Second-order probability affects hypothesis confirmation

KATYA TENTORI
University of Trento, Trento, Italy
and University College London, London, England

VINCENZO CRUPI
University of Florence, Florence, Italy

AND

DANIEL OSHERSON
Princeton University, Princeton, New Jersey

Bayesian confirmation measures give numerical expression to the impact of evidence E on a hypothesis H. All measures proposed to date are *formal*—that is, functions of the probabilities $Pr(E \wedge H)$, $Pr(E \wedge \neg H)$, $Pr(\neg E \wedge H)$, $Pr(\neg E \wedge \neg H)$, and nothing more. Experiments reported in Tentori, Crupi, and Osherson (2007) suggest that human confirmation judgment is not formal, but this earlier work leaves open the possibility that formality holds relative to a given semantic domain. The present study discredits even this weaker version of formality by demonstrating the role in confirmation judgments of a probability distribution defined over the possible values of $Pr(E \wedge H)$, $Pr(E \wedge \neg H)$, $Pr(\neg E \wedge H)$, and $Pr(\neg E \wedge \neg H)$—that is, a *second-order probability*. Specifically, when for each of the latter quantities a pointwise value is fixed with a maximal second-order probability, evidence impact is rated in accordance with formal and normatively credible confirmation measures; otherwise evidence impact is systematically judged as more moderate.

Epistemologists often frame inductive reasoning within theories of hypothesis confirmation. Bayesian accounts of confirmation qualify piece of evidence E as having a *positive*/*negative* impact on hypothesis H to the extent that verifying E *increases*/*decreases* the credibility of H. Thus, in the usual circumstances, a market plunge in Europe tomorrow increases the probability of a similar plunge in New York, whereas playing rugby decreases the credibility of being a violinist. Let us observe that change in credibility (or confirmation) is not the same as conditional probability. For, otherwise, an Italian victory in the next World Cup would maximally confirm the hypothesis that it will rain in Seattle sometime in 2020, since the conditional probability of the latter given the former is near unity. (As they seem to be completely independent, the two statements have no confirmatory impact on each other.)

Bayesian confirmation measures acknowledge the relation between impact and posterior probability by satisfying the "classificatory" condition (due to Carnap,

1950/1962, pp. 21–22) shown in (1), below. Several measures of confirmation satisfying Carnap's condition have been proposed (for discussion and comparison see Crupi, Tentori, & Gonzalez, 2007; Eells & Fitelson, 2002).[1] Two among them are given in (2) (Good, 1983; Kemeny & Oppenheim, 1952) and (3) (Crupi et al., 2007).

$$l(E, H) = \frac{Pr(E|H) - Pr(E|\neg H)}{Pr(E|H) + Pr(E|\neg H)}, \tag{2}$$

$$z(E, H) = \begin{cases} \dfrac{Pr(H|E) - Pr(H)}{Pr(\neg H)} & \text{if } Pr(H|E) \geq Pr(H), \\ \dfrac{Pr(H|E) - Pr(H)}{Pr(H)} & \text{otherwise.} \end{cases} \tag{3}$$

All Bayesian confirmation measures thus far proposed share a property identified in Tentori, Crupi, and Osherson (2007) called *formality*. A confirmation measure (with arguments E and H) is formal if, and only if, it is determined by the probability distribution over the algebra gen-

$$\text{CONF(E, H)} \begin{cases} > 0 & \text{iff } Pr(H|E) > Pr(H) \quad \text{(i.e., iff E has a } positive\ impact \text{ on H)} \\ = 0 & \text{iff } Pr(H|E) = Pr(H) \quad \text{(i.e., iff E has } no\ impact \text{ on H)} \\ < 0 & \text{iff } Pr(H|E) < Pr(H) \quad \text{(i.e., iff E has a } negative\ impact \text{ on H)} \end{cases} \tag{1}$$

K. Tentori, katya.tentori@unitn.it

erated by E and H. In other words, a formal confirmation measure depends on just the four probabilities $Pr(E \wedge H)$, $Pr(E \wedge \neg H)$, $Pr(\neg E \wedge H)$, and $Pr(\neg E \wedge \neg H)$. To illustrate, l is formal because it is a function of $Pr(E|H)$ and $Pr(E|\neg H)$, which can both be expressed in terms of the four probabilities above; specifically, $Pr(E|H) = Pr(E \wedge H) / [Pr(E \wedge H) + Pr(\neg E \wedge H)]$, and similarly for $Pr(E|\neg H)$. Likewise, z and all other Bayesian confirmation measures yet proposed are formal.

