistic analysis of inductive reasoning, there is a major conceptual difference between posterior probability and inductive confirmation (see, e.g., Carnap, 1950/1962; Fitelson, 1999). Indeed, as a matter of logical analysis, one can convincingly argue that confirmation is the very core notion in the study of induction (for a neat statement, see Fitelson, 2006). Unfortunately, with a few notable exceptions (e.g., Sides, Osherson, Bonini, & Viale, 2002), a clear distinction between posterior probability and confirmation is seldom spelled out in psychological investigations of human inductive reasoning. It should be emphasized, however, that the notion of inductive confirmation often lies behind psychological treatments of the "inductive strength" of arguments, a much more familiar label in the empirical study of reasoning (for a clear example of the connection, see Rips, 2001; see especially note 1).

Inductive confirmation is a relative notion in a very crucial sense: The credibility of a hypothesis can be changed in either a positive (confirmation in a narrow sense) or negative (disconfirmation) way by a given piece of evidence. Confirmation (in the narrow sense) thus reflects an increase from prior to posterior probability, whereas disconfirmation reflects a decrease. As a consequence, the degree of confirmation is not the same as the posterior probability. To illustrate, the probability of an otherwise very rare disease (h) can be quite low even after a relevant positive test result (e); yet h is inductively confirmed by e to the extent that its probability has risen. By the same token, the probability of the absence of the disease (not-h) can be quite high despite the positive test result (e), yet not-h is disconfirmed by e to the extent that its probability has decreased. Confirmation concerns the relationship between prior and posterior probability, so there is simply no single probability value that can capture the notion, much as the heating (or cooling) of an environment cannot be represented by any single temperature.

Although seldom highlighted or analyzed under this heading, assessments of confirmation relations are arguably common in daily life, as well as in expert practices. Consider the following example. A father is suspected of abusing his son. The child does claim he has been abused. The forensic psychiatrist, when consulted, upholds that this is evidence for guilt. But now suppose that the child is asked and does *not* report having been abused. As noticed by Dawes (2001), it may well happen that the forensic psychiatrist nonetheless interprets *this* as evidence for guilt (suggesting that violence prompted the removal). Of course, the two judgments seem inconsistent and thus untenable on a purely logical basis. The reason lies in the following entirely general and compelling principle: e is evidence for h if and only if (iff) not-e is evidence against h.

Can the latter principle be captured by considering posterior probabilities alone? The answer is no. This is because, mathematically, there is no fixed relationship constraining the values of $P(h \mid e)$ and $P(h \mid \text{not-}e)$; they can both be high, for instance, or both low. On the contrary, the notion of confirmation/disconfirmation as an increase/decrease in probability does yield the desired principle as a matter of course, for, demonstrably, e confirms h iff not-e disconfirms h—that is, $P(h \mid e) > P(h)$ iff $P(h \mid \text{not-}e) < P(h)$. As a consequence, the hypothetical judgments of the forensic psychiatrist mentioned above can only be sensibly understood and logically criticized as distinctively representing judgments of confirmation.

Confirmation by Uncertain Evidence

A natural way to measure inductive confirmation amounts to positing a function $c_{t,t+1}(h)$ mapping a relevant set of probability values from P_t and P_{t+1} onto a number that is positive, null, or negative, depending on the posterior of h being higher, equal to, or lower than its prior—that is,

$$c_{t,\,t+1}(h)\begin{cases} > 0 & \text{if} \quad P_{t+1}(h) > P_t(h). \\ = 0 & \text{if} \quad P_{t+1}(h) = P_t(h). \\ < 0 & \text{if} \quad P_{t+1}(h) < P_t(h). \end{cases} \tag{3}$$

Various alternative measures of confirmation that satisfy this basic constraint have been proposed and defended (see Crupi, Festa, & Buttasi, 2010; Crupi, Tentori, & Gonzalez, 2007; Festa, 1999; and Fitelson, 1999). As shown by Crupi, Festa, and Mastropasqua (2008), moreover, major confirmation measures can be defined in a completely general fashion—that is, not depending on the particular rule of conditionalization leading from $P_t(h)$ to $P_{t+1}(h)$. In this way, they can be readily applied when the credibility of hypothesis h is affected by a change in the probability of some relevant piece of evidence e that does not attain certainty. In what follows, we focus on the following measures of inductive confirmation. For brevity of notation, O denotes odds, so that $O_t(h) = P_t(h)/P_t(\text{not-}h)$ and $O_{t+1}(h) = P_{t+1}(h)/P_{t+1}(\text{not-}h)$.

$$L_{t,t+1}(h) = \frac{O_{t+1}(h) - O_t(h)}{O_{t+1}(h) + O_t(h)} \tag{4A}$$

$$Z_{t,t+1}(h) = \begin{cases} \dfrac{P_{t+1}(h) - P_t(h)}{1 - P_t(h)} & \text{if} \quad P_{t+1}(h) \geq P_t(h); \\ \dfrac{P_{t+1}(h) - P_t(h)}{P_t(h)} & \text{otherwise.} \end{cases} \tag{4B}$$

Measure L is strictly connected with the log likelihood ratio measure first conceived by Alan Turing (as reported by Good, 1950, pp. 62–63; see also Fitelson, 2001; Kemeny & Oppenheim, 1952).[1] Measure Z has been recently advocated by Crupi et al. (2007); other occurrences include Rescher (1958, p. 87) and Shortliffe and Buchanan (1975) (see also Mura, 2006, 2008).

Although nonequivalent in general terms, measures L and Z share a number of properties that single them out as being particularly appealing as normative models (see Crupi et al., 2007; Eells & Fitelson, 2002; Fitelson, 2006). Among other things, each of measures L and Z achieves a fixed finite maximum (minimum) value $+1$ (-1) in the limiting case of an ascertained piece of evidence e implying (contradicting) h, thus matching in a natural way the bounded, bidirectional, and symmetric rating scale employed in our experiments.

