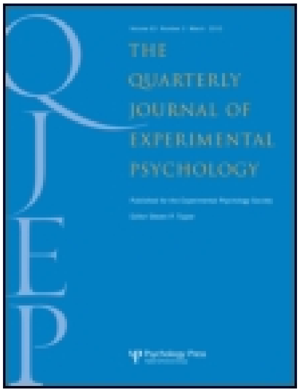
This article was downloaded by: [Universita' Milano Bicocca]
On: 17 December 2013, At: 03:17
Publisher: Routledge
Informa Ltd Registered in England and Wales Registered Number: 1072954 Registered office: Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK



The Quarterly Journal of Experimental Psychology

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/pqje20

Insensitivity and oversensitivity to answer diagnosticity in hypothesis testing

Patrice Rusconi[a] & Craig R. M. McKenzie[b]

[a] Department of Psychology, University of Milano-Bicocca, Milano, Italy

[b] Rady School of Management and Department of Psychology, University of California, San Diego, San Diego, CA, USA

Accepted author version posted online: 09 Apr 2013.Published online: 16 May 2013.

To cite this article: Patrice Rusconi & Craig R. M. McKenzie (2013) Insensitivity and oversensitivity to answer diagnosticity in hypothesis testing, The Quarterly Journal of Experimental Psychology, 66:12, 2443-2464, DOI: 10.1080/17470218.2013.793732

To link to this article: http://dx.doi.org/10.1080/17470218.2013.793732

PLEASE SCROLL DOWN FOR ARTICLE

Taylor & Francis makes every effort to ensure the accuracy of all the information (the “Content”) contained in the publications on our platform. However, Taylor & Francis, our agents, and our licensors make no representations or warranties whatsoever as to the accuracy, completeness, or suitability for any purpose of the Content. Any opinions and views expressed in this publication are the opinions and views of the authors, and are not the views of or endorsed by Taylor & Francis. The accuracy of the Content should not be relied upon and should be independently verified with primary sources of information. Taylor and Francis shall not be liable for any losses, actions, claims, proceedings, demands, costs, expenses, damages, and other liabilities whatsoever or howsoever caused arising directly or indirectly in connection with, in relation to or arising out of the use of the Content.

This article may be used for research, teaching, and private study purposes. Any substantial or systematic reproduction, redistribution, reselling, loan, sub-licensing, systematic supply, or distribution in any form to anyone is expressly forbidden. Terms & Conditions of access and use can be found at http://www.tandfonline.com/page/terms-and-conditions

THE QUARTERLY JOURNAL OF EXPERIMENTAL PSYCHOLOGY, 2013
Vol. 66, No. 12, 2443–2464, http://dx.doi.org/10.1080/17470218.2013.793732

Downloaded by [Universita' Milano Bicocca] at 03:17 17 December 2013

Insensitivity and oversensitivity to answer diagnosticity in hypothesis testing

Patrice Rusconi[1] and Craig R. M. McKenzie[2]

[1]Department of Psychology, University of Milano-Bicocca, Milano, Italy
[2]Rady School of Management and Department of Psychology, University of California, San Diego, San Diego, CA, USA

Two experiments examined how people perceive the diagnosticity of different answers ("yes" and "no") to the same question. We manipulated whether the "yes" and the "no" answers conveyed the same amount of information or not, as well as the presentation format of the probabilities of the features inquired about. In Experiment 1, participants were presented with only the percentages of occurrence of the features, which most straightforwardly apply to the diagnosticity of "yes" answers. In Experiment 2, participants received in addition the percentages of the absence of features, which serve to assess the diagnosticity of "no" answers. Consistent with previous studies, we found that participants underestimated the difference in the diagnosticity conveyed by different answers to the same question. However, participants' insensitivity was greater when the normative (Bayesian) diagnosticity of the "no" answer was higher than that of the "yes" answer. We also found oversensitivity to answer diagnosticity, whereby participants valued as differentially diagnostic two answers that were normatively equal in terms of their diagnosticity. Presenting to participants the percentages of occurrence of the features inquired about together with their complements increased their sensitivity to the diagnosticity of answers. We discuss the implications of these findings for confirmation bias in hypothesis testing.

Keywords: Hypothesis testing; Answer diagnosticity; Insensitivity; Oversensitivity; Feature-positive effect.

An efficient evaluation of the impact of the information that we receive or we acquire is of critical importance in many everyday life situations as well as in many professional contexts. Consider, for example, the case of a deluded patient who interprets the appearance of a police car in the busy street where she is walking as a cue that the police are chasing her. In this case, the patient is disregarding the probability that the police are chasing another person and not her (Hemsley & Garety, 1986). In other words, the patient fails to value the evidential strength of the incoming evidence (the appearance of the police car) under hypotheses that are different from the one she is considering. Another way to misweigh evidence is failing to appreciate that the presence of an event might convey a different amount of information than its absence (e.g., Cherubini, Rusconi, Russo, & Crippa, 2013; Fischhoff & Beyth-Marom, 1983; Rusconi, Crippa, Russo, & Cherubini, 2012; Rusconi,

Correspondence should be addressed to Patrice Rusconi, Department of Psychology, University of Milano-Bicocca, Piazza dell'Ateneo Nuovo, 1, 20126 Milano, Italy. E-mail: patrice.rusconi1@unimib.it

This work was presented in other forms in the doctoral dissertation of the first author. This material is based upon work supported by the National Science Foundation under Grants SES-1062070 and SES-0820553. We thank Jonathan Nelson for insightful suggestions on an earlier draft and Simona Sacchi for inspiring Figure 1.

© 2013 The Experimental Psychology Society

Downloaded by [Universita' Milano Bicocca] at 03:17 17 December 2013

Marelli, Russo, D'Addario, & Cherubini, 2013). For example, an investigator might fail to notice that the absence of a suspect's fingerprints from a crime scene can be more revealing than their presence if the suspect was familiar with the victim. Similarly, eyewitness nonidentifications might be at least as informative as identifications (e.g., Clark & Wells, 2008; Wells & Lindsay, 1980).

The present article is concerned with a specific case of the latter type of failure in evidence evaluation. We address the issue of people's difficulty with perceiving the informativeness that "yes" and "no" answers to the same question convey regarding the plausibility of a focal hypothesis—that is, of a hypothesis that is being tested. Understanding the psychological mechanisms underlying this process is important not only to pursue sound reasoning in scientific research, medical diagnosis, and legal contexts, but also in more mundane circumstances, such as when forming impressions of others (e.g., Evett, Devine, Hirt, & Price, 1994; Fiedler & Walther, 2004).

Symmetry and asymmetry in the diagnosticity of answers

From a normative (Bayesian) standpoint, different answers (i.e., "yes" and "no") to the same question can differ in the amount of information that they convey (i.e., how diagnostic they are). When testing the hypothesis that a new acquaintance is an extrovert you might ask her "Do you enjoy parties?". A "yes" answer to this question is about as informative as a "no" answer. Indeed, if the new acquaintance replies "yes", you will be about as confident that she is an extrovert as you will be about her introversion after a "no". This type of questions has been called "symmetric" because there is symmetry in the amount of information conveyed by the two possible answers (Cameron & Trope, 2004; Cherubini, Rusconi, Russo, Di Bari, & Sacchi, 2010; Trope & Liberman, 1996; Trope & Thompson, 1997). In contrast, if you ask "Do you organize parties at your home each week?", a "yes" answer would provide you with more information about your new acquaintance's extroversion than a "no" answer about her introversion. People can be extroverts even if they do not organize parties at their home each week. Thus, a "no" should not disconfirm strongly the extroversion hypothesis that you are testing, while a "yes" should confirm it relatively strongly. For this reason, this type of question has been labelled "asymmetric", more specifically "asymmetrically confirming" (e.g., Cameron & Trope, 2004; Cherubini et al., 2010; Trope & Liberman, 1996; Trope & Thompson, 1997). Conversely, an asymmetrically disconfirming question implies that the disconfirming answer is more informative than the confirming answer. For example, asking "Do you enjoy being alone on Saturday night?" to test one's extroversion implies anticipating a hypothesis-disconfirming ("yes") answer that is highly informative about the target's introversion and a hypothesis-confirming answer ("no") that is not as informative about the target's extroversion (e.g., Brambilla, Rusconi, Sacchi, & Cherubini, 2011, Study 2; Trope & Thompson, 1997). Figure 1 illustrates the equal or the differential diagnosticity (represented by the solid arrows) of the "yes" and "no" answers that follow symmetric and asymmetric tests, respectively.

Failure in evaluating appropriately the informativeness of different answers to the same question might lead to inefficiencies in belief revision. In fact, "optimal revision of initial beliefs depends on the diagnosticity of the specific answers received" (Slowiaczek, Klayman, Sherman, & Skov, 1992, p. 393).

Bayesian background

A widely used criterion for belief updating is Bayes' theorem, which provides a mathematical expression of how people should revise their initial confidence in a focal hypothesis in light of new evidence. A simple formulation of Bayes' theorem is given by the following equation in terms of odds (e.g., Beyth-Marom & Fischhoff, 1983; Fischhoff & Beyth-Marom, 1983):

$$\frac{p(H|D)}{p(\neg \mathrm{H}|D)} = \frac{p(H)}{p(\neg H)} \times \frac{p(D|H)}{p(D|\neg H)}$$

where, "p()" means "probability of", "|" should be read as "given that", "H" stands for the focal

Downloaded by [Universita' Milano Bicocca] at 03:17 17 December 2013

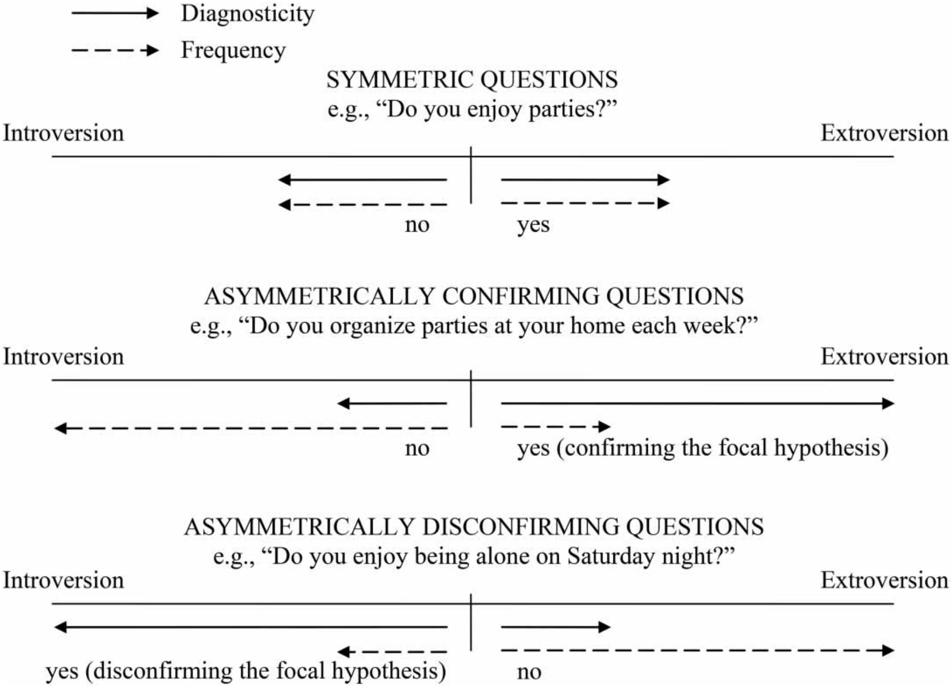

Diagnosticity
Frequency
SYMMETRIC QUESTIONS
e.g., "Do you enjoy parties?"
Introversion
Extroversion
no
yes
ASYMMETRICALLY CONFIRMING QUESTIONS
e.g., "Do you organize parties at your home each week?"
Introversion
Extroversion
no
yes (confirming the focal hypothesis)
ASYMMETRICALLY DISCONFIRMING QUESTIONS
e.g., "Do you enjoy being alone on Saturday night?"
Introversion
Extroversion
yes (disconfirming the focal hypothesis)
no