We can illustrate the implications of formality as follows. Consider the extraction of an individual from a random sample of 100 men and 100 women. Imagine you are informed that the drawn individual likes cigars. How much does this influence your opinion about the hypothesis that the individual selected is male? Imagine now that your probability estimates concerning the proportions of men versus women who like cigars are elicited and transferred to an urn with red versus blue and striped versus spotted balls. Specifically, the number of red/blue striped balls in the urn corresponds to your estimate of the number of men/women in the sample who like cigars, whereas the number of red/blue spotted balls in the urn correspond to your estimate of the number of men/women in the sample who do not like cigars. A ball is drawn from this urn. You are informed that the drawn ball is striped. How much does this influence your opinion about the drawn ball being red? Will you express the same confirmation judgment as in the isomorphic problem involving men and cigars? Because the relevant probabilities are identical across the two problems, formality requires that your respective judgments of confirmation coincide.

But, typically, they do not. Using stimuli of the foregoing kind, Tentori, Crupi, and Osherson (2007) showed that the function governing people's judgments of confirmation are not usually formal. In particular, they found that confirmation judgments in the urn setting were more extreme versions of corresponding judgments in the men/women scenario (i.e., urns produced higher values for positive impact, lower for negative impact). Confirmation seems therefore to depend on more than just $Pr(E \wedge H)$, $Pr(E \wedge \neg H)$, $Pr(\neg E \wedge H)$, and $Pr(\neg E \wedge \neg H)$.

Tentori, Crupi, and Osherson (2007) speculated that subjective assessments of confirmation may depend on the reasoner's degree of belief (or confidence) in the probabilities in play. The missing variable from formality might then be identified with a *second-order probability*—a probability distribution defined over the possible values of $Pr(E \wedge H)$, $Pr(E \wedge \neg H)$, $Pr(\neg E \wedge H)$, and $Pr(\neg E \wedge \neg H)$. Thus, unlike the sharp chances of the urn setting, the probability values elicited in the men/women scenario presumably have a probability lower than 1, perhaps because they bring to mind one's relative ignorance of the topic at issue (e.g., smoking habits). Such a lower second-order probability might trigger more cautious confirmation judgments, biased toward less extreme values.

The evidence presented by Tentori, Crupi, and Osherson (2007), however, does not definitively support the second-order probability hypothesis. The difference between confirmation judgments in the urn setting versus men/women scenario may depend on difference in content (e.g., concrete vs. abstract) without mediation by second-order probability. In this case, confirmation judgments could still be formal in a weaker sense—that is, relative to a given semantic domain. In other words, queries involving the same kind of scenario in which the relevant probabilities are equated might yield equivalent judgments of confirmation, but equivalence across different domains would not necessarily hold.

In the following experiments, we provide evidence against the weak-formality hypothesis, favoring second-order probability instead. This is achieved by comparing confirmation judgments from pairs of problems within the same semantic domain that are equivalent in terms of $Pr(E \wedge H)$, $Pr(E \wedge \neg H)$, $Pr(\neg E \wedge H)$, and $Pr(\neg E \wedge \neg H)$ but differ in terms of the probability with which such quantities can be identified.

EXPERIMENT 1

Method

Participants. Forty-four students from the University of Trento participated in the experiment (22 female; mean age, 24.9 years). In order to allow direct manipulation of second-order probability levels, all confirmation judgments were elicited in a setting involving urns. Participants were randomly divided into two groups of 22. A given participant performed his/her task twice, each time with reference to a different pair of urns, as now described. (In the following procedure, all events described as random to participants were genuine results of chance.)