Previous research has shown that, with some piece of certain evidence having been acquired, intuitive assessments of inductive confirmation can be elicited directly, because people prove able to appropriately distinguish between posteriors and degrees of confirmation (Tentori, Crupi, Bonini, & Osherson, 2007). It has also been observed that intuitive confirmation judgments based on ascertained evidence tend to conform to normatively appealing models, such as L and Z above (Crupi et al., 2007). The main goal of our inquiry is to investigate whether such conclusions can be extended to scenarios involving judgments of confirmation by uncertain evidence. The experiments below test the understanding of confirmation in human judgment—as distinct from posterior probability—in the extended and virtually unexplored domain of inductive reasoning with uncertain evidence.

EXPERIMENT 1

Experiment 1 was conceived as a first test of the descriptive adequacy of measures L and Z relative to judgments of confirmation by uncertain evidence. The degree

Table 1
The Seven Levels of Uncertain Evidence and Four Hypotheses Appearing in the Inductive Arguments Employed

Information About Uncertain Evidence
The drawn student is male with probability [100%; 80%; 70%; 50%; 30%; 20%; 0%]. female with probability [0%; 20%; 30%; 50%; 70%; 80%; 100%].
Hypotheses
The drawn student [owns a €10,000 motorbike; owns a €10,000 necklace; usually has a shaved beard; usually applies eye makeup].

of uncertainty of evidence was manipulated by a purposely devised sampling procedure.

Method

Thirty-three students (17 female, mean age = 25 years) from the University of Trento participated in Experiment 1 in exchange for course credit.

The participants performed two tasks: a confirmation task first, then a probability task.[2] A custom Java application was used for stimulus presentation and to collect participants' responses.

Confirmation task. The participants were presented seven sets of four inductive arguments each. The four arguments in a set each involved an identical piece of evidence and a different hypothesis. The probability of evidence varied across the seven sets (seven levels, one for each set, ranging between 100% and 0%; see Table 1) and was manipulated by means of the following scenario:[3]

> Consider a group of 1,000 students, **500 males** and **500 females**, randomly selected at the University of Trento. For the sake of convenience, these 1,000 students have been ordered alphabetically by their surname, from A to Z. Starting from the beginning of the alphabetical list, separation lines have been entered after each set of 10 students, as shown below. [The relevant graphical display was provided.] In this way, the 1,000 students have been divided into **100 groups**, each formed by **10 students**. In what follows we will repeatedly draw at random 1 among the 100 groups of students, then again 1 at random among the 10 students in that group. Draws will be independent at each trial (so, in principle, the same student might be selected more than once).

The gender of the drawn student represented the relevant evidence, and the double sampling procedure (i.e., first drawing a group, then a student from that group) provided a plausible way to manipulate probability. For example, participants concurred that a student drawn from a group of 8 males and 2 females had a .8 probability of being male versus a .2 probability of being female.

After the student was said to have been drawn, participants were presented a set of four inductive arguments, each involving the same information about the probability of the student being male versus female, coupled with one among four different hypotheses. An example of an argument as displayed in the experiment is shown in the following text:

INFORMATION (surely true):
the drawn student is
male with 80% probability
female with 20% probability

HYPOTHESIS (can be true or false):
the drawn student
owns a €10,000 motorbike

Let us briefly illustrate the connection between the above example and the theoretical framework presented in the introductory section. Assuming that, in the reasoner's view, the drawn student is more likely to own a €10,000 motorbike if male than otherwise—that is, that $P_t(\text{motorbike} \mid \text{male}) > P_t(\text{motorbike})$—the above argument displays a case of positive confirmatory impact from uncertain evidence concerning gender. In fact, by the information given, the probability of a piece of evidence (male) strictly confirming the hypothesis (motorbike) has risen from its baseline value $P_t(\text{male}) = .5$ to $P_{t+1}(\text{male}) = .8$, so that, by Jeffrey's rule (Equation 2), $P_{t+1}(\text{motorbike}) > P_t(\text{motorbike})$. In turn, by Equations 3 and 4, a positive value obtains as a degree of confirmation (the numerators of both measures L and Z are positive). Notably, this all occurs despite the posterior $P_{t+1}(\text{motorbike})$ presumably remaining low or moderate.

The participants were asked to estimate inductive confirmation concerning the four arguments presented. They were instructed to drag each argument icon onto an "impact scale," thus assigning it a value. The scale (see Figure 1) had two opposite directions, corresponding to positive and negative impact, respectively, as well as a neutral point in the middle, corresponding to no impact.

The participants were instructed to place the argument icon as much to the right (left) as they judged the information given about the uncertainty of evidence to increase (decrease) the plausibility of the hypothesis. Once they expressed their judgments, a novel double sampling was said to have been performed, and the participants were asked to evaluate another set of inductive arguments, and so on for all seven sets. Table 1 displays a full description of the seven levels of uncertain evidence and the four hypotheses that appeared in the inductive arguments employed.

From the findings of pilot studies, we expected the four chosen hypotheses to elicit quantitatively different judgments on both the positive and negative sides of the impact scale. Also, the chosen hypotheses were expected to span from low to moderate priors (i.e., perceived frequencies of the predicates in the overall population group). We preferred not to employ hypotheses with very high priors because they would have hindered significantly unbalanced distributions in the male versus female subgroups.

On the whole, we collected 28 confirmation judgments for each participant (seven sets × four hypotheses). The concurrent evaluation of four arguments fostered relevant comparisons and appropriate use of the quantitative scale.