Figure 1. *Examples of symmetric and asymmetric questions when inquiring about the bipolar dimension of extroversion–introversion (assuming equal priors). In these examples, the focal hypothesis is extroversion while the alternate is introversion. Arrows point to the dimension of the bipolar trait favoured by the answer. Solid arrows indicate the degree of diagnosticity of yes and no answers: The longer the arrow, the higher the answer's diagnosticity is. In contrast, dashed arrows indicate the frequency of yes and no answers: The longer the dashed arrow, the more likely the answer is. Note that whenever the prior probabilities of the hypotheses (e.g., introversion and extroversion) are equal there is always a trade-off between frequency and diagnosticity, whereby likely answers are less diagnostic than unlikely ones.*

hypothesis, "¬*H*" for the alternate ("¬" is the logical symbol for negation), and "*D*" indicates the new evidence. From the left of the equation, there are three terms: (a) the posterior odds—that is, the ratio of the probability that the focal hypothesis is true after receiving the new evidence to the probability that the alternate is true given the same evidence; (b) the prior odds that the focal hypothesis, rather than the alternate, is true prior to receiving the new evidence; (c) the likelihood ratio (LR henceforth)—that is, the ratio of the probability of occurrence of the new evidence given that the focal hypothesis is true to the probability of occurrence of the same evidence given that the focal hypothesis is false (and thus the alternate is true).

How to evaluate the informativeness of answers

From Bayes' theorem one can derive directly two measures of the evidential strength of a datum—namely, the LR and the log LR (e.g., Cherubini et al., 2010; Good, 1950, 1979; McKenzie, 2004; Slowiaczek et al., 1992). However, Nelson's (2005, 2008) studies on information gathering pointed out some flaws of these metrics. Other measures of the value of obtained evidence have been proposed in the psychological literature (e.g., Crupi, Tentori, & Gonzalez, 2007; Mastropasqua, Crupi, & Tentori, 2010; Nelson, 2005, 2008; Tentori, Crupi, Bonini, & Osherson, 2007). Among them, probability gain is a measure of error reduction

Downloaded by [Universita' Milano Bicocca] at 03:17 17 December 2013

that was found to best capture people's intuitions about information acquisition (Nelson, McKenzie, Cottrell, & Sejnowski, 2010; see however Meier & Blair, 2013, for findings showing people's preference for efficiency over probability gain). Whenever the prior probabilities of the hypotheses are equal, the values of utility predicted by probability gain are identical to those of another metric, impact. This measure quantifies the absolute change in beliefs from the prior to the posterior probabilities of the hypotheses (e.g., Nickerson, 1996). In particular, the impact of a "yes" answer (or, in an equivalent form, the evidential value of the presence of a feature) can be expressed as follows (see, e.g., Nelson, 2005, Appendix A):

$$1/2 \times \{|p(H|D) - p(H)| + |p(\neg H|D) - p(\neg H)|\}$$

while the impact of a "no" answer (or the evidential strength of the absence of a datum) is computed as:

$$1/2 \times \{|p(H|\neg D) - p(H)| + |p(\neg H|\neg D) - p(\neg H)|\}$$

When there are two exhaustive and mutually exclusive hypotheses, the impact of a "yes" answer reduces to:

$$|p(H|D) - p(H)|$$

while the impact of a "no" answer can be computed as:

$$|p(H|\neg D) - p(H)|$$

Consider, for example, the planetary scenario introduced by Skov and Sherman (1986; see also Garcia-Marques, Sherman, & Palma-Oliveira, 2001; McKenzie, 2004, 2006; Nelson, 2005, 2008; Nelson et al., 2010; Sacchi, Rusconi, Russo, Bettiga, & Cherubini, 2012, Study 3; Slowiaczek et al., 1992; Villejoubert & Mandel, 2002). On an imaginary planet, Vuma, there are two kinds of inhabitants—Gloms and Fizos—which are equally numerous (i.e., the prior probability of encountering a Glom or a Fizo is .5) and invisible to human sight. The only way to identify the creatures is by asking about some features they possess. Participants are told the distribution of probabilities of the features across Gloms and Fizos. For example, participants are told that 90% of Gloms and 50% of Fizos drink gasoline. By applying Bayes' theorem, we can calculate the posterior probabilities of encountering either a Glom or a Fizo after the receipt of an answer to the question about drinking gasoline. If the tester receives a "yes" answer, the posterior probability that the encountered creature is a Glom in light of this answer is $p(Glom|``yes") = .64$, whereas the posterior probability that the creature is a Fizo given the same answer is $p(Fizo|``yes") = .36$. If the tester receives a "no" answer to the question about drinking gasoline, the posterior probability that the encountered creature is a Glom given this answer is $p(Glom|``no") = .17$, while the posterior probability that the creature is a Fizo given the same answer is $p(Fizo|``no") = .83$. In this example we consider two exhaustive and mutually exclusive hypotheses; therefore an answer confirms one hypothesis to the same extent as it disconfirms the other (e.g., Nelson, 2005; Nickerson, 1996). Accordingly, in this example, the formula for impact reduces to the absolute value of the difference between the posterior probability of either hypothesis and .5 (that is, the prior probability of either hypothesis). In particular, the impact of a "yes" answer is $|.64 - .5| = |.36 - .5| = .14$, while the impact of a "no" is $|.83 - .5| = |.17 - .5| = .33$. Thus, the "no" (disconfirming) answer is more informative than the "yes" (confirming) answer, and for this reason the question about drinking gasoline is asymmetrically disconfirming.

Evidence of insensitivity to differentially diagnostic answers

Although the Bayesian literature on evidence evaluation is large (e.g., Beach, 1968; Casscells, Schoenberger, & Graboys, 1978; Christensen-

Downloaded by [Universita' Milano Bicocca] at 03:17 17 December 2013

Szalanski & Bushyhead, 1981; Cosmides & Tooby, 1996; Fischhoff & Beyth-Marom, 1983; Gigerenzer & Hoffrage, 1995; Hammerton, 1973; McKenzie, 1994; Slovic & Lichtenstein, 1971; Villejoubert & Mandel, 2002), only a few studies have directly investigated how people revise their beliefs in light of different answers to the same question (McKenzie, 2006; Skov & Sherman, 1986; Slowiaczek et al., 1992). Overall, they showed that people appreciate that a "yes" and a "no" answer can convey different amounts of information, but they underestimate this difference in tasks with abstract materials. This phenomenon has been called "insensitivity to answer diagnosticity" (Slowiaczek et al., 1992). The first evidence of such insensitivity came from a study outlined by Skov and Sherman (1986) in the discussion of their seminal work (p. 118). Only 43% of participants showed the asymmetry of confidence in the normatively expected direction after a "yes" answer and after a "no" answer, and many were not asymmetric enough. Slowiaczek et al. (1992) found that, on average, participants estimated a difference of 6% between the posterior probability judgements after a "yes" and after a "no", while the normative difference was 19% (Experiment 1A, Slowiaczek et al., 1992). McKenzie (2006) replicated the findings of the study by Slowiaczek et al. (1992), but only with abstract materials (i.e., planetary scenarios). When participants were presented with familiar materials (i.e., scenarios about male and female heights) the extent of insensitivity to differentially diagnostic answers decreased. Although the familiarity of the materials used in the experiments turned out to be an important moderator of people's sensitivity to the differential diagnosticity of answers, it remains unclear how people behave in tasks with abstract materials.

Skov and Sherman (1986) hinted at a possible relation between people's failure to perceive the asymmetry in informativeness of different answers and the failure to consider base rates ($p(H)$), but they did not develop this idea or test it empirically.[1] Slowiaczek et al. (1992) advanced an explanation based on participants' confusion of the assessment of answer diagnosticity with the assessment of question usefulness, which, according to the authors, might be related to the use of the representativeness heuristic (e.g., Kahneman & Tversky, 1972; Tversky & Kahneman, 1974). Indeed, a shortcut that approximates the formal evaluation of question diagnosticity is the "feature-difference heuristic" (in fact, it is tantamount to impact; Nelson, 2005, Footnote 2; Nelson, 2009; Nelson et al., 2010; Slowiaczek et al., 1992). According to this shortcut, the most useful query is the one about a feature whose probability of occurrence is maximally different under two competing hypotheses. That is, the question with the highest diagnosticity is the one about a feature for which $\left|p(D|H) - p(D|\neg H)\right|$ is maximized.[2] Note that this shortcut for assessing *question* diagnosticity entails the consideration of the constituent probabilities of the LR—that is, one of the possible measures of *answer* diagnosticity (e.g., Cherubini et al., 2010; Good, 1950, 1979; McKenzie, 2004; Nelson, 2005; Slowiaczek et al., 1992). Indeed, on Nozick's account the difference between likelihoods is considered a measure of evidential support (e.g., Gigerenzer & Hoffrage, 1995; Schum, 1994). Several studies have shown that people are sensitive to the formal diagnosticity of

[1] We can speculate that the authors wanted to hint at the account of the phenomenon they subsequently gave in the study coauthored with Slowiaczek (Slowiaczek et al., 1992), which was based on the use of the representativeness heuristic (e.g., Kahneman & Tversky, 1972, 1973; Tversky & Kahneman, 1974). Indeed, the use of this strategy entails neglecting prior probabilities (e.g., Tversky & Kahneman, 1974).

[2] In the Bayesian reasoning literature, there are other examples of the use of this (or similar) strategy in belief revision. Indeed, one of the three most frequent non-Bayesian algorithms that Gigerenzer and Hoffrage (1995) found in problems using the standard probability format or the relative frequency format was the "likelihood subtraction"—that is, $p(D|H) - p(D|\neg H)$. Moreover, Hoffrage and Gigerenzer (1998) found that physicians asked to estimate the positive predictive values of diagnostic problems frequently used a strategy similar to the feature-difference heuristic when information was presented in form of probabilities. Indeed, one of the two most prevalent strategies when physicians did not reason according to Bayes' theorem was the difference between the sensitivity of a test—that is, $p(D|H)$ —and the false-positive rate of the test—that is, $p(D|\neg H)$. Note that the likelihood subtraction, also known as ΔR, has been long known as a strategy used in covariation assessment (e.g., Jenkins & Ward, 1965; McKenzie, 1994).

questions (e.g., Cherubini et al., 2010; Skov & Sherman, 1986; Slowiaczek et al., 1992; Trope & Bassok, 1982). Thus, Slowiaczek et al. (1992) argued that people do not perceive sufficiently the difference in the diagnosticity of "yes" and "no" answers to the same question because the difference between the constituent probabilities of the LR is the same for both "yes" and "no" answers. That is:

$$\left|p(D|H) - p(D|\neg H)\right| = \left|p(\neg D|H) - p(\neg D|\neg H)\right|$$

It should be noted that all previous studies focused on the differential impact of answers to *asymmetric* questions, for which a "yes" is more informative than a "no" or vice versa. Indeed, there are no empirical investigations of whether people perceive that the "yes" and "no" answers following a *symmetric* question convey the same amount of information. The two experiments presented in this article are aimed, in part, at filling this gap. This issue is relevant because it might clarify whether people's relative insensitivity to the differential diagnosticity of answers indicates only a tendency to perceive different answers as equally diagnostic (i.e., *underestimation* of differential evidence strength), or also as a failure to appreciate when different answers convey the same amount of information (i.e., *oversensitivity* to differential evidence strength), thus representing a more general failure in information use. Furthermore, oversensitivity to answer diagnosticity has implications for confirmation bias in hypothesis testing, defined as a tendency to apportion more confidence than warranted to the focal hypothesis (e.g., McKenzie, 2004, 2006), as we describe in the General Discussion.