Procedure and Materials. The two pairs of urns presented to Group 1 are labeled (A_{10},B_{10}) and (C_{10},D_{10})—the subscript "10" denotes the number of balls in each urn—and were composed as follows:

Urn	Number of Red Balls	Number of Green Balls
A_{10}	7	3
B_{10}	3	7
C_{10}	9	1
D_{10}	1	9

The order of presentation of (A_{10},B_{10}) and (C_{10},D_{10}) was counterbalanced across participants. The procedure was the same for the two pairs; we illustrate with (A_{10},B_{10}). Participants first verified the composition of the urns by counting the number of balls of each color. It was then explained that a coin toss would determine the choice of one of the two urns, the outcome being kept hidden. The coin was tossed, one urn was covertly selected, and the other was set aside. Next, the chosen urn was shaken, the participant was asked to blindly draw a ball from it, and the color of the chosen ball was exposed. For the remainder of the trial, the chosen ball remained in view along with a reminder (both numerical and graphical) of the composition of the pair of urns at issue. Finally, participants were asked to mark a position on an "impact scale" to indicate how the outcome of the draw influenced their current conviction that a given urn (e.g., A_{10}) had been selected by the coin toss. A fresh copy of the scale was then used to express the evidential impact of the same draw for the other urn (B_{10}). The impact scale was a 20-cm-long line printed on a strip of paper (see Figure 1). It had two opposite directions (corresponding to *positive* and *negative* impacts, respectively) as well as a neutral point in the middle (corresponding to *no* impact).

The two pairs of urns presented to Group 2 are labeled (A_{100},B_{100}) and (C_{100},D_{100}), and were composed as follows:
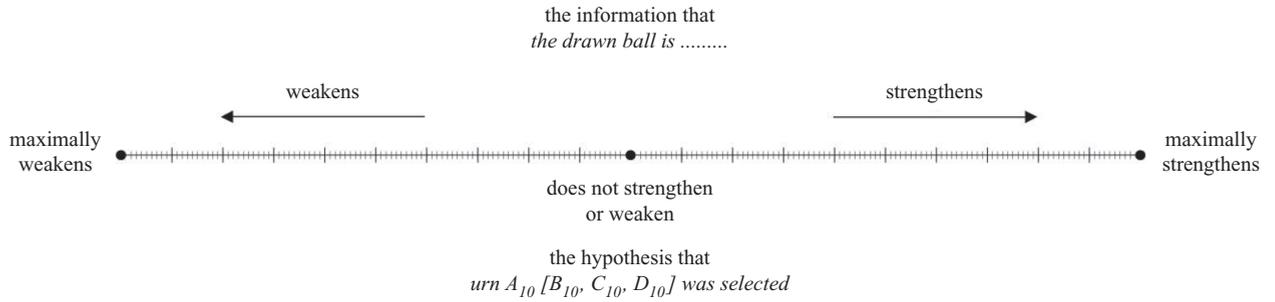
the information that
the drawn ball is ....

Figure 1. The impact scale employed. Dots were filled with the color of the drawn ball (red vs. green).

Urn	Number of Red Balls	Number of Green Balls
A_{100}	70	30
B_{100}	30	70
C_{100}	90	10
D_{100}	10	90

The order of presentation of (A_{100}, B_{100}) and (C_{100}, D_{100}) was counterbalanced across participants. The procedure was the same for the two pairs; we illustrate with (A_{100}, B_{100}). Participants first verified the composition of the urns by counting the number of balls of each color. Subsequently, they were asked to blindly draw 10 balls from one of the two urns (e.g., A_{100}) and to place them, still blindly, into another empty opaque urn, thus forming a new urn (A_{10}^*). The same was done for the other urn (B_{100}). Concerning the new pair of urns (A_{10}^*, B_{10}^*), participants then performed the same procedure as Group 1.