Probability task. After the confirmation task, participants were asked to consider again a group of 1,000 students—500 males and 500 females—and answer questions like the following, for each hypothesis. *How many **male** students out of 500 own a €10,000 motorbike? How many **male** students out of 500 **do not** own a €10,000 motorbike? How many **female** students out of 500 own a €10,000*

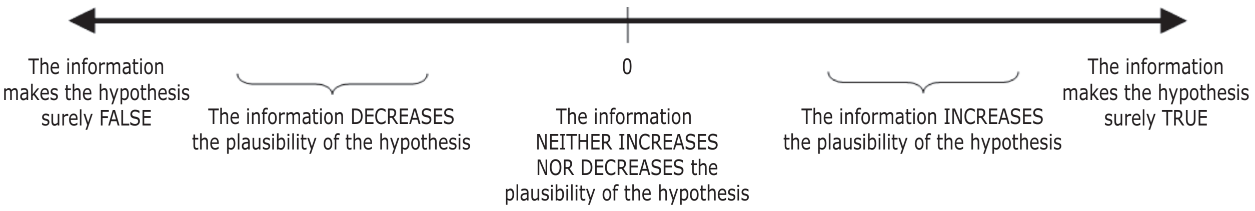

0
The information makes the hypothesis surely FALSE
The information DECREASES the plausibility of the hypothesis
The information NEITHER INCREASES NOR DECREASES the plausibility of the hypothesis
The information INCREASES the plausibility of the hypothesis
The information makes the hypothesis surely TRUE

Figure 1. The impact scale used for confirmation judgments.

motorbike? How many ***female*** *students out of 500* **do not** *own a €10,000 motorbike?*

Complementary estimates were asked for in order to foster accuracy. The participants could begin from whichever estimate they preferred. The software required each pair of complementary estimates to sum up to 500.

Results and Discussion

Following the notation used in the introductory section, h represents a hypothesis, corresponding to one of those shown in Table 1; subscripts t and $t+1$ indicate, respectively, the initial and subsequent degrees of belief concerning the statement "The drawn student is male [female]," which is denoted by e [not-e].

In order to test relevant theoretical predictions against collected judgments, quantities $P_t(h)$ and $P_{t+1}(h)$ were calculated for each of the 28 arguments presented and for each participant by means of the theorem of total probability and Jeffrey's conditionalization rule, respectively—that is,

$$P_t(h) = P_t(h \mid e)P_t(e) + P_t(h \mid \text{not-}e)P_t(\text{not-}e) \quad \text{[theorem of total probabilities]} \tag{5A}$$

and

$$P_{t+1}(h) = P_t(h \mid e)P_{t+1}(e) + P_t(h \mid \text{not-}e)P_{t+1}(\text{not-}e) \quad \text{[Jeffrey conditionalization].} \tag{5B}$$

Notice that all of the values in Equations 5A and 5B were available. The experimental procedure fixed $P_t(e)$ and $P_{t+1}(e)$. In particular, the initial probability that the drawn student was male, $P_t(e)$, was set at .5, because the participants were informed at the beginning that the overall group of 1,000 students was formed by equal numbers of males and females. $P_{t+1}(e)$ was then provided by the additional information contained in each argument as amounting to one of the seven levels of evidence uncertainty reported in Table 1. Values $P_t(h \mid e)$ and $P_t(h \mid \text{not-}e)$, on the other hand, emerged from the estimates that each participant expressed while performing the probability task and were simply obtained through division by 500 of the estimate given in response to the question about the numbers of male and female students (out of 500) satisfying hypothesis h (e.g., owning a €10,000 motorbike).

The results from the probability task, partially serving as a manipulation check for the selection of the predicates employed, are given below. Table 2 displays mean probability values obtained from participants' responses. Associations between gender and predicates, as well as overall perceived frequencies of the latter, were broadly in line with prior expectations guiding the construction of experimental materials.

Table 2
Average Values From the Probability Task in Experiment 1 (*n* = 33)

h	$P_t(h \mid \text{e})$	$P_t(h \mid \text{not-e})$	$P_t(h)$
usually applies eye makeup	.04	.85	.44
owns a €10,000 necklace	.04	.15	.09
owns a €10,000 motorbike	.19	.04	.12
usually has a shaved beard	.88	.01	.44

Note—Statement "The drawn student is male [female]" is denoted by e [not-e].

Let us now turn to the results from the confirmation task. Mean judgments across participants are displayed in Table 3. As read across the rows, these data conform to a basic feature of confirmation by uncertain evidence—that is, increasing (decreasing) amounts of confirmation for a fixed h as the new probability of a piece of confirming (disconfirming) evidence increases (for a formal analysis, see Crupi, Festa, & Mastropasqua, 2008). As read down the columns, moreover, the data also show that, for each fixed level of evidence uncertainty—excluding the special case $P_t(e) = P_{t+1}(e) = .5$ (see the discussion below)—confirmation judgments are spread over the impact scale in both the positive and negative directions. The latter point is of particular interest because it clearly shows that our results cannot be accounted for by merely considering the direction and extent of the departure of $P_{t+1}(e)$ from $P_t(e) = .5$; otherwise, absolute values would be identical across all entries within each column in Table 3. Confirmation measures L and Z, on the contrary, do capture the observed pattern. In fact, both measures rank arguments along the columns precisely as displayed in Table 3, once average probabilities from Table 2 are plugged into Equations 5A and 5B and confirmation values are subsequently computed from Equations 4A and 4B. On the whole, these remarks are an indication that observed quantitative judgments were sensitive to three basic and distinctive aspects that are all integrated in the Bayesian account of confirmation by uncertain evidence—namely, (1) the specific hypothesis h concerned (as identified, most notably, by its prior), (2) the positive versus negative connection between e and h, and (3) the degree of uncertainty of e.