More generally, taking into account the type of question (symmetric, asymmetrically confirming, or asymmetrically disconfirming) is important to elucidate the mechanisms underlying people's sensitivity to answer diagnosticity. Indeed, Slowiaczek et al. (1992, Experiment 1A) found greater insensitivity to answer diagnosticity when the answers came from asymmetrically disconfirming questions and thus when the "no" was more diagnostic than the "yes". Data from their Experiment 1A (see Slowiaczek et al., 1992, Table 2, p. 396) reveal that the estimated difference between "yes" and "no" diagnosticities was 4% for the 50–90% combination and 2% for the 90–50% combination. In contrast, when the answers came from asymmetrically confirming queries, and thus the "yes" was more diagnostic than the "no", participants were more sensitive to the differential diagnosticity of answers. Indeed, when the percentage combinations were either 50–10%, or 10–50%, the estimated difference between the informativeness of "yes" and "no" answers was 11%. From a normative (Bayesian) perspective, the difference is 19% for all percentage combinations (see Figure 2). This finding was not discussed by Slowiaczek et al. (1992) and is not accounted for by their explanation of insensitivity to answer diagnosticity. Indeed, the difference between the probabilities that constituted the LR was always 40% for both "yes" and "no" answers and for all the percentage combinations that the authors used. Accordingly, participants in Slowiaczek et al.'s (1992) study should have exhibited insensitivity to answer diagnosticity to the same extent regardless of the specific percentage combination that they received. The experiments presented in this article addressed this issue and add to the hypothesis-testing literature in two ways:

1. by considering symmetric tests—that is, the cases in which the "yes" and "no" answers are equally diagnostic (Experiments 1 and 2);
2. by using a presentation format that make explicit the likelihoods of feature absence in addition to the likelihoods of feature presence (Experiment 2).

EXPERIMENT 1

Method

Participants

One hundred and ten undergraduate students at the University of California, San Diego (66% female, mean age 20.1 years, range 18–28 years)

Downloaded by [Universita' Milano Bicocca] at 03:17 17 December 2013

Downloaded by [Universita' Milano Bicocca] at 03:17 17 December 2013

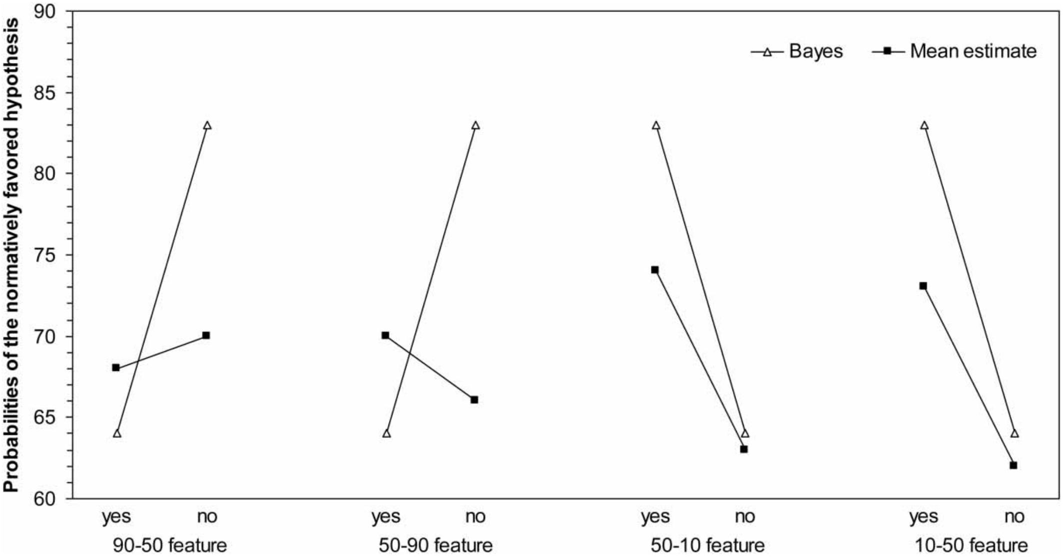


Figure 2. *Normative probabilities and mean estimated probabilities (in the form of percentages) of the hypothesis normatively favoured by the evidence drawn from Experiment 1A by Slowiaczek et al. (1992, Table 2, p. 396). The measures of variability for participants' estimates were absent in the original article. Note that participants were insensitive to answer diagnosticity for all tests. Indeed, normative slopes are steeper than participants' slopes for all the four tests shown in the figure. However, participants exhibited greater insensitivity to answer diagnosticity when the answers came from asymmetrically disconfirming tests (90–50 and 50–90 features) than when they came from asymmetrically confirming tests (50–10 and 10–50 features).*

took part in the experiment in exchange for course credit.

Materials and procedure

We set up a planetary scenario similar to the one originally introduced by Skov and Sherman (1986) and thereafter widely used in the literature on hypothesis testing (e.g., Garcia-Marques et al., 2001; McKenzie, 2004, 2006; Nelson, 2005, 2008; Nelson et al., 2010; Sacchi et al., 2012, Study 3; Slowiaczek et al., 1992; Villejoubert & Mandel, 2002). Specifically, we asked participants to imagine travelling to a planet, Vuma, where there is an equal number of two kinds of creatures, called Gloms and Fizos. Participants were presented with the answers to some questions about a series of features that Gloms and Fizos possess with different probabilities. The task was to surmise whether an encountered creature was a Glom (or a Fizo, according to the version of the questionnaire) based on the priors (50% of the creatures on the planet are Gloms, 50% are Fizos), the distributions of probabilities of the features inquired about, and the answers (i.e., “yes” or “no”) to the questions asked about these features (a sample stimulus is given in Appendix A).

We employed a 4 × 2 × 2 × 2 design. Within-participants variables were the type of test (Test 1 about a feature with probabilities of .98 and .50 under the two hypotheses, respectively; Test 2: .50–.02; Test 3: .65–.35; Test 4: .85–.15; see Table 1) and the answer that participants received (“yes” vs. “no”), whereas between-participants factors were the test order (Test 3, “yes”; Test 4, “no”; Test 1, “yes”; Test 2, “no”; Test 3, “no”; Test 4, “yes”; Test 1, “no”; Test 2, “yes”, and the reverse order) and the focal hypothesis (Glom vs. Fizo). Thus, there were four versions of the questionnaire, and each participant responded to eight problems, each presented on a separate page of the booklet. In particular, participants were asked to estimate the chances in 100 that the encountered creature was a Glom (Fizo) for each problem.

Downloaded by [Universita' Milano Bicocca] at 03:17 17 December 2013

Table 1. *The structure of the problems used in the two experiments*

Test	*Answer*	*Likelihoods*	*Impact/probability gain*	*Hypothesis supported by the answer*	*Normative probabilities of the supported hypothesis*
1	yes	.98/.5	.16	Glom	.66
	no	.02/.5	.46	Fizo	.96
2	yes	.5/.02	.46	Glom	.96
	no	.5/.98	.16	Fizo	.66
3	yes	.65/.35	.15	Glom	.65
	no	.35/.65	.15	Fizo	.65
4	yes	.85/.15	.35	Glom	.85
	no	.15/.85	.35	Fizo	.85

Note: Participants in Experiment 1 were presented only with the likelihoods (in the form of percentages) relative to the "yes" answer—that is, $p(D|Glom)$ and $p(D|Fizo)$—while in Experiment 2 they received both the likelihoods relative to the "yes" answer and those relative to the "no" answer (always in the form of percentages)—that is, $p(D|Glom)$, $p(D|Fizo)$ and their complements $p(\neg D|Glom)$ and $p(\neg D|Fizo)$.

The tests were chosen so that there were two questions (Tests 1 and 2) for which the "yes" and "no" answers conveyed a different amount of information. In Test 1 the "no" answer was more diagnostic than the "yes" answer and vice versa in Test 2 (see Table 1). In other words, the question about the 98–50 feature (Test 1) was asymmetrically disconfirming, while the query about the 50-2 feature (Test 2) was asymmetrically confirming. These types of tests have already been used in previous research on insensitivity to answer diagnosticity, and thus they were potentially useful to replicate the finding of an underestimation of the differential diagnosticity of "yes" and "no" answers to the same question. In addition, participants were presented with "yes" and "no" answers to symmetric questions, for which the "yes" answer was exactly as informative as the "no" answer. As can be seen in Table 1, both "yes" and "no" answers to Test 3 have impact of .15, while both "yes" and "no" answers to Test 4 have impact of .35. Thus, the only difference between these two tests lies in the informativeness of the answers to them: Both "yes" and "no" answers are more diagnostic following Test 4 than following Test 3. This allowed us to check whether participants were sensitive to variations in the amount of information conveyed by the answers they received, everything else kept constant. Furthermore, although Tests 3 and 4 did not allow us to assess whether participants underestimated the difference in diagnosticity between different answers, they could show whether they were calibrated or they were oversensitive to it.

Results

Following the procedure used by Slowiaczek et al. (1992) and McKenzie (2006), we recoded participants' estimates with respect to the hypothesis favoured by the answer. For example, a participant might receive a "no" answer to Test 2 (about the 50–2 feature) under the "Glom" focal hypothesis, and she or he might provide an estimate of 30% chance that the encountered creature is a Glom. This would be recoded as a 70% chance of encountering a Fizo. The hypothesis normatively supported by "yes" answers was always Glom, while it was always Fizo after "no" answers (see Table 1). Accordingly, we recoded four of the eight estimates for every participant.

Using this dependent variable, we performed a 4 × 2 × 2 × 2 mixed design analysis of variance (ANOVA). Within-participants variables were test type (Tests 1 to 4) and the answer participants received ("yes" vs. "no"), while between-groups factors were test order (one order and its reverse) and focal hypothesis (Glom vs. Fizo). The normative prediction is that there should be a significant main effect of test type but not of answer, and

Downloaded by [Universita' Milano Bicocca] at 03:17 17 December 2013

there should be a significant Test Type × Answer interaction. We present the results concerning these effects here, while we refer the reader to Appendix B for a description of the other findings.