In Group 1, the exact composition of the urns at issue is known so that maximum probability is attached to corresponding probability values connecting ball colors and urns (*maximum second-order probability*). In contrast, in Group 2 the composition of A_{10}^*, B_{10}^*, C_{10}^*, and D_{10}^* obeys hypergeometric distributions allowing in principle any possible combination of red and green balls (*low second-order probability*). The following equalities, however, were preserved by the procedure (see the Appendix for relevant calculations):

$$\Pr(red \wedge A_{10}) = \Pr(red \wedge A_{10}^*)$$
$$= \Pr(green \wedge B_{10})$$
$$= \Pr(green \wedge B_{10}^*) = .35,$$
$$\Pr(green \wedge A_{10}) = \Pr(green \wedge A_{10}^*)$$
$$= \Pr(red \wedge B_{10})$$
$$= \Pr(red \wedge B_{10}^*) = .15,$$
$$\Pr(red \wedge C_{10}) = \Pr(red \wedge C_{10}^*)$$
$$= \Pr(green \wedge D_{10})$$
$$= \Pr(green \wedge D_{10}^*) = .45,$$
$$\Pr(green \wedge C_{10}) = \Pr(green \wedge C_{10}^*)$$
$$= \Pr(red \wedge D_{10})$$
$$= \Pr(red \wedge D_{10}^*) = .05. \qquad (4)$$

The weak-formality hypothesis thus implies equal confirmation judgments across the two groups. In contrast, if low second-order probability makes confirmation less extreme, responses in Group 2 should be more cautious than those in Group 1.

Results

All pairs of urns were symmetrical in composition—for example, 7 red versus 3 green balls for A_{10}, and 3 red versus 7 green balls for B_{10}. We therefore aggregated judgments concerning positive impact (e.g., the impact of a red ball on A_{10} with that of a green ball on B_{10}), as well as judgments concerning negative impact (e.g., the impact of a green ball on A_{10} with that of a red ball on B_{10}).

There were no judgments of positive/negative impact when the posterior probability was lower/higher than the prior. This suggests that the task was appropriately understood by participants.

The mean and median confirmation judgments reported by the two groups are shown in Table 1. In line with the confidence hypothesis, Group 2 confirmation judgments are less extreme than those of Group 1. This difference is statistically significant for all four relevant matched pairs by both an independent samples t test ($p < .05$) and a Mann–Whitney U test ($p < .05$). Thus, within the same semantic domain (namely, urns), evidence can have different impacts on given hypotheses. Our results therefore contradict the weak-formality hypothesis: Lower second-order probability leads to less extreme confirmation judgments, despite all equalities in (4).

If formality is to be retained as a normative property of confirmation, it is necessary to qualify Group 1 ratings as too extreme, or Group 2 ratings as too moderate, or both. The choice obviously presupposes a normative benchmark. For this purpose, we focus on the measures l and z in (2) and (3), respectively. They have been shown to enjoy a number of desirable properties not satisfied by rivals (Crupi et al., 2007; Eells & Fitelson, 2002). Moreover, at the descriptive level, they are the best predictors of confirmation judgment (Crupi et al., 2007; Tentori, Crupi, Bonini, & Osherson, 2007). Each of l and z achieves a fixed finite maximum [minimum] value if E implies [contradicts] H.[2] We may set these two values at $+10$ and -10, respectively, corresponding to our rating scale (see Fig-

Table 1
Mean and Median Confirmation Judgments Observed in Groups 1 and 2, and Corresponding Tests for Significance of Differences

Impact	Group 1 (A_{10};B_{10})		Group 2 (A_{10}^*;B_{10}^*)		Group 1 Versus Group 2	
	Mean	Median	Mean	Median	t	z
Positive	3.8	4	2.3	2	2.2†	−2.1†
Negative	−3.6	−4	−2.1	−2	−2.4†	−2.3†
	Group 1 (C_{10};D_{10})		Group 2 (C_{10}^*;D_{10}^*)		Group 1 Versus Group 2	
	Mean	Median	Mean	Median	t	z
Positive	7.7	8	5.3	6	3.2†	−3.1†
Negative	−7.4	−8	−4.5	−5	−3.5†	−3.2†