As for the special case in which $P_{t+1}(e) = .5$, a distinctive implication of Bayesian confirmation theory, as extended to uncertain evidence, is that a null degree of confirmation applies for all four arguments. Indeed, because $P_{t+1}(e) = .5 = P_t(e)$, nothing new is actually learned concerning the evidence from t to $t+1$. As a consequence, one can easily verify that $P_{t+1}(h) = P_t(h)$, whatever their specific values may be (see Equations 5A and 5B). In this

Table 3
Average Values From the Confirmation Task in Experiment 1 (*n* = 33)

h	$P_{t+1}(e) = 1$	$P_{t+1}(e) = .8$	$P_{t+1}(e) = .7$	$P_{t+1}(e) = .5$	$P_{t+1}(e) = .3$	$P_{t+1}(e) = .2$	$P_{t+1}(e) = 0$
usually applies eye makeup	−.84	−.61	−.49	.01	.50	.56	.79
owns a €10,000 necklace	−.55	−.42	−.34	.00	.31	.36	.42
owns a €10,000 motorbike	.46	.35	.25	.00	−.37	−.38	−.55
usually has a shaved beard	.90	.66	.51	.02	−.56	−.65	−.97

Note—Statement "The drawn student is male" is denoted by e.

experiment, a "0" confirmation judgment was reported in 94% of the relevant cases (124 out of $4 \times 33 = 132$), with an overall mean departure from 0 (i.e., algebraic difference in absolute value) amounting to .01.

More generally, if confirmation by uncertain evidence is appropriately assessed, individual confirmation judgments expressed by participants—hereafter denoted by $\text{Judged}_{t,t+1}(h)$—should match the basic condition displayed for $c_{t,t+1}(h)$ in the introductory section (see Equation 3). It should, therefore, be the case that

$$\text{Judged}_{t,\,t+1}(h) \begin{cases} > 0 & \text{if} \quad P_{t+1}(h) > P_t(h). \\ = 0 & \text{if} \quad P_{t+1}(h) = P_t(h). \\ < 0 & \text{if} \quad P_{t+1}(h) < P_t(h). \end{cases} \qquad (6)$$

We thus checked whether the basic normative constraint in Equation 6 was indeed satisfied at the individual level. Overall, only 17 among $(28 \times 33) = 924$ $\text{Judged}_{t,t+1}(h)$—that is, 1.8%—violated Equation 6. We also carried out the same analysis after splitting the confirmation judgments into two subsets consisting of limiting cases of evidence uncertainty versus cases of strict evidence uncertainty, respectively. The former subset includes $(8 \times 33) = 264$ judgments with $P_{t+1}(e)$ amounting to either 100% or 0% (indicating that either *e* or not-*e* was, in fact, *certain* evidence at $t+1$). The latter subset includes all of the $(20 \times 33) = 660$ other judgments, with $P_{t+1}(e)$ amounting to intermediate values between 80% and 20% (see Table 1). In both subsets, the proportion of violations of Equation 6 was negligible (0.4% in limiting cases and 2.4% under strict uncertainty). Thus, intuitive confirmation judgments elicited in Experiment 1 largely reflect the theoretical distinction of positive, null, and negative impact, even when evidence is strictly uncertain.

A further kind of analysis was aimed at measuring the degree of association between participants' confirmation judgments and the corresponding quantitative degrees of confirmation, as predicted by measures *L* and *Z*. In line with the notation introduced earlier, let us denote any confirmation judgment as predicted by *L* and *Z* as $L_{t,t+1}(h)$ and $Z_{t,t+1}(h)$, respectively. For each participant, we first computed the 28 $L_{t,t+1}(h)$ and $Z_{t,t+1}(h)$ values by directly substituting $P_t(h)$ and $P_{t+1}(h)$ into the relevant expressions (see Equations 4A and 4B). For 2 participants, some $L_{t,t+1}(h)$ turned out to be undefined because $P_t(h)$ and $P_{t+1}(h)$ were zero for some hypotheses *h* (division by zero) and were thus excluded from the present analysis. For each of the remaining 31 participants, Pearson[4] correlations were computed between the 28 $\text{Judged}_{t,t+1}(h)$ and the corresponding 28 $L_{t,t+1}(h)$, $Z_{t,t+1}(h)$, and posterior probabilities arising from Jeffrey conditionalization. Average correlations across participants are shown in Table 4.

If participants' judgments did not appropriately reflect the distinction between confirmation and posterior, then the average correlation from posterior probability would have been close to 1. It can be seen that, on the contrary, posterior probability produced the lowest average correlation. Indeed, paired *t* tests revealed that average correlations yielded by *L* and *Z* were both reliably greater than that computed by posterior probability ($p < .01$). Thus, participants in our experiment were apparently able to assess confirmation as distinct from posterior probability. Furthermore, the high average correlations with both *L* and *Z* indicate that participants' confirmation judgments were normatively sound—that is, close to those implied by credible theoretical models, with a small but significant higher predictive accuracy of *L* than of *Z* ($p < .01$, by paired *t* test).

We also carried out the same quantitative analyses on a more detailed level by identifying three subsets of judgments. The first subset amounted to the limiting cases of evidence uncertainty as defined above—that is, with $P_{t+1}(e)$ equal to either 100% or 0%. The second and third subsets were two classes of cases of strict evidence uncertainty: $P_{t+1}(e)$ equal to either 80% or 20% and $P_{t+1}(e)$ equal to either 70% or 30%. The results closely matched those from the general analysis reported above. Average correlations with each of the measures *L* and *Z* were statistically indistinguishable across all three subsets. Within each subset, both *L* and *Z* were consistently superior predictors, as compared with posterior probability ($p < .01$, by paired *t* tests), with *L* being consistently more accurate than *Z* ($p < .05$, by paired *t* test).