There was a significant main effect of test type, $F(1, 105) = 14.78$, $MSE = 943.75$, $p < .001$, $\eta^2 = .027$, lower bound correction. The pairwise comparisons revealed a pattern that is consistent with normative considerations. Indeed, as we should expect by applying Bayes' rule (see the normative values in Table 1), participants gave higher estimates after an answer to Test 4 ($M = 68$, $SDE = 1.6$), than after answers to Test 1 ($M = 63.4$, $SDE = 1.5$), $p = .015$, and answers to Tests 2 and 3, *p*s < .001. Furthermore, there was no significant difference in the estimates after receiving the answers to asymmetric tests—that is, the answers to Test 1 and the answers to Test 2 ($M = 61$, $SDE = 1.6$), $p = .283$. Finally, participants provided significantly lower estimates after answers to Test 3 ($M = 56.9$, $SDE = 0.8$) than after answers to all other tests, *p*s ≤ .013. However, Figure 3 shows that participant estimates were overall lower than the normatively expected values.

Contrary to the normative prediction, there was a significant main effect of the answer, $F(1, 105) = 8.09$, $MSE = 1{,}621.55$, $p = .005$, $\eta^2 = .025$, lower bound correction, indicating that participants found "yes" answers ($M = 66.2$, $SDE = 1.5$) more informative than "no" answers ($M = 58.4$, $SDE = 1.8$). This finding is shown also by the slopes of the four data lines in Figure 3, which indicate an almost identical trend.

The Test Type × Answer interaction was not significant, $F(1, 105) = 1.33$, $MSE = 788.70$, $p = .251$, $\eta^2 = .002$, lower bound correction. This lack of effect is not normatively grounded: Looking at Figure 3 it is apparent that the normative prediction for the two asymmetric tests is a steep slope (top panels), while for the symmetric tests is a flat slope (bottom panels).

Using the same dependent variable as that described above, we performed a series of one-sample *t* tests comparing the normatively (Bayesian) expected and the observed difference between probabilities after a "yes" and probabilities after a "no" for each test. In particular, we computed the observed difference for each test by subtracting the estimate after a "no" from the estimate after a "yes" for each participant. A positive difference means that participants gave more weight to the "yes" answer, and a negative difference implies greater weight assigned to the "no" answer. We computed the theoretical Bayesian difference in the same manner. There is a normative difference of –30 for the asymmetrically disconfirming test (Test 1), of 30 for the asymmetrically confirming test (Test 2), and no difference for symmetric tests (Tests 3 and 4).

For the asymmetrically disconfirming question (98–50 feature), the mean estimated difference of 5.6 ($SD = 38$) was significantly lower than the normative difference of –30, $t(109) = 9.83$, $p < .001$, $d = 0.94$. Note that participants tended to value more the "yes" than the "no" answer. Thus, they tended to perceive an asymmetry in the opposite direction compared to the normatively expected direction (Figure 3, top left panel). For the asymmetrically confirming question (50–2 percentage combination), the mean difference of 10.4 ($SD = 50.2$) was significantly less than that normatively expected of 30, $t(109) = -4.1$, $p < .001$, $d = 0.39$ (Figure 3, top right panel).

For Test 3 (the 65–35 feature), the mean difference in the estimates after a "yes" and after a "no" ($M = 5.5$, $SD = 32.5$) was higher than the normatively expected difference of 0, although only marginally, $t(108) = 1.78$, $p = .078$, $d = 0.17$ (Figure 3, bottom left panel). Similarly, for Test 4 (85–15 feature), the mean difference in the estimates after a "yes" and after a "no" ($M = 8.2$, $SD = 39.2$) was significantly higher than the normatively expected null difference, $t(109) = 2.19$, $p = .031$, $d = 0.21$ (Figure 3, bottom right panel).

Discussion

The results of Experiment 1 replicate the findings of previous studies that pointed out people's insensitivity to answer diagnosticity when the "yes" and "no" answers come from asymmetric queries. The results extend previous findings by showing that people tend to be *over*sensitive to answer diagnosticity when the "yes" and "no" answers are equally

Downloaded by [Universita' Milano Bicocca] at 03:17 17 December 2013

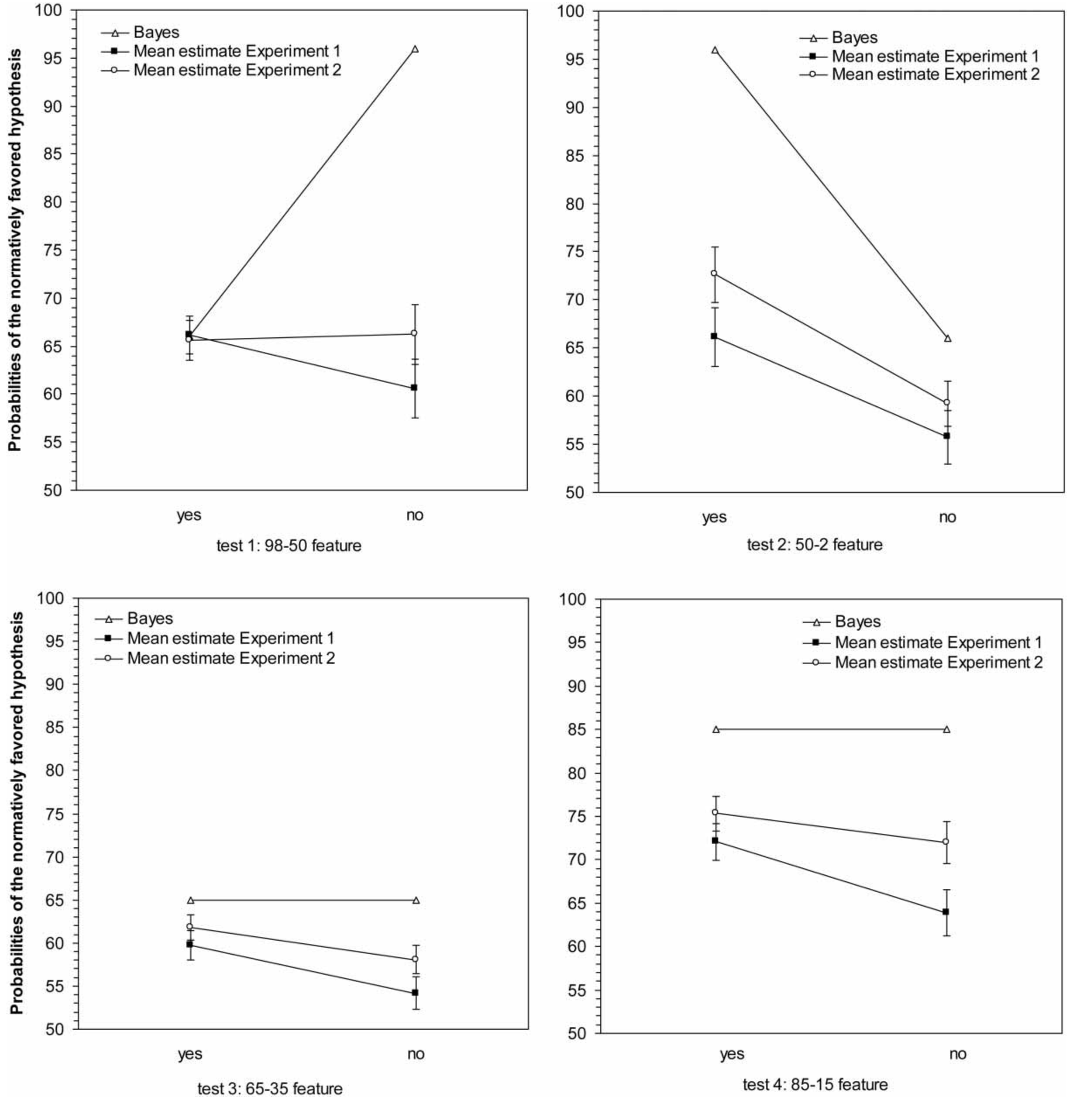


Figure 3. *Probabilities (in the form of percentages) of the hypothesis favoured by the evidence. The participants' mean estimates are compared with the Bayesian responses. For participants' mean estimates, standard error of the mean (*SEM*) bars are also shown.*

diagnostic—that is, when they come from a symmetric question.

For the asymmetric tests, the estimated difference between the diagnosticities of "yes" and "no" answers was underestimated compared to the Bayesian difference. That is, participants were insufficiently sensitive to differentially diagnostic answers. As in Slowiaczek et al. (1992), we found a different magnitude of such insensitivity as a function of the type of asymmetric question from which the answers came. In particular, as in Slowiaczek et al.'s (1992) study, participants in our study appreciated less the difference in diagnosticity between the "yes" and "no" answers for the asymmetrically disconfirming question (98–50 combination, Figure 3, top left panel) than for the asymmetrically confirming question (50–2 combination, Figure 3, top right panel).

Downloaded by [Universita' Milano Bicocca] at 03:17 17 December 2013

The greater insensitivity to answer diagnosticity for the asymmetrically disconfirming test is apparent because participants tended to exhibit an asymmetry in the counternormative direction (see Figure 3, top left panel). As for symmetric tests, participants perceived a difference in informativeness between the "yes" and the "no" answers when actually they were equally diagnostic. In particular, they perceived "yes" answers as more diagnostic than "no" answers.

Both oversensitivity to answer diagnosticity, when the answers came from symmetric queries (Tests 3–4), and the different magnitude of insensitivity found for asymmetric questions (Tests 1–2) contradict Slowiaczek et al.'s (1992) explanation of insensitivity to answer diagnosticity. According to Slowiaczek et al. (1992), people perceive the "yes" and "no" answers as more similar in diagnosticity than they actually are because the difference between the constituents of the LR is the same for both answers. However, the difference was identical in the two asymmetric tests that we used. In particular, the difference was 48% for both "yes" and "no" answers for both asymmetric tests. Furthermore, in the case of each symmetric test, the difference between the constituent probabilities of the LRs was the same for "yes" and "no" answers. Specifically, the difference was 30% for Test 3 and 70% for Test 4. If people subtract the percentages that constitute the LR when evaluating answer diagnosticity, then participants in our study would have noticed that not only the percentage differences but also the percentages by themselves were identical for the "yes" and "no" answers in the case of symmetric tests. Thus, they should have valued equally the "yes" and "no" answers to symmetric tests.

The main effect of test type revealed that participants weighed most the answers following the most diagnostic question (Test 4, 85–15 combination), and they weighed least the answers following the least diagnostic query (Test 3, 65–35 combination), while the answers to asymmetric tests (Tests 1 and 2) fell in between.[3] This finding indicates that participants were sensitive to the diagnosticity of the questions. Interestingly, the metrics proposed in the literature for computing the diagnosticity of a question are based on the expected outcomes—that is, they are weighted averages of the diagnosticities of the "yes" and "no" answers to that question (see e.g., Nelson, 2005, Appendix A; Nickerson, 1996).[4] Participants might have perceived that the answers following a question were on average more, less, or equally diagnostic compared to the answers following another question, but they failed to perceive the relative weight of the specific answers ("yes" versus "no") following a question.