†$p < .05$, one-tailed.

ure 1). It then turns out that the two measures produce the same values for the events in our experiment, as follows:

$$\begin{aligned}
\mathrm{CONF}(red, A_{10}) &= \mathrm{CONF}(red, A^*_{10}) \\
&= \mathrm{CONF}(green, B_{10}) \\
&= \mathrm{CONF}(green, B^*_{10}) = +4,
\end{aligned}$$

$$\begin{aligned}
\mathrm{CONF}(green, A_{10}) &= \mathrm{CONF}(green, A^*_{10}) \\
&= \mathrm{CONF}(red, B_{10}) \\
&= \mathrm{CONF}(red, B^*_{10}) = -4,
\end{aligned}$$

$$\begin{aligned}
\mathrm{CONF}(red, C_{10}) &= \mathrm{CONF}(red, C^*_{10}) \\
&= \mathrm{CONF}(green, D_{10}) \\
&= \mathrm{CONF}(green, D^*_{10}) = +8,
\end{aligned}$$

$$\begin{aligned}
\mathrm{CONF}(green, C_{10}) &= \mathrm{CONF}(green, C^*_{10}) \\
&= \mathrm{CONF}(red, D_{10}) \\
&= \mathrm{CONF}(red, D^*_{10}) = -8.
\end{aligned}$$

Comparison with Table 1 reveals that the confirmation judgments of Group 1 are remarkably close to those advocated by l and z. Group 2, in contrast, diverges from the l, z recommendations ($p < .05$ for all four relevant matched pairs by both one-sample t test and one-sample Wilcoxon test). If formality is adopted as a normative desideratum, it thus appears that confirmation assessment is sound in Group 1 (*maximum second-order probability*) but too moderate in Group 2 (*low second-order probability*). None of this, of course, is a reason to adopt formality as a desirable property of confirmation measures.

CONTROL EXPERIMENTS

Before embracing the conjecture that second-order probability affects confirmation, we must address two potential ambiguities in the procedure of Experiment 1.

First, it is necessary to ensure that the equalities in (4) hold up, not just objectively, but subjectively, as well. In fact, one might wonder if participants in Group 2 fully understood the nature of the sampling process and the resulting probability distributions. In particular, had subjectively expected proportions of red versus green balls been more regressive (less extreme) than the correct ones, this could account for the results observed quite apart from a direct effect of second-order probability.[3]

To rule out this possibility, we ran a separate experiment. A new group of 22 participants (10 female; mean age, 28.7 years) were presented with the same procedure as Group 2 concerning urns A_{100} and C_{100}. Once A^*_{10} and C^*_{10} were formed, participants were asked to indicate the most likely sample composition of A^*_{10} [C^*_{10}] (in terms of the number of red/green balls) and the likelihood of drawing a red ball from A^*_{10} [C^*_{10}]. The order of the two questions was counterbalanced; for half of the participants we introduced A_{100} first, for the other half, C_{100}. All participants but 2 (91%) identified "7 red balls and 3 green balls" as the most likely composition for A^*_{10}, and all but 2 (91%) identified "9 red balls and 1 green ball" as the most likely composition for C^*_{10}. Moreover, mean and median estimates for a red ball being drawn from A^*_{10} (.684 and .7, respectively) as well as from C^*_{10} (.898 and .9, respectively) were statistically indistinguishable from the corresponding objec-

Table 2
Mean and Median Confirmation Judgments Observed in Group 3, and Corresponding Tests for Significance of Differences With Groups 1 and 2

Impact	Group 3 (A_{100};B_{100})		Group 1 Versus Group 3		Group 2 Versus Group 3	
	Mean	Median	t	z	t	z
Positive	3.6	4	0.2	−0.2	−2.2*	−2.3*
Negative	−3.3	−3	−0.4	−0.5	2.2*	−2.2*

	Group 3 (C_{100};D_{100})		Group 1 Versus Group 3		Group 2 Versus Group 3	
	Mean	Median	t	z	t	z
Positive	7.0	8	1.2	−1.3	−2.1*	−2.1*
Negative	−6.2	−8	−1.5	−1.5	1.8*	−1.9*

*$p < .05$, one-tailed.

tive values (.7 and .9) by one-sample t test and one-sample Wilcoxon test. We can conclude that the procedure adopted for Group 2 in Experiment 1 conveys a transparent representation of the sampling process, producing appropriate assessments of the probability values appearing in (4).