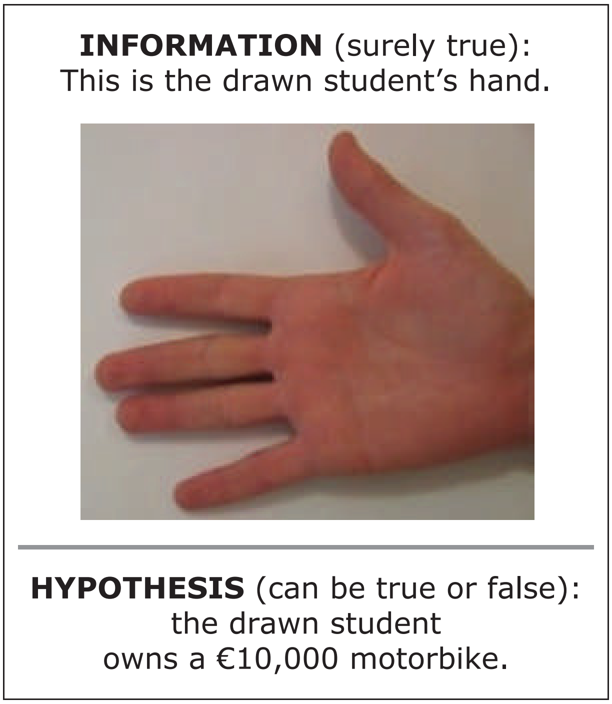
EXPERIMENT 2

Experiment 1 employed inductive arguments in which the probability of evidence was explicitly provided (e.g., "The drawn student is male with 80% probability, female with 20% probability"). Results show that the participants' judgments largely conformed to plausible normative models. However, in most inductive arguments from real life, people have to deal with uncertain evidence while not being given any numerical measure of belief by some

Table 4
Average Pearson Correlations Between Judged Confirmation and Confirmation Predicted by *L* and *Z*, and Between Judged Confirmation and Posterior Probability Computed by Jeffrey Conditionalization

	Predicted Confirmation (*L*)	Predicted Confirmation (*Z*)	Posterior Probability (Jeffrey Conditionalization)
Judged confirmation	.913*	.903*	.662

Note—Each value is the average of 31 Pearson correlations (1 per participant) involving 28 observations. *Reliably greater than the average correlation for posterior probability at $p < .01$ by paired *t* test.



INFORMATION (surely true):
This is the drawn student's hand.

HYPOTHESIS (can be true or false):
the drawn student
owns a €10,000 motorbike.

Figure 2. Sample stimulus from Experiment 2.

external source. As a test of generality, in Experiment 2 the uncertainty of evidence was manipulated indirectly by means of ambiguous pictures.

Method

Thirty-four students (15 female, mean age = 26 years) from the University of Trento participated in Experiment 2 in exchange for course credit. None had participated in Experiment 1. As in Experiment 1, the participants performed a confirmation task followed by a probability task presented via a custom Java application.

Confirmation task. The confirmation task was basically the same as that in Experiment 1, but it differed in the way in which evidential uncertainty was manipulated. In Experiment 2, participants were presented the following scenario:

> Consider a group of 1,000 students, **500 males** and **500 females**, randomly selected at the University of Trento. In what follows, we will repeatedly draw at random one among the 1,000 students, and we will show you a picture of her/his hand. Draws will be independent at each trial (so, in principle, the same student might be selected more than once).

As can be seen, no double sampling procedure was involved in this scenario. The student was said to have been directly drawn from the larger sample of 1,000. The uncertainty of evidence concerning the student's gender was implicitly manipulated via the picture of her/his hand. We selected pictures displaying more or less relevant cues to gender, according to the findings of a pilot study, thus determining more or less extreme departures of the probability of being male/female from the initial base-rate level of .5. Pictures were also selected as *not* displaying cues that could possibly count as relevant *direct* evidence for any of the hypotheses (i.e., independent of gender). For example, the hands chosen and pictured exhibited no rings, tattoos, or nail enamel. (See the Results and Discussion section for further discussion of this point.)

At each trial, an enlarged picture of the hand appeared on the screen for 10 sec, and the participants were prompted to look at it very carefully and in detail. The picture then automatically reduced in size (but could still be widened just by clicking on it), and the participants were asked to answer the following questions: (1) *In light of the picture, do you think the drawn student is male or female?* [Participants had to choose one option: *male* vs. *female.*] (2) *What is the probability that your previous answer is correct?* [Participants had to place the cursor on a sliding bar ranging from 50% to 100%.]

The responses to the questions above provided an estimate of participants' perceived degrees of uncertainty about the evidence concerning gender. Afterward, a set of four inductive arguments was presented, while a reminder in the top right corner of the screen reported the degree of uncertainty previously assigned to the evidence. As in Experiment 1, the participants had to estimate inductive confirmation. The hypotheses, as well as the scale employed and the rest of the procedure, were the same as in Experiment 1. An example of inductive argument as displayed in Experiment 2 appears in Figure 2.

Probability task. The probability task was exactly the same as in Experiment 1.

Results and Discussion

As Table 5 shows, the results from the probability task in Experiment 2 largely reproduce those in Experiment 1.

With regard to the confirmation task, notice that, unlike in Experiment 1, no analysis of mean values across participants is appropriate here. This is because different levels of evidential uncertainty were not uniformly provided in the procedure, but rather were individually judged on the basis of the interpretation of pictures.

Only 7 participants in Experiment 2 assigned $P_{t+1}(e) = .5$ as reflecting their assessment of some of the pictures of hands they were being shown. As a consequence, only 48 arguments required a null degree of confirmation as implied by $P_t(e) = P_{t+1}(e)$. Thirty-eight of these arguments (79%) actually elicited "0" responses, with a mean departure from 0 (i.e., algebraic difference in absolute value) amounting to .09. It should be noted, moreover, that 1 participant accounted for half of the 10 normatively inconsistent judgments. If this single participant is left out of this analysis, the proportion of appropriate "0" responses rises to 87.5%, with a mean departure from 0 of just .01.

On the whole, $(28 \times 34) = 952$ Judged$_{t,t+1}(h)$ were collected in Experiment 2. Sixty-three of them (6.6%) violated Equation 6 above (i.e., the basic normative distinction of positive, null, and negative impact). Based on the participants' own interpretations of the pictures displayed, limiting cases of evidence uncertainty (i.e., with Judged $P_{t+1}(e)$ amounting to either 100% or 0%) were a small minority, namely 56 (5.9%) judgments out of 952. The proportions

Table 5
Average Values From the Probability Task in Experiment 2 (n = 34)

h	$P_t(h \mid e)$	$P_t(h \mid \text{not-}e)$	$P_t(h)$
usually applies eye makeup	.01	.84	.43
owns a €10,000 necklace	.03	.10	.07
owns a €10,000 motorbike	.16	.07	.11
usually has a shaved beard	.88	.003	.44

Note—Statement "The drawn student is male [female]" is denoted by *e* [not-*e*].