For each question, participants were affected more by "yes" answers than by "no" answers. This finding is revealed by the significant main effect of answer and is evident in Figure 3. In other words, participants seemed as though they were influenced by a form of the feature-positive effect, whereby people overweigh the presence, as opposed to the absence, of features (e.g., Cherubini et al., 2013; Jenkins & Sainsbury, 1969, 1970; Newman, Wolff, & Hearst, 1980; Rusconi, Crippa, et al., 2012; Rusconi et al., 2013). This finding is in

[3] The diagnosticity of the questions that we used in our task followed this decreasing order: Test 4 (85–15 combination) > Test 1 (98–50 combination) = Test 2 (50–2 combination) > Test 3 (65–35 combination). This order is derived both from the application of the feature-difference heuristic and from the calculation of question diagnosticity according to several norms (namely $\log_{10}$-diagnosticity, information gain, Kullback–Leibler distance, probability gain, and impact, see e.g., Nelson, 2005, 2008). However, an exception is the expected maximum LR (e.g., Good, 1950), often called Bayesian diagnosticity (e.g., Nelson, 2005, 2008, 2009), according to which the least useful question is Test 3 (65–35 combination), as for the other norms, but the most useful queries are the asymmetric tests (Tests 1–2) not Test 4 (85–15 combination).

[4] The idea of selecting questions to ask based on the weighted average of the utilities of the possible outcomes was suggested by Alan Turing (1912–1954) in 1940 (cited in Good & Card, 1971, p. 182) and then by Good (1950). In a similar way, Savage (1954) proposed to use expected subjective utility to select questions. One exception to this approach based on the weighted average of the expected outcomes' utility is presented in Martignon, Katsikopoulos, and Woike (2008). These authors proposed to use "fast and frugal trees" for categorization in contexts with limited resources. These trees are heuristics that bypass the computation of probabilities and define the diagnosticity of a question on the basis of the best possible outcome—that is, the possible outcome that would be the most useful.

Downloaded by [Universita' Milano Bicocca] at 03:17 17 December 2013

keeping with Slowiaczek et al.'s (1992) study, in which the authors found that participants weighted the "yes" answers more than the "no" answers regardless of their actual informativeness in Experiments 1A, 2B, and 2C.

EXPERIMENT 2

Experiment 2 addressed the issue of a possible effect of the format of the presented information on participants' estimates after a "yes" and after a "no". In Bayesian terms, the ability to differentiate (or equate, in case of symmetric tests) the "yes" and the "no" answers in terms of their different (equal) informativeness entails the computation of both the LR for the "yes" answer and that for the "no" answer. People might encounter more difficulties in considering the LR for "no" than for "yes" when evaluating answers because of the well-known difficulty to process negative information relative to positive information (e.g., Cherubini et al., 2013; Hearst, 1991; Van Wallendael, 1995; Wason, 1959, 1961). Participants in Experiment 1 had to calculate from the presented percentages their complements to compute the LRs for "no" answers. This required a further step of processing compared to the evaluation of the impact of "yes" answers. Accordingly, we hypothesized that by adding the probabilities of the absence of the features to the probabilities of their presence, participants would be more sensitive to the actual informativeness of "yes" and "no" answers. We tested this hypothesis in Experiment 2. Formally speaking, participants in Experiment 1 were presented with $p(D|\mathrm{H})$, where "H" stands for both the hypotheses (i.e., both Gloms and Fizos), while in Experiment 2 they received both $p(D|\mathrm{H})$ and $p(\neg D|H)$ (see Appendix A; this procedure was drawn from Cherubini et al., 2013; see also Rusconi, Crippa, et al., 2012, Study 3).

Method

Participants

Ninety-four undergraduate students at the University of California, San Diego (68% female, mean age 20.2 years, range 17–28 years) took part in the experiment in exchange for course credit.

Materials and procedure

Design, materials, instructions, and procedure were exactly the same as those in Experiment 1, with the exception of the addition of the probabilities of the absence of the features beside the probabilities of their presence (see Appendix A). For instance, when presenting to participants Test 2, we gave them both the 50–2 percentage combination, indicating the probabilities of the presence of the feature (i.e., drinking gasoline) in Gloms and Fizos, and its complement, the 50–98 combination, indicating the probabilities of the absence of the same feature in the two groups.

Results

We used the same recoding of participants' estimates as that used in Experiment 1 (see also McKenzie, 2006; Slowiaczek et al., 1992). We subjected the recoded estimates to the same 4 × 2 × 2 × 2 mixed design ANOVA as that run in Experiment 1. We remind the reader that, according to Bayes' theorem, there should be a significant main effect of test type, but not of answer, and there should be a significant Test Type × Answer interaction. As in Experiment 1, we report the results concerning these effects here, while we present the description of the other findings of the ANOVA in Appendix C. We found a significant main effect of test type, $F(1,\ 90) = 23.01$, $MSE = 784.58$, $p < .001$, $\eta^2 = .053$, lower bound correction, showing that participants' estimates were higher for Test 4 ($M = 73.7$, $SDE = 1.5$), than for all other test types, all $p\mathrm{s} < .001$. Furthermore, there was not a significant difference in the estimates after receiving the answers to asymmetric tests—that is, the answers to Test 1 ($M = 65.7$, $SDE = 1.5$) and the answers to Test 2 ($M = 66.1$, $SDE = 1.6$), $p = .819$. Finally, participants provided significantly lower estimates for Test 3 ($M = 59.9$, $SDE = 0.8$) than for all other test types, all $p\mathrm{s} < .001$. Although the pattern is similar to the one found in Experiment 1, the estimates are overall higher, as shown in Figure 3.

Downloaded by [Universita' Milano Bicocca] at 03:17 17 December 2013

Contrary to the normative prediction, and as in Experiment 1, there was a significant, although weaker, main effect of answer, $F(1, 90) = 5.52$, $MSE = 947.87$, $p = .021$, $\eta^2 = .015$, lower bound correction, reflecting greater weight assigned to the "yes" answer ($M = 69$, $SDE = 1.4$) than to the "no" answer ($M = 63.7$, $SDE = 1.6$).

Crucially, and contrary to Experiment 1, there was a Test Type × Answer interaction, $F(1, 90) = 8.05$, $MSE = 633.30$, $p = .006$, $\eta^2 = .015$. Pairwise comparisons showed that only for Test 2 (50–2 combination) was there a significant difference between the estimates after a "yes" ($M = 73$, $SDE = 2.1$) and the estimates after a "no" ($M = 59.2$, $SDE = 2.3$), $p < .001$. In contrast, Test 1 (98–50 combination) was perceived as though it was symmetric, because the difference between the estimates after the "yes" ($M = 65.5$, $SDE = 2.1$) and after the "no" ($M = 65.8$, $SDE = 2.3$) was not significant, $p = .917$. As for Test 3 (65–35 combination), the difference between participants' estimates after a "yes" ($M = 61.9$, $SDE = 1.3$) and after a "no" ($M = 57.9$, $SDE = 1.5$) did not reach statistical significance, $p = .099$. Also for the other symmetric test (85–15 combination) the difference between the estimates after a "yes" ($M = 75.5$, $SDE = 1.8$) and after a "no" ($M = 71.9$, $SDE = 2.3$) was not significant, $p = .233$. It should be noted that, overall, participants' responses in this experiment are closer to Bayesian responses than they are in Experiment 1 (see Figure 3).

Following the same procedure as that used in Experiment 1, we performed a series of one-sample *t* tests to compare the Bayesian and the observed difference between probabilities after a "yes" and probabilities after a "no" for each test type. For asymmetric tests, we found again an insensitivity to answer diagnosticity, although weaker than in Experiment 1. Specifically, the mean difference of –0.6 ($SD = 33.5$) between estimates after the "yes" and estimates after the "no" following the asymmetrically disconfirming question (98–50 feature) was significantly less than the normative difference of –30, $t(93) = 8.51$, $p < .001$, $d = 0.88$. Nonetheless, compared to Experiment 1, participants tended to value the diagnosticity of the "no" answer more appropriately (Figure 3, top left panel). Furthermore, the mean perceived difference of 13.4 ($SD = 36.3$) when the question was asymmetrically confirming (50–2 feature) was significantly less than the normative difference of 30, $t(93) = -4.44$, $p < .001$, $d = 0.46$ (Figure 3, top right panel).

Contrary to Experiment 1, we found that the difference between participants' estimates after the "yes" and after the "no" answers for symmetric tests did not differ significantly from the normatively expected null difference. In particular, the mean difference between the estimates after a "yes" and those after a "no" for Test 3 (65–35 combination; $M = 3.8$, $SD = 26.9$) was not significantly different from the normatively expected difference of 0, $t(93) = 1.36$, $p = .178$, $d = 0.14$ (Figure 3, bottom left panel). In a similar vein, for Test 4 (85–15 percentage combination), the mean difference between the estimates after the "yes" and the estimates after the "no" ($M = 3.4$, $SD = 31$) was not significantly different from the Bayesian null difference, $t(93) = 1.05$, $p = .298$, $d = 0.11$ (Figure 3, bottom right panel).

Discussion

Experiment 2 indicated that participants tended to exhibit a symmetry of estimates when evaluating "yes" and "no" answers to symmetric tests (Tests 3 and 4). Indeed, both the pairwise comparisons and the follow-up *t* tests revealed that the difference in the estimates after the "yes" and "no" answers to symmetric tests was not significantly different from the normatively expected null difference. The answers to the asymmetrically disconfirming query (about the 98–50 feature) were perceived as equally diagnostic, too, but there was a slight tendency to perceive a greater weight of the "no" versus the "yes", in line with the normative direction of the asymmetry (see Figure 3, top left panel). Furthermore, participants perceived differently the "yes" and "no" answers to the asymmetrically confirming question (50–2 percentage combination).

Overall, participants benefited from the manipulation of the presentation format of the probabilistic

Downloaded by [Universita' Milano Bicocca] at 03:17 17 December 2013

information that they received. Presenting to participants both the percentages that are useful to compute the LR for the "yes" answers and their complements, which are used to compute the LR for the "no" answers, had the effect of sensitizing them to answer diagnosticity. Indeed, contrary to Experiment 1, the Test Type × Answer interaction was significant, indicating that participants did not value the diagnosticities of "yes" and "no" answers in the same way regardless of the question from which the answers came.

As in Experiment 1, participants provided the highest estimates after receiving the answers to the most diagnostic question (Test 4, 85–15 combination), and the lowest estimates when the answers came from the least diagnostic query (Test 3, 65–35 combination), while the estimates after the receipt of the answers to asymmetric tests (Tests 1 and 2) fell in between. The significant main effect of test type revealed this finding. However, participants' sensitivity to the differential diagnosticity of questions leaves unresolved why they are insensitive to answer diagnosticity. Indeed, the different magnitude of insensitivity to answer diagnosticity found for the asymmetrically confirming versus the asymmetrically disconfirming test runs counter to Slowiaczek et al.'s (1992) explanation based on people's confusion of answer diagnosticity assessment with question diagnosticity assessment. In fact, for both types of asymmetric tests the difference in the percentages that constitute the LR was 48% for both "yes" and "no" answers, and the two asymmetric questions were equally diagnostic regardless of the specific metric used to assess question usefulness. Furthermore, according to this account, presenting both the percentages that are useful to compute the LR for the "yes" answer and those relative to the "no" answer should have decreased, instead of increased, participants' sensitivity to differential answer diagnosticity when the answers came from asymmetric tests because participants could straightforwardly determine that the differences in the LR constituents were identical for the "yes" and "no" answers.