A second concern is related to a documented tendency, labeled "ratio bias," to judge the probability of an event as different when expressed as a ratio of small (e.g., 1:10) versus large (e.g., 10:100) numbers (see, e.g., Denes-Raj & Epstein, 1994; Kirkpatrick & Epstein, 1992). The proportion of red versus green balls initially provided to Groups 1 and 2 was the same, but it referred to samples of different size—namely, a 10-ball urn for Group 1 versus a 100-ball urn for Group 2. In order to check for the possible effect of this difference in numerosity, we carried out a further control experiment. Twenty-two students from the University of Trento participated in this experiment (10 female; mean age, 24.4 years), henceforth called Group 3. None had participated in the previous experiments. The procedure for Group 3 was identical to that for Group 1 except that the hundred-ball urns A_{100}, B_{100}, C_{100}, D_{100} initially presented to Group 2 were used in place of A_{10}, B_{10}, C_{10}, D_{10}.

The confirmation judgments expressed by Group 3 are reported in Table 2, and compared to the judgments of the two groups from Experiment 1. It can be seen that the Group 3 judgments differ significantly by both independent samples t test ($p < .05$) and Mann–Whitney U test ($p < .05$) from those of Group 2, and are more extreme. On the other hand, Group 3 judgments do not differ significantly from those of Group 1, despite the difference in the number of balls per urn. This pattern of results rules out the possibility that the contrasting judgments between Groups 1 and 2 are a consequence of the different numbers of balls presented at the outset. The results from Group 3 once again discredit weak formality, and support the hypothesis that judged confirmation depends partly on second-order probability.

DISCUSSION

A confirmation measure CONF is *formal* if CONF(E,H) is determined by the four probabilities Pr(E∧H), Pr(E∧¬H),

$Pr(\neg E \wedge H)$, and $Pr(\neg E \wedge \neg H)$. Previous experimentation (Tentori, Crupi, & Osherson, 2007) suggests that judgments of evidential impact cannot be modeled by a formal confirmation measure. But this finding leaves open the possibility of *weak formality*—namely, that a formal confirmation measure accurately predicts confirmation judgments within a given semantic domain (e.g., urns). The present investigation, however, appears to close off this possibility. Pairs of urn problems were exhibited that agree on the four probabilities displayed in (4) but nonetheless generate distinct confirmation judgments. Specifically, when for each of those quantities a pointwise value was fixed with a maximal second-order probability, participants rated evidence impact in accordance with normatively credible confirmation measures [namely, l and z shown in (2) and (3), respectively]; otherwise, rated impact was systematically more moderate. Two control experiments allowed us to exclude potential misunderstandings as causes of this effect, by demonstrating that assessments of the probabilities in (4) were suitably aligned across the two conditions of Experiment 1. As a consequence, the results obtained in Experiment 1 appear to document a genuine violation of weak formality.

It is well known that ambiguity about probabilities is a potent variable in preference among bets (Einhorn & Hogarth, 1985, 1986; Ellsberg, 1961), and that a common procedure for manipulating ambiguity is to vary the decision-maker's confidence in probability estimates (Heath & Tversky, 1991). Compared to earlier studies, however, the present investigation is different in both method and implications.