Table 6
Average Pearson Correlations Between Judged Confirmation and Confirmation Predicted by *L* and *Z*, and Between Judged Confirmation and Posterior Probability Computed by Jeffrey Conditionalization

	Predicted Confirmation (L)	Predicted Confirmation (Z)	Posterior Probability (Jeffrey Conditionalization)
Judged confirmation	.902*	.893*	.605

Note—Each value is the average of 34 Pearson correlations (1 per participant) involving 28 observations. *Reliably greater than the average correlation for posterior probability at $p < .01$ by paired *t* test.

of violations of Equation 6 in the latter set and among all remaining judgments involving strict evidence uncertainty were 5.4% and 6.7%, respectively. Overall, although still minor, departures from Equation 6 were somewhat more common than in Experiment 1 (*z* test for proportion, $p < .01$), presumably reflecting an increased difficulty of the task. The pattern arising from quantitative analyses was nevertheless very similar to that in Experiment 1.

Average Pearson correlations from L, Z, and posterior probability are shown in Table 6. Once again, both L and Z yielded very high average correlations, significantly greater than that with posterior probability ($p < .01$, by paired *t* tests). Much as in Experiment 1, moreover, the higher average correlation of measure L than of Z also reached statistical significance ($p < .05$). Also, as in Experiment 1, the results are not inflated by limiting cases of evidence uncertainty, because all significance tests remain unaffected under strict evidence uncertainty—that is, by the removal of the 5 participants who sometimes provided extreme values for $P_{t+1}(e)$.

As a final point, let us come back to the indirect manipulation of the uncertainty of evidence concerning gender by means of ambiguous pictures. As anticipated in the Method section, for this manipulation to reliably serve the aims of the experiment, the pictures shown should not have provided any significant *direct* hint with regard to the hypotheses at issue (i.e., independent of gender). Any such direct impact—if present, despite our deliberate precautions—would have run strongly counter to the descriptive accuracy of confirmation models as applied in the experiment. Consider the artificial but illustrative case of a seemingly female hand that appears to be smeared with motor oil. Participants would have quite reasonably seen the corresponding argument as a *confirmation* of the motorbike hypothesis. And clearly, predictions derived from confirmation measures in our experiment would have blatantly failed to capture judgments of this kind, since they only had gender as an (uncertain) evidence input. As a consequence, the high correlations obtained from L and Z—perfectly in line with those in Experiment 1—also provide independent empirical support for the validity of the major methodological novelty of Experiment 2.[5]

GENERAL DISCUSSION

Ever since the work of chief Bayesian theorists such as Keynes (1921), Carnap (1950/1962), and Good (1950), a basic component of inductive reasoning has been identified in the notion of evidence prompting a change in belief—namely, confirmation—as distinct from final belief per se. In the philosophy of science and in epistemology, the debate on the issue has been lasting (see, e.g., Earman, 1992; Fitelson, 1999). In the psychological literature, on the other hand, Bayesian confirmation has occurred sparsely and indirectly, often by different names. It has been invoked, for instance, in discussions concerning the reality of the "conjunction fallacy" (see Crupi, Fitelson, & Tentori, 2008; Sides et al., 2002) and related phenomena (see Lagnado & Shanks, 2002), as well as in inquiries into various aspects of the perception of chance (e.g., Tenenbaum & Griffiths, 2001). A specific principle of confirmation theory has been experimentally studied by Lo, Sides, Rozelle, and Osherson (2002) and found to be largely adhered to in children's reasoning. Bayesian confirmation also yields formal and conceptual connections with models of the value of information (Nelson, 2005) involved in a number of established research areas in psychology, such as Wason's selection task (see Fitelson, 2010; Klayman & Ha, 1987; McKenzie & Mikkelsen, 2000; Nickerson, 1996; Oaksford & Chater, 1994, 2003).

The experiments reported here extend recent studies explicitly devoted to the psychology of confirmation (Crupi et al., 2007; Tentori et al., 2007; Tentori, Crupi, & Osherson, 2010). Tentori et al. (2007), in particular, employed an urn setting with sequential draws, where relevant evidence (the color of drawn balls) was certain (indeed, was established by the participants themselves by direct observation). In Tentori et al. (2007), intuitive judgments of confirmation reflected to a remarkable extent the formal notion as represented by normatively appealing accounts, such as measures L and Z (see also Crupi et al., 2007). The present experiments replicate this basic finding in a different setting and generalize it to the assessment of confirmation by uncertain evidence. As implied by the results of both Experiments 1 and 2, confirmation measures L and Z accurately capture inductive reasoning from naive reasoners even when uncertain evidence is at issue (as is often the case in real settings), and both models outperform posterior probability as computed from Jeffrey conditionalization, here employed for comparison as a competing, although theoretically spurious, predictor of confirmation judgments.

In order to better appreciate the results reported here, it is useful to consider the following points about the procedures adopted, showing, in our view, how the present results significantly extend current knowledge of human inductive reasoning. First, participants were not faced with problems involving artificially devised predicates (such as the color of balls or the composition of urns, as in Tentori et al., 2007) or blank (i.e., semantically opaque) properties, as is common in other experimental paradigms for the study of inductive reasoning (e.g., Osherson et al., 1990). Rather, real-world and transparent hypotheses were employed.

Second, convergent results were obtained with two different ways of manipulating evidence uncertainty—that is, directly providing probabilistic information (Experiment 1) versus relying on the interpretation of ambiguous pictures conveying uncertainty (Experiment 2).