In contrast, what seemed to underlie participant residual insensitivity to answer diagnosticity was the tendency to weigh more the "yes" answers than the "no" answers. This interpretation is supported by the significant main effect of answer (see also Figure 3). Nonetheless, this effect was weaker than in Experiment 1, probably reflecting the debiasing effect of the manipulation of the presentation format of the percentage combinations.

GENERAL DISCUSSION

Two experiments showed that people can value two answers ("yes" and "no") to the same question as differentially informative although they are equally diagnostic (and equally frequent, see Figure 1) from a normative (Bayesian) standpoint. That is, we provided evidence for people's *over*sensitivity to answer diagnosticity when the answers come from a symmetric question. We also provided evidence for the reliability of the *in*sensitivity to answer diagnosticity found in previous similar studies. Indeed, when the "yes" and "no" answers to a question convey a different amount of information (i.e., when the question is asymmetric), people tend to perceive the two answers as more similar in terms of their diagnosticity than they actually are. However, both the findings of previous studies (see Figure 2) and the data presented in this article (see Figure 3) indicate that insensitivity to answer diagnosticity varies as a function of the type of question.

In particular, we found greater insensitivity for the asymmetrically disconfirming question (98–50 combination) than for the asymmetrically confirming question (50-2 combination). That is, participants had more difficulties in perceiving that a "no" answer was more informative than a "yes" answer (as it is the case for the 98–50 test) than vice versa (as it is in the case for the 50–2 combination). This finding cannot be explained in terms of participants' confusion of the assessment of answer diagnosticity with the assessment of question usefulness (Slowiaczek et al., 1992). According to this explanation, people perceive two answers as more similar than they actually are because the difference between the probabilities that constitute the LR is the same for both "yes"

Downloaded by [Universita' Milano Bicocca] at 03:17 17 December 2013

and "no" answers. However, in our experiments the difference between the percentages was 48% for both "yes" and "no" answers regardless of whether the question was asymmetrically confirming or asymmetrically disconfirming. In fact, the two asymmetric tests were equally diagnostic. Rather, the greater difficulty in belief updating in light of the answers to the asymmetrically disconfirming question reveals people's tendency to overweigh the evidential strength of "yes" versus "no" answers. We found this form of the feature-positive effect (e.g., Cherubini et al., 2013; Jenkins & Sainsbury, 1969, 1970; Newman et al., 1980; Rusconi, Crippa, et al., 2012; Rusconi et al., 2013) in both experiments, as revealed by the significant main effects of answer. This finding is in keeping with previous similar experiments (Slowiaczek et al., 1992, Experiments 1A, 2B, and 2C).

Taking into account whether the question asked is asymmetrically confirming or asymmetrically disconfirming when investigating answer diagnosticity assessment has implications for confirmation bias —that is, the tendency to apportion unwarranted confidence to the focal hypothesis (e.g., McKenzie, 2004, 2006). The current knowledge about confirmation bias indicates that it originates from a combination of biases at the testing stage and biases at the evaluation stage of hypothesis development (e.g., Klayman, 1995; McKenzie 2004, 2006; Poletiek, 2001). It has been argued that one of these combinations is the preference for asking asymmetrically disconfirming questions and a failure to perceive that the hypothesis-disconfirming answer to this kind of question is more diagnostic than the hypothesis-confirming answer (e.g., Klayman, 1995; McKenzie, 2004, 2006; Slowiaczek et al., 1992). In our experiments, participants failed to perceive the greater diagnosticity of the disconfirming "no" answer than of the confirming "yes" answer to Test 1 (98–50 combination). Therefore, we found corroborating evidence for the bias in the evaluation part of this testing/evaluation combination.

Some authors have argued that confirmation bias can originate from a preference for asking asymmetrically confirming questions (e.g., Cameron & Trope, 2004; Poletiek & Berndsen, 2000; Trope & Thompson, 1997). For example, Trope and Thompson state: "The testing strategy becomes biased in favor of the hypothesis when the questions are asymmetric, namely, when hypothesis-consistent answers are more diagnostic than hypothesis-inconsistent answers" (Trope & Thompson, 1997, p. 240)—that is, when the questions are asymmetrically confirming. Note that, from a Bayesian perspective, a highly diagnostic outcome is rare whenever the prior probabilities of the hypotheses being considered are equal (see Figure 1, also see the diagnosticity/frequency trade-off, e.g., McKenzie, 2006; Poletiek, 2001, chapters 1 and 2; Poletiek & Berndsen, 2000; Rusconi, Sacchi, Toscano, & Cherubini, 2012; Sacchi, Rusconi, Bonomi, & Cherubini, 2013). Accordingly, an asymmetrically confirming query cannot foster confirmation bias because the hypothesis-confirming answer is more rare than the hypothesis-disconfirming answer although it is more diagnostic. However, according to these authors, when tasks activate strong a priori beliefs (e.g., stereotypes) this diagnosticity/frequency trade-off might not occur because "strong category-based expectancies increase the subjective likelihood that the target will provide the more diagnostic expectancy-consistent answer rather than the less diagnostic expectancy-inconsistent answer" (Trope & Thompson, 1997, p. 230). In other words, people would perceive that a hypothesis-confirming answer to an asymmetrically confirming question is both highly diagnostic and highly likely, and thus confirmation bias would be possible. The insensitivity that we found for the asymmetrically confirming query (Test 2, 50–2 combination), if replicated under the circumstances explained above, would weaken confirmation bias. Indeed, participants in our experiments did not value the hypothesis-confirming "yes" answer as more diagnostic than the hypothesis-disconfirming "no" answer as much as we would expect based on normative considerations.

Hence, while insensitivity to answer diagnosticity might favour confirmation bias when combined with an asymmetrically disconfirming testing strategy, it might have a debiasing effect

Downloaded by [Universita' Milano Bicocca] at 03:17 17 December 2013

in some circumstances (e.g., intergroup contexts) when it is combined with an asymmetrically confirming testing strategy.

Slowiaczek et al. (1992) argued that "symmetrical questions (70–30, 20–80) are not prone to the inferential errors we document, because 'yes' and 'no' answers are equally diagnostic" (Slowiaczek et al., 1992, p. 402). The authors probably referred to the errors due to the combination of asymmetrically disconfirming testing and insensitivity to answer diagnosticity described above. However, another combination that might lead to confirmation is positive testing (asking questions about features that are more likely to occur if the focal hypothesis is true than if it is false) and the feature-positive effect in the evaluation stage of hypothesis development (e.g., Klayman, 1995; McKenzie, 2004, 2006). We found evidence for the feature-positive effect because people weighed more the "yes" answers than the "no" answers to symmetric queries in Experiment 1. Therefore, oversensitivity to answer diagnosticity (in the form of assigning more weight than warranted to "yes" than to "no" answers) combined with a preference for asking symmetric positive questions can also lead to confirmation bias.

Our experiments also add to the literature by showing that insensitivity and oversensitivity to answer diagnosticity in abstract tasks of hypothesis testing is moderated by the presentation format of the percentages given to participants. The feature-positive effect found in Experiment 1 suggests that participants might have difficulty in assessing the diagnosticity of "no" answers. Indeed, "no" answers require a further step of processing compared to "yes" answers: People have to calculate the complements of the percentages that constitute the LR for "yes" answers. In Experiment 2, we presented to participants the percentages needed to compute the LR for "yes" answers along with their complements. This manipulation had a debiasing effect. Participants were still affected by the feature-positive effect, but to a lesser extent than in Experiment 1. Contrary to Experiment 1, the Test Type × Answer interaction was significant, indicating that participants appreciated that the relative weight of "yes" and "no" answers differed as a function of the type of question. In fact, contrary to Experiment 1, participants exhibited the symmetry of estimates that we would expect for normative reasons in the case of symmetric tests (65–35 and 85–15 features). For the asymmetrically disconfirming question (98–50 feature), there was a slight tendency to perceive the greater informativeness of the "no" answer than the "yes" answer. Finally, for the asymmetrically confirming question (50–2 feature), the insensitivity was less pronounced than in Experiment 1. These findings suggest that people's difficulty to revise in a Bayesian way their initial beliefs in light of different answers to the same question might reside, at least in part, in the failure to infer correctly the likelihoods of the nonoccurrence of the features inquired about.

More generally, the present data suggest that people's non-Bayesian use of the obtained evidence in this kind of task might be due to two reasons. First, overall, participants' revisions in light of the received answers appeared to be conservative (e.g., Edwards, 1968; Phillips & Edwards, 1966)—that is, their posterior probability estimates were closer to the prior probability of .5 than normatively expected (see Figure 3). Second, participants found the "yes" answers more informative than the "no" answers—that is, they exhibited a feature-positive effect (e.g., Cherubini et al., 2013; Jenkins & Sainsbury, 1969, 1970; Newman et al., 1980; Rusconi, Crippa, et al., 2012; Rusconi et al., 2013).

Newman et al. (1980) proposed that the feature-positive effect might have evolved because occurrences of natural events are relatively rare and thus more informative than nonoccurrences of events. Accordingly, one can hypothesize that participants in our experiments had difficulties in processing the words-and-numbers scenarios, and thus they might have exploited their knowledge about real-world relationships. If this is the case, they might have assumed that feature presences (i.e., "yes" answers) were rare and thus more informative than feature absences (i.e., "no" answers; e.g., McKenzie & Chase, 2012; McKenzie & Mikkelsen, 2000, 2007; for a discussion of this argument see McKenzie, 2006, p. 580).

Downloaded by [Universita' Milano Bicocca] at 03:17 17 December 2013

In conclusion, we showed that people's insensitivity to answer diagnosticity is more malleable than previously argued (e.g., Slowiaczek et al., 1992). Taking into account the type of question from which the answers come revealed that people exhibit not only insensitivity but also oversensitivity to answer diagnosticity. Furthermore, insensitivity to answer diagnosticity per se is a more nuanced phenomenon than previously thought. Indeed, people are less sensitive when the normative diagnosticity of "no" answers is higher than that of "yes" answers than vice versa. Finally, the way in which the relevant information is presented to participants might enhance their sensitivity to the amount of information that the answers convey. Further empirical investigations are in need to elucidate the mechanisms underlying evidence evaluation in hypothesis testing. In the current experiments, for example, we used an unfamiliar scenario, and previous studies have shown that sensitivity to answer diagnosticity increases when familiar materials are used (McKenzie, 2006). However, biases observed using abstract materials did not disappear or reverse with concrete materials (McKenzie, 2006). Thus, results using abstract materials might provide a useful guide of what to expect when using concrete materials. Furthermore, they suggest possible ways to reduce inefficiencies in evidence evaluation in the specific situations in which previous knowledge does not play a key role, and contextual cues are limited and uncertain, as it might be when professionals have to evaluate statistical write-ups.

In the present experiments, we used words-and-numbers scenarios. This is not the information format that most enhances Bayesian reasoning (e.g., Cosmides & Tooby, 1996; Gigerenzer & Hoffrage, 1995; Meder & Nelson, 2012; Nelson, 2009). Therefore, it remains to be investigated whether other formats, such as icon arrays (e.g., Galesic, Garcia-Retamero, & Gigerenzer, 2009), or allowing participants to learn the environmental probabilities through experience (e.g., Meder & Nelson, 2012; Nelson, 2009; Nelson et al., 2010) have a greater debiasing effect. Finally, future studies might consider individual differences in insensitivity and oversensitivity to answer diagnosticity. In particular, it would be interesting to investigate whether statistical numeracy (assessed, for example, through the Berlin Numeracy Test introduced by Cokely, Galesic, Schulz, Ghazal, & Garcia-Retamero, 2012) moderates such phenomena.