First, our experimental manipulation is different from those employed in studying the effects of ambiguity. In particular, ambiguity is typically induced by limiting knowledge of the process that generates outcomes, or is simulated by hypothetical scenarios. For example, Ellsberg (1961) presented urns containing unspecified numbers of colored balls and did not explain how they were obtained. In Hogarth and Kunreuther (1989), as well as in Heath and Tversky (1991), participants were directly asked to assume they were experiencing "confidence" versus "considerable uncertainty" about the probability of a hypothetical event. In both procedures, the detailed circumstances yielding ambiguity remain unexplained to participants (in the latter case, participants did not even generate a probability estimate or associated confidence on their own). The possibility is thus left open that participants' mistrust or ignorance played a role in their reaction to the ambiguity. In contrast, our manipulation of second-order probability did not rely on incomplete information, but was instead based on a transparent stochastic process under participants' control.

Second, our experiments document the effect of second-order probability in a purely cognitive context, inasmuch as participants were not asked to choose among uncertain prospects; that is, second-order probability was shown to affect judgment of evidential impact, which concerns change in belief rather than perceived value. The separable impact of ambiguity on belief versus choice is suggested by previous studies showing that ambiguity aversion in bets is affected by variables extraneous to confirmation, such as fear of negative evaluation (Curley, Yates, & Abrams, 1986). Heath and Tversky (1991) documented further discrepancies between choice and judgment when ambiguity is involved. It remains for further inquiry to clarify the psychological relation between ambiguity in choice versus second-order probability in confirmation.

One is free to assimilate the violation of (weak) formality to the list of other complaints about human reasoning under uncertainty (see, e.g., Girotto & Gonzalez, 2001; Tentori, Bonini, & Osherson, 2004). There is a difference, however, between violating coherence constraints when estimating probabilities and contradicting a class of confirmation measures. In the former case, criticism of human reasoning is backed by theorems linking incoherence to suboptimal outcomes (e.g., guaranteed worse penalties under any *proper scoring rule*; see Predd et al., 2009). Arguments for or against given confirmation measures, in contrast, are typically based on intuition about "clear cases" (see, e.g., Eells & Fitelson, 2002). Widespread tendencies in "lay" judgment should therefore be taken seriously.

In any event, the focus here is descriptive rather than normative. Our results imply that no confirmation measure among those commonly considered is entirely accurate as a model of human intuition about evidential impact, even within a given semantic domain. To model the absence of weak formality in confirmation judgment, consider the following parametric family of measures, appearing in Crupi et al. (2007, p. 233):

$$z_\alpha(E, H) = \begin{cases} z(E, H)^\alpha & \text{if } P(H|E) \geq P(H), \\ -\,|\,z(E, H)\,|^\alpha & \text{otherwise.} \end{cases}$$

The parameter α may be set to unity when second-order probability is maximal and assume higher values as it decreases. This scheme yields underrated assessments of evidential impact from low second-order probability, thus capturing the qualitative phenomenon documented here. The precise relation between α and second-order probability, as well as a quantitative test of the foregoing account, are topics for further inquiry.

REFERENCES

CARNAP, R. (1962). *Logical foundations of probability* (2nd ed.). Chicago: University of Chicago Press. (Original work published 1950)
CHENG, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review*, **104**, 367-405.
CRUPI, V., TENTORI, K., & GONZALEZ, M. (2007). On Bayesian theories of evidential support: Theoretical and empirical issues. *Philosophy of Science*, **74**, 229-252.
CURLEY, S. P., YATES, J. F., & ABRAMS, R. A. (1986). Psychological

sources of ambiguity avoidance. *Organizational Behavior & Human Decision Processes*, **38**, 230-256.

DENES-RAJ, V., & EPSTEIN, S. (1994). Conflict between intuitive and rational processing: When people behave against their better judgment. *Journal of Personality & Social Psychology*, **66**, 819-829.

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

EINHORN, H. J., & HOGARTH, R. M. (1985). Ambiguity and uncertainty in probabilistic inference. *Psychological Review*, **92**, 433-461.

EINHORN, H. J., & HOGARTH, R. M. (1986). Decision under ambiguity. *Journal of Business*, **59**, S225-S250.