Finally, and more generally, the relative difficulty of the task, which makes participants' performance noticeable, should be mentioned. A sound confirmation judgment always reflects the quantitative relationship between two distinct variables (i.e., prior and posterior probability). By their responses, our participants proved to be able to integrate the degree of evidence uncertainty into this sophisticated assessment.

Notably, none of our conclusions above presupposes the descriptive validity of Jeffrey conditionalization per se—an interesting issue deserving investigation on its own. By way of analogy, one can remark that the perception of an environment getting hotter (or cooler) is distinct from a numerical estimate of the final temperature. The former could well be appropriate even if the latter is biased for whatever reason (say, an anchoring effect). Indeed, Tentori et al. (2007) reported that, in their urn setting involving ascertained evidence, elicited confirmation judgments were in line with normatively appealing Bayesian confirmation measure L, despite subjective assessments of posteriors being prone to well-known biases (i.e., *conservatorism*; see Edwards, 1968; Slovic & Lichtenstein, 1971).

Beyond a generally high correlation with observed judgments, our experiments also document a slight but significant advantage of measure L over measure Z in terms of descriptive accuracy. Interestingly, Crupi et al. (2007) had reported a similar but reversed pattern: L and Z turned out to be very good predictors, with a slight but significant advantage for the latter. Measures L and Z thus appear to be close competitors in capturing confirmation assessment in human reasoning. More definite conclusions about their respective merits also remain an aim for further research.

To conclude, we shall notice that a growing trend of claims depicts various aspects of human inductive reasoning involving certain evidence as appropriately captured by sophisticated models arising from the Bayesian approach and involving normatively sound principles (see, e.g., Crupi, Tentori, & Lombardi, 2009; Griffiths & Tenenbaum, 2006; Kemp & Tenenbaum, 2009; Oaksford & Chater, 2007; but for a critical view, see Sloman & Fernbach, 2008). As far as inductive confirmation by uncertain evidence is concerned, however, available normative models had not yet been examined through empirical testing. The experiments reported here open up this line of investigation, providing evidence that those models prove psychologically tenable.

AUTHOR NOTE

The research was supported by a grant from the SMC/Fondazione Cassa di Risparmio di Trento e Rovereto for the CIMeC (University of Trento) Research Project on Inductive Reasoning, by the Marie Curie Grant INDUCT (221350), and by a grant from the Spanish Department of Science and Innovation (FFI2008-01169/FISO). Correspondence concerning this article should be addressed to K. Tentori, CIMeC, DiSCoF, University of Trento, Corso Bettini 31, Rovereto, Italy (e-mail: katya.tentori@unitn.it).

REFERENCES

Baron, J. (2008). *Thinking and deciding* (4th ed.). New York: Cambridge University Press.

Blok, S. V., Medin, D. L., & Osherson, D. N. (2007). Induction as conditional probability judgment. *Memory & Cognition*, **35**, 1353-1364.

Blok, S. V., Osherson, D., & Medin, D. L. (2007). From similarity to chance. In A. Feeney & E. Heit (Eds.), *Inductive reasoning* (pp. 137-166). New York: Cambridge University Press.

Carnap, R. (1962). *Logical foundations of probability*. Chicago: University of Chicago Press. (Original work published 1950)

Chan, H., & Darwiche, A. (2005). On the revision of probabilistic beliefs using uncertain evidence. *Artificial Intelligence*, **163**, 67-90.

Crupi, V., Festa, R., & Buttasi, C. (2010). Towards a grammar of Bayesian confirmation. In M. Suárez, M. Dorato, & M. Rédei (Eds.), *EPSA epistemology and methodology of science: Vol. 1* (pp. 73-93). Berlin: Springer.

Crupi, V., Festa, R., & Mastropasqua, T. (2008). Bayesian confirmation by uncertain evidence: A reply to Huber (2005). *British Journal for the Philosophy of Science*, **59**, 201-211.

Crupi, V., Fitelson, B., & Tentori, K. (2008). Probability, confirmation, and the conjunction fallacy. *Thinking & Reasoning*, **14**, 182-199.

Crupi, V., Tentori, K., & Gonzalez, M. (2007). On Bayesian measures of evidential support: Theoretical and empirical issues. *Philosophy of Science*, **74**, 229-252.

Crupi, V., Tentori, K., & Lombardi, L. (2009). Pseudodiagnosticity revisited. *Psychological Review*, **116**, 971-985.

Dawes, R. M. (2001). *Everyday irrationality: How pseudo-scientists, lunatics, and the rest of us systematically fail to think rationally*. Boulder, CO: Westview.

Earman, J. (1992). *Bayes or bust? A critical examination of Bayesian confirmation theory*. 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: Wiley.

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

Festa, R. (1999). Bayesian confirmation. In M. C. Galavotti & A. Pagnini (Eds.), *Experience, reality, and scientific explanation: Essays in honor of Merrilee and Wesley Salmon* (pp. 55-87). Dordrecht, The Netherlands: Kluwer.

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

Fitelson, B. (2001). A Bayesian account of independent evidence with applications. *Philosophy of Science*, **68**, S123-S140.

Fitelson, B. (2006). Inductive logic. In S. Sarkar & J. Pfeifer (Eds.), *Philosophy of science: An encyclopedia*. New York: Routledge.

Fitelson, B. (2010). *The Wason task(s) and the paradox of confirmation*. Manuscript in preparation.

Good, I. J. (1950). *Probability and the weighing of evidence*. London: Griffin.

Griffiths, T. L., & Tenenbaum, J. B. (2006). Optimal predictions in everyday cognition. *Psychological Science*, **17**, 767-773.

Hartmann, S. (2008). Modeling in philosophy of science. In M. Frauchiger & W. K. Essler (Eds.), *Representation, evidence, and justification: Themes from Suppes* (pp. 95-121). Frankfurt: Ontos.

Hastie, R., & Dawes, R. M. (2001). *Rational choice in an uncertain world: The psychology of judgment and decision making*. Thousand Oaks, CA: Sage.