Original manuscript received 10 August 2012
Accepted revision received 28 March 2013
First published online 16 May 2013

REFERENCES

Beach, L. R. (1968). Probability magnitudes and conservative revision of subjective probabilities. *Journal of Experimental Psychology*, *77*, 57–63. doi:10.1037/h0025800

Beyth-Marom, R., & Fischhoff, B. (1983). Diagnosticity and pseudodiagnosticity. *Journal of Personality and Social Psychology*, *45*, 1185–1195. doi:10.1037//0022-3514.45.6.1185

Brambilla, M., Rusconi, P., Sacchi, S., & Cherubini, P. (2011). Looking for honesty: The primary role of morality (vs. sociability and competence) in information gathering. *European Journal of Social Psychology*, *41*, 135–143. doi:10.1002/ejsp.744

Cameron, J. A., & Trope, Y. (2004). Stereotype-biased search and processing of information about group members. *Social Cognition*, *22*, 650–672. doi:10.1521/soco.22.6.650.54818

Casscells, W., Schoenberger, A., & Graboys, T. B. (1978). Interpretation by physicians of clinical laboratory results. *New England Journal of Medicine*, *299*, 999–1001. doi:10.1056/NEJM197811022991808

Cherubini, P., Rusconi, P., Russo, S., & Crippa, F. (2013). Missing the dog that failed to bark in the nighttime: On the overestimation of occurrences over non-occurrences in hypothesis testing. *Psychological Research*, *77*, 348–370. doi:10.1007/s00426-012-0430-3

Cherubini, P., Rusconi, P., Russo, S., Di Bari, S., & Sacchi, S. (2010). Preferences for different questions when testing hypotheses in an abstract task: Positivity does play a role, asymmetry does not. *Acta Psychologica*, *134*, 162–174. doi:10.1016/j.actpsy.2010.01.007

Christensen-Szalanski, J. J. J., & Bushyhead, J. B. (1981). Physicians' use of probabilistic information in a real clinical setting. *Journal of Experimental*

Psychology: Human Perception and Performance, *7,* 928–935. doi:10.1037//0096-1523.7.4.928

Clark, S. E., & Wells, G. L. (2008). On the diagnosticity of multiple-witness identifications. *Law and Human Behavior*, *32,* 406–422. doi:10.1007/s10979-007-9115-7

Cokely, E. T., Galesic, M., Schulz, E., Ghazal, S., & Garcia-Retamero, R. (2012). Measuring risk literacy: The Berlin Numeracy Test. *Judgment and Decision Making*, *7,* 25–47.

Cosmides, L., & Tooby, J. (1996). Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. *Cognition*, *58,* 1–73. doi:10.1016/0010-0277(95)00664-8

Crupi, V., Tentori, K., & Gonzalez, M. (2007). On Bayesian measures of evidential support: Theoretical and empirical issues. *Philosophy of Science*, *74,* 229–252. doi:10.1086/520779

Edwards, W. (1968). Conservatism in human information processing. In B. Kleinmuntz (Ed.), *Formal representation of human judgment* (pp. 17–52). New York, NY: Wiley.

Evett, S. R., Devine, P. G., Hirt, E. R., & Price, J. (1994). The role of the hypothesis and the evidence in the trait hypothesis testing process. *Journal of Experimental Social Psychology*, *30,* 456–481. doi:10.1006/jesp.1994.1022

Fiedler, K., & Walther, E. (2004). *Stereotyping as inductive hypothesis testing*. Hove, UK: Psychology Press.

Fischhoff, B., & Beyth-Marom, R. (1983). Hypothesis evaluation from a Bayesian perspective. *Psychological Review*, *90,* 239–260. doi:10.1037//0033-295X.90.3.239

Galesic, M., Garcia-Retamero, R., & Gigerenzer, G. (2009). Using icon arrays to communicate medical risks: Overcoming low numeracy. *Health Psychology*, *28,* 210–216. doi:10.1037/a0014474

Garcia-Marques, L., Sherman, S. J., & Palma-Oliveira, J. M. (2001). Hypothesis testing and the perception of diagnosticity. *Journal of Experimental Social Psychology*, *37,* 183–200. doi:10.1006/jesp.2000.1441

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review*, *102,* 684–704. doi:10.1037//0033-295X.102.4.684

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

Good, I. J. (1979). Studies in the history of probability and statistics: XXXVII. A. M. Turing's statistical work in World War II. *Biometrika*, *66,* 393–396. doi:10.1093/biomet/66.2.393

Good, I. J., & Card, W. I. (1971). The diagnostic process with special reference to errors. *Methods of Information in Medicine*, *10,* 176–188.

Hammerton, M. (1973). A case of radical probability estimation. *Journal of Experimental Psychology*, *101,* 252–254. doi:10.1037/h0035224

Hearst, E. (1991). Psychology and nothing. *American Scientist*, *79,* 432–443.

Hemsley, D. R., & Garety, P. A. (1986). The formation of maintenance of delusions: A Bayesian analysis. *British Journal of Psychiatry*, *149,* 51–56. doi:10.1192/bjp.149.1.51

Hoffrage, U., & Gigerenzer, G. (1998). Using natural frequencies to improve diagnostic inferences. *Academic Medicine*, *73,* 538–540. doi:10.1097/00001888-199805000-00024

Jenkins, H. M., & Sainsbury, R. S. (1969). The development of stimulus control through differential reinforcement. In N. J. Mackintosh & W. K. Honig (Eds.), *Fundamental issues in associative learning* (pp. 123–161). Halifax: Dalhousie University Press.

Jenkins, H. M., & Sainsbury, R. S. (1970). Discrimination learning with the distinctive feature on positive or negative trials. In D. Mostofsky (Ed.), *Attention: Contemporary theory and analysis* (pp. 239–275). New York, NY: Appleton-Century-Crofts.

Jenkins, H. M., & Ward, W. C. (1965). Judgment of contingency between responses and outcomes. *Psychological Monographs: General and Applied*, *79,* 1–17. doi:10.1037/h0093874

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, *3,* 430–454. doi:10.1016/0010-0285(72)90016-3

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, *80,* 237–251. doi:10.1037/h0034747

Klayman, J. (1995). Varieties of confirmation bias. *Psychology of Learning and Motivation*, *32*, 385–418. doi:10.1016/S0079-7421(08)60315-1

Martignon, L., Katsikopoulos, K. V., & Woike, J. K. (2008). Categorization with limited resources: A family of simple heuristics. *Journal of Mathematical Psychology*, *52,* 352–361. doi:10.1016/j.jmp.2008.04.003

Mastropasqua, T., Crupi, V., & Tentori, K. (2010). Broadening the study of inductive reasoning: Confirmation judgments with uncertain evidence.

Downloaded by [Universita' Milano Bicocca] at 03:17 17 December 2013

Downloaded by [Universita' Milano Bicocca] at 03:17 17 December 2013

Memory & Cognition, *38,* 941–950. doi:10.3758/MC.38.7.941

McKenzie, C. R. M. (1994). The accuracy of intuitive judgment strategies: Covariation assessment and Bayesian inference. *Cognitive Psychology*, *26,* 209–239. doi:10.1006/cogp.1994.1007

McKenzie, C. R. M. (2004). Hypothesis testing and evaluation. In D. J. Koehler & N. Harvey (Eds.), *Blackwell handbook of judgment and decision making* (pp. 200–219). Oxford: Blackwell.

McKenzie, C. R. M. (2006). Increased sensitivity to differentially diagnostic answers using familiar materials: Implications for confirmation bias. *Memory & Cognition*, *34,* 577–588. doi:10.3758/BF03193581

McKenzie, C. R. M., & Chase, V. M. (2012). Why rare things are precious: The importance of rarity in lay inference, In P. M. Todd, G. Gigerenzer, & The ABC Research Group (Eds.), *Ecological rationality: Intelligence in the world* (pp. 309–334). Oxford: Oxford University Press

McKenzie, C. R. M., & Mikkelsen, L. A. (2000). The psychological side of Hempel's paradox of confirmation. *Psychonomic Bulletin & Review*, *7,* 360–366. doi:10.3758/BF03212994

McKenzie, C. R. M., & Mikkelsen, L. A. (2007). A Bayesian view of covariation assessment. *Cognitive Psychology*, *54,* 33–61. doi:10.1016/j.cogpsych.2006.04.004

Meder, B., & Nelson, J. D. (2012). Information search with situation-specific reward functions. *Judgment and Decision Making*, *7,* 119–148.

Meier, K. M., & Blair, M. R. (2013). Waiting and weighting: Information sampling is a balance between efficiency and error-reduction. *Cognition*, *126,* 319–325. http://dx.doi.org/10.1016/j.cognition.2012.09.014

Nelson, J. D. (2005). Finding useful questions: On Bayesian diagnosticity, probability, impact, and information gain. *Psychological Review*, *112,* 979–999. doi:10.1037/0033-295X.112.4.979

Nelson, J. D. (2008). Towards a rational theory of human information acquisition. In N. Chater & M. Oaksford (Eds.), *The probabilistic mind* (pp. 143–163). Oxford: Oxford University Press.

Nelson, J. D. (2009). Naïve optimality: Subjects' heuristics can be better motivated than experimenters' optimal models. *Behavioral and Brain Sciences*, *32,* 94–95. doi:10.1017/S0140525X09000405

Nelson, J. D., McKenzie, C. R. M., Cottrell, G. W., & Sejnowski, T. J. (2010). Experience matters: Information acquisition optimizes probability gain. *Psychological Science*, *21,* 960–969. doi:10.1177/0956797610372637

Newman, J., Wolff, W. T., & Hearst, E. (1980). The feature-positive effect in adult human subjects. *Journal of Experimental Psychology: Human Learning and Memory*, *6,* 630–650. doi:10.1037/0278-7393.6.5.630

Nickerson, R. S. (1996). Hempel's paradox and Wason's selection task: Logical and psychological puzzles of confirmation. *Thinking & Reasoning*, *2,* 1–31. doi:10.1080/135467896394546

Phillips, L. D., & Edwards, W. (1966). Conservatism in a simple probability inference task. *Journal of Experimental Psychology*, *72,* 346–354. doi:10.1037/h0023653

Poletiek, F. H. (2001). *Hypothesis-testing behaviour*. Hove, UK: Psychology Press.