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

FITELSON, B., & HITCHCOCK, C. (in press). Probabilistic measures of causal strength. In F. Russo & J. Williamson (Eds.), *Causality in the sciences*. Oxford: Oxford University Press.

GIROTTO, V., & GONZALEZ, M. (2001). Solving probabilistic and statistical problems: A matter of information structure and question form. *Cognition*, **78**, 247-276.

GOOD, I. J. (1983). *Good thinking*. Minneapolis: University of Minnesota Press.

HEATH, C., & TVERSKY, A. (1991). Preference and beliefs: Ambiguity and the competence in choice under uncertainty. *Journal of Risk & Uncertainty*, **4**, 5-28.

HOGARTH, R. M., & KUNREUTHER, H. C. (1989). Risk, ambiguity, and insurance. *Journal of Risk & Uncertainty*, **2**, 5-35.

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

KIRKPATRICK, L. A., & EPSTEIN, S. (1992). Cognitive–experiential self-theory and subjective probability: Further evidence for two conceptual systems. *Journal of Personality & Social Psychology*, **63**, 534-544.

PREDD, J., SEIRINGER, R., LIEB, E. H., OSHERSON, D., POOR, V., & KULKARNI, S. (2009). Probabilistic coherence and proper scoring rules. *IEEE Transactions on Information Theory*, **55**, 4786-4792.

TENTORI, K., BONINI, N., & OSHERSON, D. (2004). The conjunction fallacy: A misunderstanding about conjunction? *Cognitive Science*, **28**, 467-477.

TENTORI, K., CRUPI, V., BONINI, N., & OSHERSON, D. (2007). Comparison of confirmation measures. *Cognition*, **103**, 107-119.

TENTORI, K., CRUPI, V., & OSHERSON, D. (2007). Determinants of confirmation. *Psychonomic Bulletin & Review*, **14**, 877-883.

NOTES

1. Confirmation measures are in some respects related to the measures of strength appearing in psychological studies of causal judgment (e.g., Cheng, 1997). See Tentori, Crupi, Bonini, and Osherson (2007) and Fitelson and Hitchcock (in press) for discussion.

2. Both measures l and z range between -1 and $+1$. Notice that, when E implies H, $Pr(H|E) = 1$ and $Pr(E|\neg H) = 0$. On the other hand, when E contradicts H, $Pr(H|E) = Pr(E|H) = 0$. Simple algebra shows that $l(E,H) = z(E,H) = +1$ in the former case, and $l(E,H) = z(E,H) = -1$ in the latter.

3. We thank an anonymous reviewer for prompting us to address this concern explicitly.

APPENDIX

Participants in Group 2 were required to sample $n = 10$ balls without replacement from a population of $N = 100$ balls, of which R are red and $N–R$ are green. In this set up, the discrete probability distribution over the number r of red balls in sample n is the hypergeometric distribution, whereby the expected value of r amounts to $R \times (n/N)$.

As a consequence, probabilities of a red ball being drawn from A_{10}^*, B_{10}^*, C_{10}^*, and D_{10}^* are, respectively, as follows:

$$Pr(red|A_{10}^*) = [70 \times (10/100)]/10 = .7$$
$$Pr(red|B_{10}^*) = [30 \times (10/100)]/10 = .3$$
$$Pr(red|C_{10}^*) = [90 \times (10/100)]/10 = .9$$
$$Pr(red|D_{10}^*) = [10 \times (10/100)]/10 = .1$$

Thus:

$$Pr(red \wedge A_{10}^*) = Pr(red|A_{10}^*) \times Pr(A_{10}^*) = .7 \times .5 = .35$$
$$Pr(red \wedge B_{10}^*) = Pr(red|B_{10}^*) \times Pr(B_{10}^*) = .3 \times .5 = .15$$
$$Pr(red \wedge C_{10}^*) = Pr(red|C_{10}^*) \times Pr(C_{10}^*) = .9 \times .5 = .45$$
$$Pr(red \wedge D_{10}^*) = Pr(red|D_{10}^*) \times Pr(D_{10}^*) = .1 \times .5 = .05$$