Heit, E. (1998). A Bayesian analysis of some forms of inductive reasoning. In M. Oaksford & N. Chater (Eds.), *Rational models of cognition* (pp. 248-274). Oxford: Oxford University Press.

Howson, C., & Urbach, P. (2006). *Scientific reasoning: The Bayesian approach* (3rd ed.). La Salle, IL: Open Court.

Jeffrey, R. (1965). *The logic of decision* (2nd ed.). Chicago: University of Chicago Press.

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

Jeffrey, R. (2004). *Subjective probability: The real thing*. Cambridge: Cambridge University Press.

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

Kemp, C., & Tenenbaum, J. B. (2009). Structured statistical models of inductive reasoning. *Psychological Review*, **116**, 20-58.

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

Klayman, J. M., & Ha, Y.-W. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review*, **94**, 211-228.

Lagnado, D. A., & Shanks, D. R. (2002). Probability judgment in hierarchical learning: A conflict between predictiveness and coherence. *Cognition*, **83**, 81-112.

Lange, M. (2000). Is Jeffrey conditionalization defective by virtue of being non-commutative? Remarks on the sameness of sensory experiences. *Synthese*, **123**, 393-403.

Lo, Y., Sides, A., Rozelle, J., & Osherson, D. (2002). Evidential diversity and premise probability in young children's inductive judgment. *Cognitive Science*, **26**, 181-206.

McKenzie, C. R. M., & Mikkelsen, L. A. (2000). The psychological side of Hempel's paradox of confirmation. *Psychonomic Bulletin & Review*, **7**, 360-366.

Medin, D. L., Coley, J. D., Storms, G., & Hayes, B. K. (2003). A relevance theory of induction. *Psychonomic Bulletin & Review*, **10**, 517-532.

Mura, A. (2006). Deductive probability, physical probability and partial entailment. In M. Alai & G. Tarozzi (Eds.), *Karl Popper: Philosopher of science* (pp. 181-202). Soveria Mannelli, Italy: Rubbettino.

Mura, A. (2008). Can logical probability be viewed as a measure of degrees of partial entailment? *Logic & Philosophy of Science*, **6**, 25-33.

Nelson, J. D. (2005). Finding useful questions: On Bayesian diagnosticity, probability, impact, and information gain. *Psychological Review*, **112**, 979-999.

Nickerson, R. S. (1996). Hempel's paradox and Wason's selection task: Logical and psychological puzzles of confirmation. *Thinking & Reasoning*, **2**, 1-31.

Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review*, **101**, 608-631.

Oaksford, M., & Chater, N. (2003). Optimal data selection: Revision, review, and reevaluation. *Psychonomic Bulletin & Review*, **10**, 289-318.

Oaksford, M., & Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford: Oxford University Press.

Osherson, D. N., Smith, E. E., Wilkie, O., López, A., & Shafir, E. (1990). Category-based induction. *Psychological Review*, **97**, 185-200.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Mateo, CA: Morgan Kaufmann.

Rescher, N. (1958). A theory of evidence. *Philosophy of Science*, **25**, 83-94.

Rips, L. J. (2001). Two kinds of reasoning. *Psychological Science*, **12**, 129-134.

Shortliffe, E. H., & Buchanan, B. G. (1975). A model of inexact reasoning in medicine. *Mathematical Biosciences*, **23**, 351-379.

Sides, A., Osherson, D., Bonini, N., & Viale, R. (2002). On the reality of the conjunction fallacy. *Memory & Cognition*, **30**, 191-198.

Sloman, S. A., & Fernbach, P. M. (2008). The value of rational analysis: An assessment of causal reasoning and learning. In N. Chater & M. Oaksford (Eds.), *The probabilistic mind: Prospects for rational models of cognition* (pp. 485-500). New York: Oxford University Press.

Sloman, S. A., & Lagnado, D. A. (2005). The problem of induction. In K. J. Holyoak & R. G. Morrison (Eds.), *Cambridge handbook of thinking and reasoning* (pp. 95-116). Cambridge: Cambridge University Press.

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

Tenenbaum, J. B., & Griffiths, T. L. (2001). The rational basis of representativeness. In J. D. Moore & K. Stenning (Eds.), *Proceedings of the 23rd Annual Conference of the Cognitive Science Society* (pp. 1036-1041). Mahwah, NJ: Erlbaum.

Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences*, **10**, 309-318.

Tentori, K., Crupi, V., Bonini, N., & Osherson, D. (2007). Comparison of confirmation measures. *Cognition*, **103**, 107-119.

Tentori, K., Crupi, V., & Osherson, D. (2010). Second-order probability affects hypothesis confirmation. *Psychonomic Bulletin & Review*, **17**, 129-134.

Tomic, W., & Kingma, J. (1998). Accelerating intelligence development through inductive reasoning training. *Advances in Cognition & Educational Practice*, **5**, 291-305.

Wagner, C. G. (2002). Probability kinematics and commutativity. *Philosophy of Science*, **69**, 266-278.

NOTES

1. Indeed, under strict Bayesian conditionalization, $L_{t,t+1}(h) = \tanh \{\frac{1}{2} \ln[P(e|h)/P(e|\text{not-}h)]\}$.

2. Confirmation judgments represented the ultimate variable of interest of this study, with probability estimates providing relevant data for the empirical testing of Bayesian confirmation measures. As a consequence, task order was fixed for all participants in an effort to preserve the intuitive and naive character of confirmation assessments from any risk of carryover effects.

3. All materials are translated from the Italian originals.

4. We assume $Judged_{t,t+1}(h)$ to lie on an interval scale, because participants expressed their confirmation judgments through a continuous scale.

5. We owe acknowledgment to an anonymous reviewer for prompting us to discuss this point and for providing the useful motor oil illustration.

(Manuscript received September 30, 2009;
revision accepted for publication March 4, 2010.)