Poletiek, F. H., & Berndsen, M. (2000). Hypothesis testing as risk behaviour with regard to beliefs. *Journal of Behavioral Decision Making*, *13,* 107–123. doi:10.1002/(SICI)1099-0771(200001/03)13:1<107::AID-BDM349>3.0.CO;2-P

Rusconi, P., Crippa, F., Russo, S., & Cherubini, P. (2012). Moderators of the feature-positive effect in abstract hypothesis-evaluation tasks. *Canadian Journal of Experimental Psychology*, *66,* 181–192. doi:10.1037/a0028173

Rusconi, P., Marelli, M., Russo, S., D'Addario, M., & Cherubini, P. (2013). Integration of base rates and new information in an abstract hypothesis-testing task. *British Journal of Psychology*, *104*, 193–211. doi:10.1111/j2044-8295.2012.02112.x

Rusconi, P., Sacchi, S., Toscano, A., & Cherubini, P. (2012). Confirming expectations in asymmetric and symmetric social hypothesis testing. *Experimental Psychology*, *59,* 243–250. doi:10.1027/1618-3169/a000149

Sacchi, S., Rusconi, P., Bonomi, M., & Cherubini, P. (2013). Effects of asymmetric questions on impression formation: A trade-off between evidence diagnosticity and frequency. *Social Psychology*. Advance online publication. doi:10.1027/1864-9335/a000158

Sacchi, S., Rusconi, P., Russo, S., Bettiga, R., & Cherubini, P. (2012). New knowledge for old credences: Asymmetric information search about in-group and out-group members. *British Journal of Social Psychology*, *51,* 606–625. doi:10.1111/j.2044-8309.2011.02026.x

Savage, L. J. (1954). *The foundations of statistics*. New York, NY: John Wiley & Sons.

Downloaded by [Universita' Milano Bicocca] at 03:17 17 December 2013

Schum, D. A. (1994). *The evidential foundations of probabilistic reasoning*. New York, NY: John Wiley & Sons.

Skov, R. B., & Sherman, S. J. (1986). Information-gathering processes: Diagnosticity, hypothesis-confirmatory strategies, and perceived hypothesis confirmation. *Journal of Experimental Social Psychology*, *22,* 93–121. doi:10.1016/0022-1031(86)90031-4

Slovic, P., & Lichtenstein, S. (1971). Comparison of Bayesian and regression approaches to the study of information processing in judgment. *Organizational Behavior and Human Performance*, *6*, 649–744. doi:10.1016/0030-5073(71)90033-X

Slowiaczek, L. M., Klayman, J., Sherman, S. J., & Skov, R. B. (1992). Information selection and use in hypothesis testing: What is a good question, and what is a good answer? *Memory & Cognition*, *20,* 392–405.

Tentori, K., Crupi, V., Bonini, N., & Osherson, D. (2007). Comparison of confirmation measures. *Cognition*, *103,* 107–119. doi:10.1016/j.cognition.2005.09.006

Trope, Y., & Bassok, M. (1982). Confirmatory and diagnosing strategies in social information gathering. *Journal of Personality and Social Psychology*, *43,* 22–34. doi:10.1037/0022-3514.43.1.22

Trope, Y., & Liberman, A. (1996). Social hypothesis-testing: Cognitive and motivational mechanisms. In E. T. Higgins & A. W. Kruglanski (Eds.), *Social psychology: Handbook of basic principles* (pp. 239–270). New York, NY: Guilford Press.

Trope, Y., & Thompson, E. P. (1997). Looking for truth in all the wrong places? Asymmetric search of individuating information about stereotyped group members. *Journal of Personality and Social Psychology*, *73,* 229–241. doi:10.1037/0022-3514.73.2.229

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185,* 1124–1131. doi:10.1126/science.185.4157.1124

Van Wallendael, L. R. (1995). Implicit diagnosticity in an information-buying task. How do we use the information that we bring with us to a problem? *Journal of Behavioral Decision Making*, *8,* 245–264. doi:10.1002/bdm.3960080403

Villejoubert, G., & Mandel, D. R. (2002). The inverse fallacy: An account of deviations from Bayes's theorem and the additivity principle. *Memory & Cognition*, *30,* 171–178. doi:10.3758/BF03195278

Wason, P. C. (1959). The processing of positive and negative information. *Quarterly Journal of Experimental Psychology*, *11,* 92–107. doi:10.1080/17470215908416296

Wason, P. C. (1961). Response to affirmative and negative binary statements. *British Journal of Psychology*, *52,* 133–142. doi:10.1111/j.2044-8295.1961.tb00775.x

Wells, G. L., & Lindsay, R. C. L. (1980). On estimating the diagnosticity of eyewitness nonidentifications. *Psychological Bulletin*, *88,* 776–784. doi:10.1037//0033-2909.88.3.776

Downloaded by [Universita' Milano Bicocca] at 03:17 17 December 2013

APPENDIX A

Sample stimulus from Experiments 1 and 2

Imagine that you have traveled to a planet called Vuma, where there are two types of invisible creatures, Gloms and Fizos. Both types are equally common. That is, 50% of creatures are Gloms and 50% are Fizos. You are told the proportion of Gloms and of Fizos who possess a certain feature. You meet eight creatures and you are asked to estimate the likelihood that it is a Glom [Fizo] based on their answers to a question about a feature. Assume that each creature truthfully answers “yes” or “no” to the question.

Imagine you encounter a creature. Recall that on the planet Vuma 50% of creatures are Gloms and 50% are Fizos.

Experiment 1 version:

	Have gills
Gloms	65%
Fizos	35%

	Have gills	
	YES	NO
Gloms	65%	35%
Fizos	35%	65%

Experiment 2 version:

The creature is asked: “Do you have gills?”.

It answers: “Yes, I do”.

Please estimate the chances in 100 that this creature is a *Glom* [*Fizo*].

There are _____ chances in 100 that this creature is a *Glom* [*Fizo*].

APPENDIX B

Report of the other effects that emerged from the ANOVA performed in Experiment 1

The significant main effect of the answer was qualified by a significant Answer × Focal Hypothesis interaction, $F(1, 105) = 54.49$, $MSE = 1{,}621.55$, $p < .001$, $\eta^2 = .170$, lower bound correction, indicating that, when the focal hypothesis was Glom the “no” answer ($M = 70.7$, $SDE = 2.5$) had more impact than the “yes” answer ($M = 58.4$, $SDE = 2.1$), while when the focal hypothesis was Fizo the “yes” answer ($M = 74$, $SDE = 2.1$) had more impact than the “no” answer ($M = 46.1$, $SDE = 2.5$). This pattern is opposite to the normatively expected one because under the Glom focal hypothesis, “yes” answers lead to higher posterior probabilities than “no” answers for all tests, while under the Fizo focal hypothesis the reverse holds true (see Table 1). Note, however, that this finding probably originates as a by-product of our recoding of the dependent variable.

As expected normatively, there was no effect of test order, $F(1, 105) = 0.34$, $MSE = 759.39$, $p = .563$, $\eta^2 = .000$. However, there was a significant main effect of the focal hypothesis, $F(1, 105) = 5.77$, $MSE = 759.39$, $p = .018$, $\eta^2 = .001$, with higher estimates provided when the focal hypothesis was Glom ($M = 64.5$, $SDE = 1.3$) than when it was Fizo ($M = 60.1$, $SDE = 1.3$). This finding is in contrast with the application of Bayes’ rule because the posterior probabilities are overall balanced across focal hypotheses (see Table 1). This effect was qualified by a significant Test Type × Focal Hypothesis interaction, $F(1, 105) = 33.64$, $MSE = 943.75$, $p < .001$, $\eta^2 = .061$, lower bound correction. Pairwise comparisons revealed that estimates were significantly higher when the focal hypothesis was Glom ($M = 74.6$, $SDE = 2.1$) rather than Fizo ($M = 52.1$, $SDE = 2.1$), $p < .001$, for the 98–50 test. In contrast, estimates were significantly higher when the focal hypothesis was Fizo ($M = 66.7$, $SDE = 2.3$) rather than Glom ($M = 55.2$, $SDE = 2.3$), $p = .001$, for the 50–2 test. This pattern of results relative to asymmetric tests is opposite to the normative one: Estimates should be higher under the Fizo focal hypothesis than under the Glom focal hypothesis for the 98–50 test, and vice versa for the 50–2 test (see Table 1). Again, note that our recoding of the dependent variable might explain the counternormative findings concerning the effect of focal hypothesis. However, there were no significant differences between the estimates under the Glom hypothesis and those under the Fizo hypothesis for either of the symmetric tests, $ps \geq .123$, a finding that is consistent with the application of Bayes’ rule (see Table 1).

There were other significant interactions, which we report without further discussion for the sake of concision and because they were tiny effects. Specifically, there were significant three-way interactions among test type, answer, and focal hypothesis, $F(1, 105) = 4.07$, $MSE = 788.70$, $p = .046$, $\eta^2 = .006$, lower bound correction, and among test type, answer, and test order, $F(1, 105) = 5.79$, $MSE = 788.70$, $p = .018$, $\eta^2 = .009$, lower bound correction. Finally, the three-way interaction among test type, test order, and focal hypothesis was marginally significant, $F(1, 105) = 3.34$, $MSE = 943.75$, $p = .070$, $\eta^2 = .006$, lower bound correction.

APPENDIX C

Report of the other effects that emerged from the ANOVA performed in Experiment 2

There was a significant interaction between answer and focal hypothesis, $F(1, 90) = 43.00$, $MSE = 947.87$, $p < .001$, $\eta^2 = .120$, lower bound correction, showing that when the focal hypothesis was Glom, the “no” answer ($M = 71.2$, $SDE = 2.2$) was weighed heavier than the “yes” answer ($M = 61.7$, $SDE = 1.9$), $p = .003$, while when the focal hypothesis was Fizo, the “yes” answer ($M = 76.3$, $SDE = 2$) shifted the estimates more than the “no” answer ($M = 56.3$, $SDE = 2.2$), $p < .001$. The pattern shown by this interaction mimics the one found in Experiment 1 and is opposite to the normative prediction because we would expect higher posterior probabilities after “yes” answers than after “no” answers under the Glom focal hypothesis, and the reverse under the Fizo focal hypothesis (see Table 1). Again, this finding might be a by-product of our recoding of the dependent variable.

As expected normatively, neither the main effect of focal hypothesis nor the main effect of test order was significant, Fs < 1, ps $\geq .549$, $\eta^2 = .000$. However, as in Experiment 1, there was a significant Test Type × Focal Hypothesis interaction, $F(1, 90) = 59.52$, $MSE = 784.58$, $p < .001$, $\eta^2 = .138$, lower bound correction, which followed only partially the normative prescription. Indeed, pairwise comparisons revealed that estimates were higher under the Glom focal hypothesis ($M = 78.1$, $SDE = 2.2$) than under the Fizo focal hypothesis ($M = 53.2$, $SDE = 2.2$), $p < .001$, for the 98–50 test, while we would expect the opposite for normative reasons. Indeed, posterior probabilities under the Fizo focal hypothesis are overall higher than those under the Glom focal hypothesis (see Table 1). Furthermore, estimates were higher under the Fizo focal hypothesis ($M = 75.5$, $SDE = 2.3$) than under the Glom focal hypothesis ($M = 56.8$, $SDE = 2.2$), $p < .001$, for the 50–2 test. Again, Bayes’ rule predicts the opposite pattern (see Table 1). However, as in Experiment 1, estimates did not differ significantly under the two focal hypotheses for both symmetric tests, ps $\geq .088$, consistent with the normative prediction (see Table 1).

Finally, there was a significant three-way interaction among answer, test order, and focal hypothesis, $F(1, 90) = 4.00$, $MSE = 947.87$, $p = .048$, $\eta^2 = .011$, lower bound correction, which we do not discuss further for the sake of concision and because the effect was small.

Downloaded by [Universita' Milano Bicocca] at 03:17 17 December 2013